TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2014

105

Joaquín Sanz Remón

Tackling complexity in biological systems: multi-scale approaches to tuberculosis infection

Departamento
Física Teórica

Director/es
Moreno Vega, Yamir

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad Zaragoza
1542

Tesis Doctoral

TACKLING COMPLEXITY IN BIOLOGICAL SYSTEMS: MULTI-SCALE APPROACHES TO TUBERCULOSIS INFECTION

Autor

Joaquín Sanz Remón

Director/es

Moreno Vega, Yamir

UNIVERSIDAD DE ZARAGOZA

Física Teórica

2014

Tackling complexity in biological systems: multi-scale approaches to tuberculosis infection

Memoria de tesis presentada por
Joaquín Sanz Remón

Director
Yamir Moreno Vega

Universidad de Zaragoza
Facultad de Ciencias
Departamento de Física Teórica
Instituto de Biocomputación y Física de Sistemas Complejos BIFI

A Maitane,
porque la espera forma parte de la alegría.

¿Conoces esa enfermedad febril que se apodera de nosotros en las miserias frías,
esa nostalgia de un país ignorado, esa angustia de la curiosidad?

Charles Baudelaire.
Invitación al viaje.

Agradecimientos

Casi cinco años después de comenzar este trabajo, me siento a teclear unas líneas de gratitud para con aquellos que me han ayudado a lo largo de este tiempo. Uno suele tener escasas ocasiones de reconocer públicamente el apoyo, el cariño y la ayuda recibida –que en este caso no es poca– y es mi entender que no hay que desperdiciarlas.

Mi director de tesis ha sido el Doctor Yamir Moreno Vega. Nos encontramos en 2009, y tan sólo las personas que me conocen bien saben hasta qué punto la oportunidad que me brindó planteándome esta tesis cambió mis perspectivas y mi situación personal desde aquellos días desagradables hasta hoy. De aquello ha pasado un lustro y por el camino ha habido tiempo para todo, incluidas las correspondientes travesías por el desierto. No obstante, el resultado final es el texto que estas líneas abren y la certeza de que, muy principalmente gracias a él y a su confianza en mi trabajo, termino esta tesis habiendo aprendido un montón enorme de Ciencia durante este periodo; una Ciencia declamada en singular, entendida como un sólo cuerpo en el cual poco sitio cabe para fronteras entre disciplinas -Física, Química, Biología o Sociología-, las cuales se difuminan, se intentan cruzar con placer y ambición y quedan a menudo diluidas a la categoría de meras discusiones semánticas. El privilegio de haber podido aprender el oficio en estas condiciones es enorme, como enorme es el sentimiento de gratitud y de deuda que siento para con Yamir.

He tenido otros Maestros, además de él, durante mis estudios de Física. Querría acordarme de Juanjo Mazo y Fernando Falo, a través de cuyas clases entré de facto en esta maravillosa madriguera de conejo que es la Dinámica no lineal y sus aplicaciones en Biología; y, sobre todo, querría agradecer todo su apoyo y aprecio a Mario Floría, de quien aprendí durante las primeras fases de esta tesis muchas cosas, no sólo sobre ciencia, sino también sobre cómo hacerla sin perder el equilibrio.

Y he tenido también grandes Maestros antes de acudir a la Universidad. Muchos son los nombres que me vienen a la cabeza, maestros y profesores de la escuela pública, profesionales espectaculares a los que debo, seguramente, toda la curiosidad científica, espíritu crítico e independencia que pueda atesorar. Me quiero acordar de Fernando de la Cueva, de Eva Bajén, de Joaquín Marco y Joaquín Bueno, de Pilar Vázquez, de Paco Durán, porque creo que les gustará saber que les estoy profundamente agradecido y que los años mejoran la perspectiva y le permiten a uno ver lo díficil que es hacer un buen trabajo dentro de un aula, algo en lo que todos los mencionados son o eran alucinantes. Muy especialmente quiero acordarme de Encarna Temiño, de quien me siento terriblemente orgulloso de haber sido alumno, y quien es absolutamente responsable de esta querencia mía por la Biología que ni el tiempo ni la Física han querido aplacar. Estoy convencido de que no puede ni imaginar hasta qué punto la vividez del recuerdo de lo que nos enseñaba en sus clases –infinitamente más mérito suyo que mío– me ha ayudado a navegar sin desnortarme durante mi trabajo en esta tesis.

Además de a ellos, muchísimo es lo que debo a la gente del grupo de genética de mycobacterias de la facultad de Medicina de Zaragoza, quienes, como un Virgilio de

muchas cabezas, han ejercido conmigo de cicerones de paciencia infinita a través de los años, enseñándome algunos de los más oscuros rincones de la microbiología, genética y epidemiología de la tuberculosis, infiernos poblados por la bacteria que de algún modo protagoniza este texto, responsable de más muerte en los últimos dos siglos que cualquier otro ser vivo. Quiero agradecer muy especialmente a Maria José Iglesias y Sofía Samper todas las veces que me han atendido, ayudándome a resolver cuantas dudas tuviera y a abrirme paso entre la literatura médica y epidemiológica. Sobre todo quiero agradecer a Carlos Martín su interés en nuestro trabajo durante este tiempo, su apoyo constante y muy especialmente en estos últimos meses de buscar laboratorio para continuar investigando y solicitar becas postdoctorales. Compartir labor con Carlos ha sido un auténtico placer, e intentar aportar nuestro *teórico* grano de arena a un proyecto de un calado tan enorme como el suyo un improbable y alucinante privilegio que dificilmente podré agradecerle lo suficiente alguna vez.

Asimismo ha sido un inmenso placer compartir el trabajo tanto con el resto de coautores de los artículos que aquí presento como con los compañeros del Instituto. El BIFI es un extraño lugar lleno de buena gente donde nunca va a faltar alguien que te selle un documento un 13 de Agosto, te aconseje gustoso sobre cómo paralelizar un código, o te ayude lo mismo a hacer una web que a escudriñar el hermético material suplementario de un artículo de Biología molecular. Especialmente, le debo más que agradecimiento a Sergio Arregui, que decidió poner el oído un par de tardes a lo que le contaba sobre genes, bacterias y enfermedades, de lo cual confío que termine por no arrepentirse demasiado. Sin su ayuda una gran parte de este texto, sencillamente, no podría existir aún.

A mis compañeros del CosNet lab tengo que agradecerles algo tan simple y esencial como la compañía. El cierzo pega duro en Juslibol, y habría sido peor si la cotidianidad que construimos entre todos hubiera sido menos sencilla. Igualmente, quiero agradecer la acogida a la gente del ISI en Turín, así como a los integrantes del grupo de dinámica no lineal y láseres en Tarrasa, porque, si es un lujo raro sentirse en casa cuando uno está de estancia, a mi me ha pasado en dos lugares diferentes. Con Jordi García Ojalvo estoy particularmente en deuda, además de por el trabajo que tenemos a medias -doceavo y ausente capítulo de este texto-, por haberme hecho tan sencilla mi estancia en Barcelona, y por dos o tres lecturas e ideas que han tenido una influencia central en partes del trabajo que nada tienen que ver con lo que exploráramos en aquella agradable primavera de 2012 en la capital catalana.

En otro orden de cosas, necesito especialmente dar las gracias a la persona que más me ha ayudado -en la acepción más literal y noble del verbo- durante estos años: mi tío Santiago, quien -no os quepa duda- sabe más que todos nosotros juntos. Asimismo, estoy agradecido a un pequeño puñado de amigos de Ejea, por cada vez que me llaman por enésima vez, sin importar cuantas veces no conteste, y a los muchachos del *clut*, virtuosos del dardo, por permitirme colocarle el sistema de referencia a todo esto, cada Jueves a las ocho y media de la tarde en el número 2 de Ángel Ganivet.

Por último, estoy aún más agradecido a mi familia -tierna e impagable escuela de complejidad-, muy especialmente a mis hermanos y a mi madre.

Mi madre es una mujer de casa humilde que un día decidió vivir cinco vidas además

de la suya, una por cada hijo que sacó adelante. En ella, la generosidad funciona como el corazón, orgánica e inevitablemente, y es su nombre completo Gloria Remón Villarreal.

Luis Rosales dejó escrito que el agradecimiento es camastrón y llega a misa tarde. Yo voy a estar siempre agradecido a mi padre, que probablemente habría dado por buena la gravedad del aforismo, con *somardonería*, de haber venido a dar ante estas líneas.

Abbreviatures

List of abbreviatures used in this thesis:

- AFRH: Africa, high HIV (Region formed by african countries with more than 4% of HIV prevalence).
- AFRL: Africa, low HIV (Region formed by african countries with less than 4% of HIV prevalence).
- AFV: adolescent focused vaccine.
- AIDS: acquired immunodeficiency syndrome.
- BA: binding assay.
- BCG: bacillus Calmette-Guérin (vaccine).
- BZA: benzamide.
- cfp10: culture filtrate protein of 10 kDa.
- DETA-NO: diethylenetriamine nitric oxide adduct (A NO donor).
- EMR: east Mediterranean region.
- EMSA: electrophoretic mobility shift assay.
- ER: Erdos-Renyi (network).
- ESAT-6: early secreted antigenic target of 6 kDa.
- FBL: feedback loop.
- FR: folding rate.
- FFL: feed-forward loop.
- GFP: green fluorescent protein.
- HIV: human immunodeficiency virus.
- HMF: heterogeneous mean field.
- IGRA: interferon-gamma release assay.
- KO: knock-out mutant.
- KronEM: Kronecker expectation-maximization algorithm.

- LMA: Levemberg-Marquardt algorithm.
- MA: microarray.
- MF: mean field.
- $M\phi$: macrophage.
- MMC: mitomycin C.
- mRNA: messenger ribonucleic acid.
- *MTB*: *Mycobacterium tuberculosis*.
- NAM: nicotinamide.
- NFV: newborn focused vaccine.
- PCL: poorly characterized link.
- PDIM phthiocerol dimycocerosates.
- PPIN: protein-protein interaction network.
- PZA: pyrazinamide.
- qRT-PCR: quantitative real time polymerase chain reaction.
- SBM: stochastic block model.
- SDS: sodium dodecyl sulfate (a detergent).
- SEAR south East Asia Region.
- SEIR: susceptible-exposed-infected-recovered (model).
- SF: scale-free (network).
- SIG: sigma factor.
- SIM: single input module.
- SIR: susceptible-infected-recovered.
- SIS: susceptible-infected-susceptible.
- SL: suspicious link.
- SOM: single output module.
- TB: tuberculosis.
- TELC: target expression levels comparison.

- TF: transcription factor.
- TRN: transcriptional regulatory network.
- TSP: triad significance profile.
- TST tuberculin skin test.
- WCL: well characterized link.
- WPR west Pacific Region.

Resumen

La tuberculosis (TB), –pese a ser comúnmente considerada como una enfermedad en vías de erradicación, propia de tiempos peores– es, junto a malaria y SIDA, una de las tres mayores causas de mortandad a nivel mundial, con devastadoras cifras en países subdesarrollados de África y Asia. Según estimaciones de la Organización Mundial de la Salud (OMS), en 2012 casi 9 millones de personas desarrollaron la enfermedad en todo el mundo, y 1.3 millones murieron por su causa. Su agente etiológico -el bacilo *Mycobacterium tuberculosis* (*MTB*)– es un parásito humano cuyo sofisticado ciclo de vida involucra mecanismos que tienen lugar a muy variadas escalas temporales y espaciales, convirtiéndolo en un interesante objeto de estudio desde el punto de vista de la Física y la Biología de sistemas complejos, tanto por los retos conceptuales que su estudio plantea como por el potencial impacto científico y social que éste implica [2].

En este contexto se enmarca la presente Tesis Doctoral, a través de la cual se propone el estudio de diferentes aspectos relacionados con el proceso de infección tuberculosa que tienen lugar a diferentes escalas, sirviéndonos para ello de métodos analíticos y computacionales tomados de la Física de Sistemas y Redes Complejas. Dichos aspectos abarcan desde el estudio de la adaptación al medio de células individuales del patógeno, hasta el estudio de la propagación de la enfermedad sobre poblaciones distribuidas en continentes enteros. En lo que respecta a la Biología del bacilo, hemos usado herramientas de análisis de Teoría de Redes Complejas para caracterizar redes de interacciones bio-moleculares, tanto redes de transcripción génetica como redes de interacciones entre proteínas. En lo que respecta al nivel de poblaciones humanas hemos trabajado en la incorporación de determinadas propiedades de las poblaciones humanas sobre modelos clásicos de propagación de la TB, con el propósito de avanzar hacia un conocimiento más detallado de los factores que afectan la propagación de la enfermedad a escala global. Nuestro objetivo aquí es el desarrollo de herramientas cada vez más fiables para la evaluación de nuevas intervenciones epidemiológicas; particularmente vacunas preventivas contra la TB.

Esta tesis se estructura en seis partes: una introducción (parte I); dos partes dedicadas al estudio de redes de interacciones bio-moleculares (partes II y III); dos partes más dedicadas al desarrollo de modelos epidemiológicos de propagación de enfermedades (partes IV y V) y una parte final para las conclusiones VI.

En la introducción expondremos los principales conceptos utilizados, o simplemente mencionados, durante el resto de capítulos. Por una parte, esto incluye un breve repaso de los principales métodos, modelos y conceptos de la Teoría de Redes Complejos y sus aplicaciones en Biología y Medicina (capítulo 1), los cuales constituyen los fundamentos metodológicos utilizados en este texto. Por otra parte, una vez que los métodos que utilizaremos han sido contextualizados, discutiremos las principales características biológicas del principal objeto de estudio de esta Tesis: el patógeno *MTB* y la

enfermedad de la que es responsable (capítulo 2).

En las secciones siguientes, los contenidos de cada parte del texto son desglosados.

Partes II y III: *Interactomas* celulares como redes complejas

A buen seguro, entre todas las escalas biológicas implicadas en el proceso de infección de la TB, el nivel celular es aquel sobre el cual una cantidad más grande de datos experimentales ha sido recopilada durante la última década. La generalización y el abaratamiento de las técnicas de secuenciación genómica; junto al florecimiento de diversas técnicas experimentales en transcriptómica y proteómica, ha generado una proliferación masiva de datos experimentales que ha sido explotada sólo parcialmente. En este sentido, la investigación en TB ha asistido durante los últimos años a la compilación de los primeros *datasets* relativos a *interactomas* celulares con ánimo de completitud, desde redes de regulación transcripcional hasta redes de interacciones entre proteínas y mapas metabólicos [89, 4]. Las técnicas experimentales siguen avanzando en precisión, precio, flexibilidad y alcance, y una fracción cada vez mayor de los datos generados se dispone libremente en bases de datos *on line* [5]. El procesamiento e integración de dichos datos de origen diverso alrededor nuevos modelos basados en redes ofrece múltiples posibilidades, y sin duda contribuirá a mejorar nuestro conocimiento de la enfermedad desde una perspectiva sistémica.

En este contexto, en la parte II de esta Tesis, caracterizamos y estudiamos diversas redes bio-moleculares en *MTB*. El punto de partida de esta parte del trabajo (capítulo 3), consiste en la compilación de la red de regulación transcripcional del bacilo [3]. Reuniendo toda la información disponible en la literatura sobre regulación transcripcional en *MTB* alrededor de un único *dataset*, construimos una red de regulación que, además de constituir un recurso útil por sí mismo, nos permite analizar la maquinaria de control genético en la bacteria desde una perspectiva de red.

Además, en el capítulo 4, estudiamos cómo la respuesta a los diversos estímulos ambientales que *MTB* debe afrontar durante su ciclo de vida se refleja en la estructura de su red de interacciones entre proteínas. Para completar esta tarea, hemos estudiado cómo dicha red se transforma a consecuencia de los diferentes perfiles de expresión genética que la bacteria reproduce en respuesta a diferentes condiciones experimentales. En este punto hemos sido capaces de descubrir un elevado grado de coherencia en la estructura topológica de la red en respuesta a algunos de los principales *stresses* que la bacteria afronta al entrar en la célula huésped, así cómo el rol privilegiado que algunos de los principales antígenos de la bacteria juegan en este contexto.

Por otro lado, en la parte III de esta Tesis, nuestro principal objetivo es el de utilizar la red de regulación genética del bacilo, junto a otros sistemas similares, para estudiar la topología de este tipo de redes desde una perspectiva más genérica, revisando ciertos resultados clásicos en el campo y extendiendo el rango de utilidad de otros. En esta parte hemos prestado especial atención al papel jugado por la incertidumbre e incompletitud que acompañan, de manera inherente, a la información a partir de la

cual se construyen estas redes. Específicamente, en el capítulo 5 nos dedicamos a caracterizar el efecto que la incompletitud de los datos puede llegar a ejercer sobre el resultado de ciertos análisis topológicos en redes genéticas [6]; mientras que, en el capítulo 6, proponemos nuevos métodos para la medida de la fiabilidad de los enlaces en redes complejas dirigidas [7].

Partes IV y V: modelos epidemiológicos

En los últimos años, la Teoría de Redes ha realizado una serie de relevantes aportaciones metodológicas a la epidemiología matemática que han permitido describir la relación existente entre ciertas características de las poblaciones humanas y la facilidad con la que las enfermedades infecciosas se transmiten a través de ellas. Aspectos como las estructuras de las redes de contactos, los patrones de movilidad e incluso los cambios en el comportamiento causados por la irrupción de una enfermedad, han sido profundamente estudiados, y su sorprendente influencia sobre la propagación de enfermedades, caracterizada [8, 9, 10]. Adicionalmente, importantes *datasets* en relación a estas cuestiones han sido compilados y utilizados por la comunidad científica en la implementación de modelos de propagación de enfermedades cada vez más sofisticados [11, 12, 13], lo que ha producido relevantes avances en el campo [14]. Esta confluencia entre avances teóricos y disponibilidad de datos ha hecho posible la aparición de nuevos modelos guiados por datos con elevadas capacidades predictivas [15, 16]; permitiendo su uso por las autoridades, en foros relacionados con la salud pública y la vigilancia y control de enfermedades infecciosas.

No obstante, todos estos avances no han tenido lugar del mismo modo para todas las enfermedades; sino que la gran mayoría de ellos han sido aplicados hasta el momento al estudio de enfermedades de ciclo corto como la gripe [15], caracterizadas por periodos infecciosos mucho menores que la esperanza de vida de los individuos afectados. Pese a la relevancia de este tipo de enfermedades; existen otro tipo de agentes infecciosos cuya propagación no puede ser descrita utilizando las mismas herramientas de modelización. Específicamente, las descripciones de ciertas enfermedades persistentes como la TB suelen basarse en descripciones menos sofisticadas.

Mi principal propósito en este punto ha consistido en extender el marco conceptual de la epidemiología matemática sobre poblaciones estructuradas hacia determinados escenarios epidemiológicos estrechamente relacionados con la TB que han recibido una menor atención hasta la fecha. Esto incluye el desarrollo de modelos epidemiológicos genéricos, de clara inspiración teórica, así como la implementación de modelos guiados por datos útiles para propósitos cuantitativos. Los resultados de mi investigación en este campo se presentan en esta Tesis en las partes IV y V; correspondientes a los capítulos 7, 8, 9, 10 y 11.

De este modo, en una primera fase correspondiente a la parte IV, nos centramos en el desarrollo de modelos matemáticos de propagación de enfermedades desde un punto de vista teórico; a través de los cuales estudiamos la influencia de ciertos aspectos característicos de la propagación de la TB, como son sus largos periodos de

latencia (persistencia) así como sus fuertes interacciones con otros agentes infecciosos: (principalmente, el VIH); cuando las redes de contactos sobre las cuales el patógeno se propaga son heterogéneas. En esta linea, comenzamos derivando, en el capítulo 7, un modelo general para la descripción de enfermedades persistentes sobre poblaciones estructuradas [17, 18]. Los largos periodos de latencia que caracterizan este tipo de enfermedades nos obligan a considerar, de manera explícita, las variaciones de población que, por razones ajenas a la enfermedad, ocurren en las poblaciones bajo estudio mientras se propaga la infección. En el contexto de nuestro modelo, comprobamos como la consideración de dichas variaciones de población introduce nuevas propiedades y plantea nuevos problemas que somos capaces de resolver. Algo similar ocurre, de hecho, cuando consideramos interacciones entre enfermedades que se propagan sobre topologías complejas. En el capítulo 8 exploramos este escenario, en el cual caracterizamos los umbrales epidémicos de parejas de enfermedades propagándose bajo dichas circunstancias.

En una fase ulterior de la investigación, correspondiente a la parte V, nos centramos en la implementación de modelos específicos para TB, guiados por datos, y cuyo objetivo es la evaluación cuantitativa de vacunas preventivas; un tema de actualidad científica, dada la cantidad de proyectos de nuevas vacunas que hoy en día se encuentran siendo testados en diferentes fases de desarrollo clínico [20]. En este sentido, en el capítulo 9, se presenta un nuevo modelo de propagación de la TB, que, tomando como base una serie de trabajos previos de C. Dye y colaboradores [21, 22], incluye una descripción detallada de todos los proceso de infección, reactivación, re-infección, tratamiento y recuperación relevantes en este caso, sobre un esquema de modelización estructurado por edades. A partir de este punto, incorporamos una serie de nuevos ingredientes a nuestro modelo, como patrones de contactos heterogéneos entre distintos grupos de edad, así como un acoplamiento explícito con la evolución demográfica esperada para las próximas décadas en las diferentes regiones del mundo [23]. Confrontando nuestro modelo con datos públicamente disponibles sobre demografía e incidencia de la TB [23, 24] somos capaces de demostrar como esta serie de aspectos, ignorados hasta ahora en los modelos de propagación de TB son capaces de ejercer una influencia poderosa sobre las predicciones de los modelos de propagación. Esto hace evidente la necesidad de mejorar los modelos actualmente disponibles; especialmente si se pretende realizar estimaciones de impacto fiables para las nuevas vacunas actualmente en desarrollo.

Aprovechando el modelo desarrollado en el capítulo 9, en los capítulos 10 y 11 exploramos determinadas cuestiones abiertas en el campo del desarrollo de nuevas vacunas contra la TB, entre las cuales destaca a qué edad debería aplicarse una vacuna antituberculosa para maximizar su impacto. En estos capítulos exploramos cómo la respuesta a dicha pregunta permanece sujeta a determinados factores como la capacidad de los modelos de reproducir las tasas de incidencia de la enfermedad en cada grupo de edad por separado, los distintos niveles de persistencia en el tiempo que las nuevas vacunas puedan ofrecer o, notablemente, la exposición previa de los individuos vacunados a antígenos mycobacterianos.

Summary

Tuberculosis -even if commonly conceived as an old-fashioned disease, proper of past, worse times-, is, along HIV and malaria, one of the three single leading causes of death worldwide; specially in underdeveloped african and asian countries. According WHO estimations, in 2012, almost 9 million people developed the disease all around the world, and 1.3 million died for its cause [1]. Its causative agent, –the bacillus *Mycobacterium tuberculosis*– is a human obligate parasite whose elusive life cycle involves many different spatial and temporal scales, which ultimately make it an optimal target of a research program based on complex systems science, both with respect to the conceptual challenges it poses and to the potential impact that such a project might offer [2].

In this context, the principal purpose of this Thesis is the use of certain methods and tools taken from Physics of complex systems and networks to the study and modeling of tuberculosis disease (TB) at different biological levels, spanning from the study of cellular interactomes within single bacterial cells to the study of the spreading dynamics of TB on top of human societies. From the pathogen side, complex networks analysis (both transcriptional regulatory networks and networks of protein-protein interactions) is used to define systems of bio-molecular interactions within a single pathogen. At hosts' population level, I focus on the incorporation of certain complex features of human societies on top of classical spreading models of TB epidemiology, with the purpose of increasing current understanding of the factors affecting TB burden at a global scale, and deepening in the impact forecasting of novel epidemiological interventions such as preventive vaccines.

This thesis is structured in six parts: an introduction (part I); two parts devoted to the study of bio-molecular networks (parts II and III); two additional parts focused on the development of mathematical models of disease spreading (parts IV and V) and a section for the final conclusions (part VI).

In the introduction, for the sake of completeness, I review the main concepts used, or just mentioned, in the rest of the chapters. On the one hand, this includes a brief review of the most important concepts of complex networks theory and its applications on Biology and disease spreading modeling (chapter 1), which constitute the methodological foundations of the research conducted in this Thesis. On the other hand, once the methods used are put into context, the focus is shifted on the system under study: the pathogen *MTB* and the disease it causes: its life-cycle and epidemiological implications (chapter 2.

In the following paragraphs, the contents of the different parts of the text are summarized.

Parts II and III: Network-based analysis of cell interactomes

Probably, among all biological levels involved in TB infection, the cellular is the one for which a more overwhelming amount of experimental data has been achieved during the past decade. Generalization –or cheapening– of genome sequencing tools, transcriptomics and proteomic techniques have resulted in a particular big-bang of data that has been only partially exploited so far, specially from a systemic perspective. In this context, the first genome wide interactome datasets of the bacterium [3, 4] are being collected, and the amount and quality of the data available is continuously increasing because of both the use of better and more accurate experimental techniques and the generalization of the use of on-line databases [5], which allow repurposing of experimental information. The gathering of these data of disparate origin (and initial purposes) around new network-based models offers many possibilities, and undoubtedly will contribute to the improvement of our current understanding of the pathogen from a systemic perspective.

In this context, in part II, we present a study of diverse bio-molecular networks of *MTB*. The starting point of this part of the Thesis is the compilation of the TRN of the bacillus, in chapter 3. By gathering all the information processed in this chapter around a unique, coherent dataset, we constructed a network that, beyond constituting a useful resource by itself, allowed us to analyze the transcriptional machinery of the bacterium from a networked perspective.

In chapter 4, we study of how the response to the disparate environmental stimuli that MTB has to face during its life cycle reflects into the structure of the PPIN of the bacterium. In order to do so, we have characterized how the PPIN of MTB, as compiled in a genome-wide two-hybrid assay [4], is transformed as a consequence of the different expression profiles that follow the exposure to different stimuli. By doing so, we have discovered a high coherence in the topological structure of the bacterial PPIN in response to the main stresses associated to the pathogen's environment when invading host's immune cells, and a central role reserved for ESAT-6 family of antigens in the network.

In the part III of this Thesis, our first scope consisted in using the compiled TRN of MTB, along with other related networks, to study the topology of TRNs from a general perspective, revising certain classical results on the field and expanding the applicability of others. In this part, we have payed a central attention to the role of data uncertainty and incompleteness on the structure of TRNs. More precisely, in chapter 5 we have come to isolate some of the most relevant caveats of topological analysis on TRNs inherited from the lack of data reliability [6]. Similarly, in chapter 6, we propose some network-based methods to overcome these issues in a general context [7], by studying the performance of different algorithms for the measurement of links reliabilities in complex directed networks.

Parts IV and V: epidemiological models

In the last years, network theory has provided mathematical epidemiologists with new theoretical tools to understand certain features of human populations and how these influence the spreading of infectious diseases. Aspects such as contact networks structures, mobility patterns or behavioral responses against disease have been studied, and their striking influences on the overall disease spreading addressed [8, 9, 10]. Furthermore, relevant datasets regarding all these questions have been rendered available to the research community [11, 12, 13], which has constituted an optimum complement responsible of much progress in the field [14]. This confluence between theory and data availability has allowed the appearance of data-driven epidemic models with meaningful predictive skills [15, 16], so bringing all those scientific developments closer to epidemiological and public health policy grounds.

However, most of these advances have been used so far to describe short-cycle diseases like influenza [15], for which the typical infectious periods are much shorter than individuals' lifetime. Despite of that, many relevant real diseases or epidemiological situations cannot be described under such narratives, which makes their description to lie a step behind. More specifically, current descriptions of persistent diseases, like TB, are grounded on simpler modeling schemes.

Our main purpose in this context has been to extend the conceptual framework of disease spreading on structured populations to some relevant epidemiological scenarios tightly related to TB which have been neglected so far. This includes the development of generic models according to a general, theoretical perspective and the implementation of data-driven, complex models useful for quantitative purposes. The results obtained are presented in this Thesis in parts IV and V; chapters 7, 8, 9, 10 and 11.

Thus, in a first phase, corresponding to part IV, we focus on the development of general epidemic models whereby we studied the influence of aspects relevant to TB spreading, such as long persistence times or disease-disease interactions, when the epidemics being modeled spread through complex, structured populations. In this line, we start by developing, in chapter 7, a general model of persistent infections on structured populations [17, 18]. Long latency periods that characterize persistent diseases like TB are typically longer than single individuals' life expectance, which forces us to consider open populations subject to variation due to natural causes, not related to TB (i.e. births or natural death fluxes). In the context of these models, we see how considering such open populations on models with contact heterogeneities results into new features and challenges -both from the analytical and the numerical points of view- which we are able to successfully address. Something similar occurs in what regards the effects of disease-disease interactions on the spreading dynamics, studied in chapter 8, where we address how this kind of phenomenology influences epidemic phenomena when the heterogeneous nature of human populations is considered [19].

In a second phase, corresponding to part V, we focused on the development of specific data-driven models of TB spreading for the evaluation of novel epidemic interventions, principally age-focused preventive vaccines, motivated by the numerous candidates that are currently under development, being tested in clinical trials [20].

To accomplish that goal we present, in chapter 9, a novel TB spreading model that takes as a basis previous works developed by C.Dye et al. [21, 22] for the description of the disease, which include a detailed description of all relevant processes of infection, re-infection and endogenous or exogenous reactivation, and an age-structured description of the spreading dynamics. From that starting point, we have incorporated some new ingredients, like non-homogeneous population mixing patterns [12] and disparate data driven scenarios of demographic evolution for the populations under study, addressing the influence of these relevant aspects on model outcomes. By confronting our model to publicly available data on trans-national demography and TB burden levels [23, 24], we show how these so-far neglected model ingredients strongly influence model predictions; which makes evident the need of refining current state-of-the-art tools for TB spreading modeling, specially if a reliable estimation of novel vaccines impacts is intended.

Taking advantage of the model developed in chapter 9, we explore, in chapters 10 and 11, different effects of TB spreading dynamics that may exert relevance influences on vaccine impact evaluations. In the context of the vaccine development pipeline, a number of open questions remain, as for example the age segments on which successful candidates should be applied. In this context, in chapters 10 and 11, we evaluate how the answer to this question is subject to factors like models' ability to offer proper age-distributed descriptions of TB burden, the different possible patterns of vaccine persistence or, very relevantly, the presence of prior sensitization to environmental mycobacteria in the populations under study.

Publications related to this Thesis

Part of the work presented in this Thesis (chapters 3, 5, 6, 7 and 8) corresponds to the following published works:

- J Sanz, J Navarro, J Arbués, C Martín, P Marijuán and Y Moreno. The transcriptional regulatory network of *Mycobacterium tuberculosis*. *PLoS One*, **6**, 7, e22178, (2011).
- J Sanz, LM Floría and Y Moreno. Dynamics of persistent infections in homogeneous populations. *Int. J. Bifourc. Chaos*, **22**, 7, 125164 (8 pp.) (2012).
- J Sanz, LM Floría and Y Moreno. Spreading of persistent infections in heterogeneous populations. *Phys. Rev. E*, **81**, 056108/1-056108/9, (2010).
- J Sanz, C Xia, S Meloni and Y Moreno. Dynamics of Interacting diseases. *Phys. Rev. X*, in press, (2014).
- J Sanz, E Cozzo, J Borge-Holthoefer and Y Moreno. Topological effects of data incompleteness of gene regulatory networks. *BMC Sys. Biol.*, **6**, 110, (2012).
- J Sanz, E Cozzo and Y Moreno. Data reliability in complex directed networks. *J. Stat. Mech.* **P12008**, (2013).

The rest of this Thesis corresponds to more recent, yet unpublished work (chapters 4, 9, 10 and 11). Other works published during the Thesis period, not mentioned in the text are the following:

- CY Xia, Z Wang, J Sanz, S Meloni, and Y. Moreno, Effects of delayed recovery and nonuniform transmission on the spreading of diseases in complex networks. *Physica A*, **392**, 1577 (2013).
- E Cozzo, J Sanz, and Y Moreno, Dynamics of Biomolecular Networks. Chapter contribution to the Encyclopedia of Molecular Cell Biology and Molecular Medicine. (Edited by Robert A. Meyers), ISBN: 978-3-527-32607-5 (Wiley-VCH, Weinheim, 2012).

Contents

Part I
Introduction

To begin with, the art of jigsaw puzzles seems of little substance, easily exhausted, wholly dealt with by a basic introduction to Gestalt: the perceived object – we may be dealing with a perceptual act, the acquisition of a skill, a physiological system, or, as in the present case, a wooden jigsaw puzzle – is not a sum of elements to be distinguished from each other and analysed discretely, but a pattern, that is to say a form, a structure: the element's existence does not precede the existence of the whole, it comes neither before nor after it, for the parts do not determine the pattern, but the pattern determines the parts: knowledge of the pattern and of its laws, of the set and its structure, could not possibly be derived from discrete knowledge of the elements that compose it. That means that you can look at a piece of a puzzle for three whole days, you can believe that you know all there is to know about its colouring and shape, and be no further on than when you started. The only thing that counts is the ability to link this piece to other pieces, and in that sense the art of the jigsaw puzzle has something in common with the art of go. The pieces are readable, take on a sense, only when assembled; in isolation, a puzzle piece means nothing – just an impossible question, an opaque challenge. But as soon as you have succeeded, after minutes of trial and error, or after a prodigious half-second flash of inspiration, in fitting it into one of its neighbours, the piece disappears, ceases to exist as a piece. The intense difficulty preceding this link-up – which the English word puzzle indicates so well – not only loses its raison d' être, it seems never to have had any reason, so obvious does the solution appear. The two pieces so miraculously conjoined are henceforth one, which in its turn will be a source of error, hesitation, dismay, and expectation.

George Perec
Life, a User' s manual, 1978
(translation by D. Bellos in 1987)

Chapter 1

Applications of networks theory in Medicine and Biology

Complexity is the most genuine fingerprint of life. Its manifestations can be recurrently identified at virtually all scales, ranging from simple biomolecules, organelles, cells and tissues to organisms, populations and, finally, the entire Biosphere. In all these levels, biological systems are usually described as large ensembles of agents hatching complex networks of interactions among them, whose structure and dynamics are in turn coupled to other systems, often belonging to different scales. In that sense, as George Perec wrote [25], Biology has much to do with the art of solving jigsaw puzzles: each piece by itself is hardly informative of the truly nature of the whole picture. Infectious diseases are not an exception to this rule. From pathogens' biochemistry to host-pathogen cross talk processes and, finally, the transmission mechanisms that take place over hosts' populations, bio-molecular, medical and epidemiological research adapts its focus on a large range of scales, both spatial and temporal, where complexity is pervasively found.

In the last decades, complexity sciences have emerged as a precious tool that promises to allow the study of entire biological entities and processes as a whole, overcoming classical reductionist approaches according to which one was forced to split systems into isolated sub-elements of smaller entity, as the only way to gain meaningful insight of their biological functions. The circumstances that have allowed scientific community to start dreaming at this new way of studying Biology have to do with the confluence of different critical advances in experimental, theoretical and computational grounds. On the one hand, the advent of high throughput techniques to cell Biology has rendered available to the scientific community an unprecedented and massive amount of valuable data about genomics, transcriptomics and proteomics, on top of which reconstruct a faithful portrait of cellular Biology at a global scale. Complementarily, the very availability of this new source of information has motivated the development of new theoretical and computational tools, and their applications on a field whose denomination -(Complex) Systems Biology, computational Biology, theoretical Biology- is yet subject to certain contingency.

The birth of complex networks' theory, in the last years of XXth century, and the development of its applications program in Biology have arguably constituted one of the theoretical advances of deeper implications in the field. In the precise context of the study of infectious diseases, complex networks have brought with them relevant breakthroughs, from epidemiology –by providing the conceptual background for the characterization of contact networks relevant for disease transmission– to cell Biology, where gene regulatory systems, protein interactome datasets and metabolic maps are systematically studied as complex networks with relevant results that find direct

translation on Medicine, Pharmacology and Agriculture.

In this chapter, we will introduce the principal concepts of complex networks theory, which has constituted one of the main theoretical basis for the research conducted around this thesis, as well as its principal applications on the study of biological systems.

1.1 About complex networks

In a first approximation, networks science can be defined as the scientific discipline devoted to the study and modeling of whatever system composed by a multitude of agents interacting trough channels which share some set of topological, non trivial properties which are assimilable neither to random connection patterns nor to regular ones. These properties may include heterogeneous connectivity distributions, short average path lengths between nodes, specific local (and intermediate) aggregation patterns, exotic scaling features etc; as we will review in the following lines. These properties have been shown to recurrently appear when real systems are modeled as networks, in fields as disparate as Biology, computer science, Sociology and engineering, making pointless their description using classical models borrowed to Graph Theory, as regular lattices and random graphs. Furthermore, the emergence of these topological footprints of complexity exerts a crucial influence on the dynamical processes that take place over the systems themselves.

1.1.1 Concepts and elementary definitions

A complex network is nothing else but a particular type of *graph*, i.e. a mathematical object $\mathcal{G} = \{\mathcal{N}, \mathcal{L}\}$ defined as a set of $N > 0$ elements or nodes, connected through L links among them. We can denote the i-th node as n_i ($i \in [1, N]$) and the j-th link, connecting nodes n_α and n_β as $l_j = (n_\alpha, n_\beta)$ ($j \in [1, L]$ and both $\{\alpha, \beta\} \in [1, N]$). Probably, the most elementary object codifying all the information within a network is the *adjacency matrix*; a NxN matrix whose entry $A_{i,j} = 1$ if the link between n_i and n_j exists, and 0 otherwise:

$$A_{i,j} = \begin{cases} 1 & \text{if } \exists \quad l_k = (n_i, n_j) \\ 0 & \text{if } \nexists \quad l_k = (n_i, n_j) \end{cases} \tag{1.1}$$

As it is obvious, in different grounds, nodes and links represent different concepts. For example, in a protein-protein interactions network (PPIN), nodes will be proteins, and links will connect pairs of nodes that are able of physically interact to form –even transitorily– a protein complex. Instead, if we think about a social network, nodes will represent persons and links social interactions among them. The strong resemblance that appear among systems as disparate as these, once described as complex networks, is probably the most striking fact in the discipline, and allow us to use a similar toolbox to study a disparate range of systems that traditionally belonged to diverse disciplines.

In the most simple case, links constitute un-ordered couples of nodes, so representing

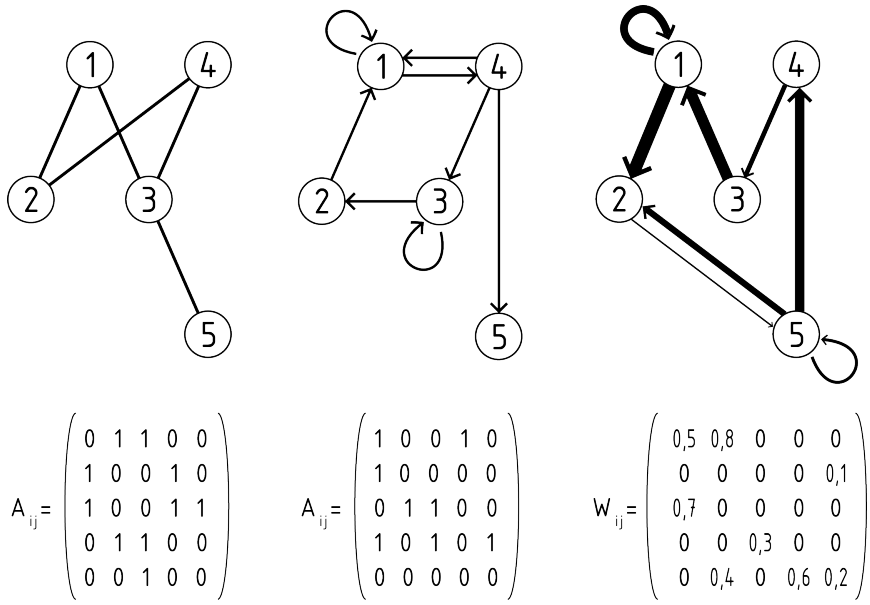


FIGURE 1.1: Types of networks according to links directionality and weights. Left: undirected, unweighted network . Center: directed, unweighted network. Right: weighted directed network. Below the graphs, we represent the corresponding adjacency matrices $A_{i,j}$ or, for the case of the weighted network, the weights matrix $w_{i,j}$

symmetric interactions between them. In such cases, the adjacency matrix is symmetric ($A_{i,j} = A_{j,i}$) and the sum over both indexes of the adjacency matrix entries is twice the number of links: $\sum_{i,j} A_{i,j} = 2L$. In that case, the maximum number of links that a network with N nodes can admit is, in principle $N(N-1)/2$. This is true for networks in which 1) self-interactions are not allowed (i.e. links of type $l = (n_\alpha, n_\alpha)$ are forbidden, and so $A_{ii} = 0 \forall i$) and 2) there cannot be multiple links between a given pair of nodes.

In case self-interactions, but no multiple links are allowed, the maximum number of links is trivially equal to $N(N+1)/2$. That is the case, for example, of the PPIN, in which a protein can interact with itself to form a homo-dimeric complex, but, almost by definition, multiple links are not allowed. In case multiple links are allowed, the upper bound for the number of links in lost. An example of such network could be recovered from a list of interaction events stablished between a set of individuals. These could be radio calls among a set of radio operators [26] or e-mails interchanged among a set of employees of an University [27]. In these cases, each interaction (radio call or e-mail) can be listed as a link, and so, limitless links might appear both in the whole network and between two given nodes of it.

From the idea of networks allowing multiple links, the transition to weighted graphs naturally emerges. In the networks mentioned before, the number of links between a given pair of nodes can be conceived as the intensity of *the* interaction between them. This conceptual step is naturally described by allowing the adjacency matrix to

adopt any natural value: $A_{i,j} \in [0, \infty)$. The situation naturally suggest to perform a normalization process, for example, dividing $A_{i,j}$ by L, to recover a rational number -instead of a natural one- for representing the intensity, or *weight* of the interaction between n_i and n_j. The kind of networks in which interactions are labeled with rational numbers, instead boolean (or integer) variables, are known as *weighted networks*. In that case, the adjacency matrix is substituted by a weight matrix $w_{i,j}$ of the same dimension, containing the intensity of the interaction –normalized or not–, between nodes n_i and n_j. It is worth noticing that weighted networks can be reconstructed from disparate types of data beyond the normalization process mentioned before for systems containing multiple links. Many natural and engineered systems can spontaneously be described as weighted networks of *real* weights, as the intensity of the interactions can be explicitly measured as a continuous variable. In these systems, the notion of number of links in the network loses importance, as typically the whole weight matrix (or a large part of it) may contain non-zero values, representing some degree of residual interaction among most nodes in the system: different thresholds levels imposed on interaction weights will give different numbers of relevant (or discarded) interactions. Furthermore, networks contemplating negative wights (or values of the adjacency matrix) are commonly denoted as *signed networks*, and in these, interactions can be attractive or repulsive, represent correlation or anti-correlation etc.

In some occasions, the systems being modeled consist of nodes that establish asymmetric interactions among them. An example is constituted by gene regulatory networks, in which nodes are genes and links are regulatory interactions that point from the regulatory gene to the gene whose activity is regulated. As a matter of fact, the examples introduced two paragraphs before (i.e. radio calls or e-mail networks) can also be considered (should, if the information is available) as an example of this, as in these cases, the communication is asymmetric as there is an information sender and an information receiver. Networks of asymmetric interactions are noted as *directed networks*, and within them links are ordered nodes' pairs and the adjacency matrix (or the weight matrix, if applies) is not symmetric any more: $A_{i,j} \neq A_{j,i}$, or $w_{i,j} \neq w_{j,i}$, generally speaking. In directed networks not containing multiple links, the maximum number of links possible are $N(N-1)$, if self-loops are forbidden, and N^2 otherwise.

Finally, as a last, particularly relevant type of systems, *bipartite networks* consists of two different sets of nodes $\mathcal{N}$ and $\mathcal{M}$ belonging to two distinct categories and a set of links connecting pairs of element, one of each group: $\mathcal{G}_{\mathcal{B}} = \{\mathcal{N}, \mathcal{M}, \mathcal{L}\}$. Such a system will have N nodes of the first class, M nodes of the second class and L link connecting pairs of nodes of different sets. The i-th mode in $\mathcal{N}$ is n_i ($i \in [1, N]$), the j-th node in $\mathcal{M}$ is m_j ($j \in [1, M]$), and an eventual link connecting them could be $l_k = (n_i, m_j)$ with $k \in [1, L]$. Relevant systems described as bipartite networks are metabolic maps. Accordingly, metabolites taking part in diverse chemical reactions occuring within a cell are associated to first-class nodes, while nodes of the second class represent the chemical reactions. Generally speaking, a metabolic network is both bi-partite and directed, as the participation of a metabolite in a reaction as a reactive is represented as a link from the metabolite to the reaction and viceversa: links from reaction to metabolites represent the formation of products. Furthermore, this representation of

metabolic systems as networks spontaneously offers and adequate framework for the description of chemical equilibria, as in these case, a metabolite is both reactive and product of the same reaction, and so, it both sends and receives a link from the reaction representing the equilibrium. Adjacency (or weight) matrices of bipartite networks are not squared, but of NxM dimension. Maximum number of links of a bipartite network is NM, if they are not directed and $2NM$ otherwise. Very often, bipartite networks are "bypassed", which means that one of the node's categories is disregarded and two nodes of the other set become connected if they contacted some node of the eliminated category.

In the following sections, unless the contrary is stated, we will refer to undirected networks.

1.1.2 A brief history of networks theory: from systems to models.

From the foundations of Graph Theory –commonly dated in 1736, with the popular problem of Köninsberg bridges solved by L.Euler [28]–, more than 250 years passed until the notion of complex network was firstly used, at the end of the XXth century. The reason for this *dormancy* period for the discipline, which for the rest can be considered as a particular branch of Graph Theory, is related simply to the impossibility of observing any system assimilable to a complex network until very recently. Whenever system size does not guarantee the emergence of complex features of any type in any system, for a footprint of complexity to be identified in a networked system, a minimum number of interacting agents is necessary. For example, a network of four nodes can never be considered complex, as no statistically relevant connection pattern, modular structure, clustering, or correlation of any type could be confidently addressed by an elementary problem of lack of resolution. Thus, the identification of these topological traits, which are privative of complex networks would not be possible until systems of multiple interacting agents of enough size (probably more than 1000, although for some purposes this is yet too small) could be characterized and analyzed in a systematic way. These tasks require computational resources that, even if nowadays are trivially affordable by a laptop or even a phone, have resulted privative until the advent of computers and modern telecommunications, which have allowed massive generation and gathering of data and their further analysis.

Indeed, the first systems characterized as complex networks were assembled and analyzed in the last decade of the XXth century [29, 30]. These first examples included collaboration networks (in Science and cinema) built up upon on-line databases, the US power grid, the synaptic wiring of the netamatode *C.elegans*, Internet itself and the world wide web [29, 31]. Strikingly, these systems were found to share a common set of *exotic* properties, that have attracted massive attention since then. Some of these topological treats that were identified first are the small-world property, high clustering levels and scale freedom. We will review in the following lines the definition of these unequivocal characteristics of complex networks and the way researchers have tried since then to build up relevant network models capturing them.

Small path-lengths: the Erdös Renyi random graph model.

A cardinal property of networks is the distance that separate a typical couple of nodes in the system. More formally speaking, in any network we can define the *average path length* as follows:

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j \neq i} d_{i,j} \tag{1.2}$$

where $d_{i,j}$ is the distance between nodes n_i and n_j. if the network is disconnected, some nodes are isolated from others, and, in that cases $d_{i,j}$ is not always defined. A possibility for recovering a meaningful average path length is to average $d_{i,j}$ just among the nodes' couples for which $d_{i,j}$ is defined or, instead, to average the inverse: $d_{i,j}^{-1}$, for all node pairs, which is called *network efficiency*.

Strikingly, all the networks analyzed in those first amazing years of networks science presented strikingly low average path lengths, when compared with systems sizes. As an example, a network of $N = 209293$ co-authors of scientific papers in neurosciences had an average path length of just $\langle d \rangle = 6$. This observation made the experiments of the famous psychologist S.Milgram regain popularity, through which the scientist found that the average number of steps for a package to get from randomly selected people in Omaha, Nebraska to reach one target person in Boston (an stranger person, for the package initial senders) were less than six [32]. These experiments, conducted in 1967, were not the only *vintage* scientific contribution rescued from the analysis of these novel complex networks, but the random graph model of Erdös and Renyi, developed at the end of the fifties [33, 34], also was.

According to the Erdös-Renyi (ER) random graph model, a network is built just by assigning, to each possible link in the system, a probability p to occur. The most relevant feature of this model is that, in the point $p_c = 1/N$, it presents a *phase transition*: ER networks with $p > p_c$ present a *giant component* whose size is a macroscopic fraction of the entire system within which each node is reachable from any other node (i.e. $d_{i,j}$ distances are defined); instead, for $p < p_c$, the system splits up into a myriad of small, disconnected clusters none of which reach a macroscopic size.

For ER graphs above the critical threshold p_c, average path lengths scales at a logarithmic pace with system's size: $\langle d \rangle \simeq \log(N)$. For example, a network of the same number of nodes and links of the network of neuroscientists collaborations mentioned before, built following ER model, presents even a lower average path length: $\langle d \rangle = 5.01$ [31]. This means that classical ER model is able to essentially explain low path lengths found in real systems.

High clustering: random graphs

Another relevant feature of real systems is the notion of transitivity, i.e. the tendency for nodes that share common neighbors in the network to form links among them. This property, that finds its more immediate interpretation in social grounds (*my friends' friends are my friends*), implies the appearance of *triangles* in the network, and its

mathematical definition is indeed the quotient between the number of triangles in a network and the number of connected triplets:

$$T = \frac{\#\ \ triangles}{\#\ \ connected\ \ \ triplets} \tag{1.3}$$

Closely related to this idea, we define clustering coefficient of a node in an undirected network as the number of links existing E_i among neighbors of node i, divided by the maximum number possible of that number, defined by $k_i(k_i - 1)$, where k_i is the connectivity, or degree of n_i, that is, its number of neighbors.

$$c(i) = \frac{E_i}{k_i(k_i - 1)/2} \tag{1.4}$$

From there, network clustering is obtained by averaging local clustering on the whole system: $C = \langle c(i) \rangle$.

The problem with clustering was that real networks showed *high* levels of clustering systematically higher than those produced by ER model. For example, the network of neuroscience collaborations presents a clustering equal to $C = 0.76$, while the associated ER network of same number of links and nodes presents only $C = 5.5 \cdot 10^{-5}$, four orders of magnitude less.

This does not mean, though, that there is no classical graph model able to produce networks with high clustering. Indeed, maybe the most elementary network model: regular lattices, resent high levels of clustering. For example, let us think on a ring of size N (with $4 << N$), in which each node is connected to its 4 closets neighbors (2 on each side). In this regular network, it is straightforward to show that all nodes have the same clustering: $C = 1/2$.

But the picture is not complete yet. Regular models are able to produce high clustered networks, but, when regular models of the same sizes of real systems were built (i.e. same numbers of nodes and links), average path lengths were too high to reproduce those of real systems. The order of regular systems seemed to favor clustering, but not reduced path lengths, and the randomness of ER networks worked in the opposite way. The question, at this point, seem obvious: is there a way to produce a network with reduced path lengths and high clustering *at the same time*? That kind of systems were named *small-world* networks.

The small-world network model by Watts and Strogatz

The answer is obviously affirmative. In 1998 [29] Watts and Strogatz described a model able to produce small-world networks following a rather simple approach: if you know where to find *red* objects and you also know where to find *big* stuff, when looking for *big, red* things you shall probably look at the middle of those spots.

Somehow following this idea, they conceived a model in which, starting from a ring regular lattice as the one described in the previous section (N nodes on a ring and K neighbors per node, $K/2$ on each side),one fraction $p \in [0, 1]$ of links is randomly rewired. By proceeding this way, authors elegantly build a mono-parametric family

of models that spans from a regular network, when $p = 0$, to a random-regular graph (RRG) in which although all nodes retain the same number of links, their neighbors are randomly chosen among the nodes set. The difference from WS model with $p = 1$ to ER network consists on that, in WS, the total number of links, as well as the number of links per node are fixed, which is not true for ER graphs.

The most relevant feature of this model is that, even for *small* values of p (i.e. after the rewiring of a reduced amount of links), average path lengths are drastically reduced from the case of $p = 0$ (i.e. the regular lattice), because of that rewired links acts as *shortcuts* that allow one to navigate through the system in a much more efficient way. In that regime, however, the local surroundings of most nodes yet resembles that of the regular network, and so, network clustering is substantially high.

By this way, Watts-Strogatz (WS) method constitutes the first model able to generate networks both clustered and of small average path lengths, something privative of previous models taken from Graph Theory, as regular lattices or random graphs. For that reason, it is commonly considered as the first model of complex networks theory.

The importance of the degree distribution: scale free networks and preferential attachment

A relevant question, though, remained unnoticed in the first works by Watts and Strogatz that, as time passed by, turned into a crucial property of complex networks able to exert strong influences on their behavior, dynamics and robustness: the degree distribution.

As we have already say, we denote the number of neighbors of a node (the number of links) as its connectivity or degree k_i. From there, we can average over all nodes in a network and define its average degree $\langle k \rangle = 2L/N$. Analogously, we can define *in* and *out* degrees of a node (k_i^{in} and k_i^{out}) in a directed system, as the number of incoming and out-going links it has, and, in the most basic case, the average *in* and *out* degrees of the network will verify: $\langle k^{in} \rangle = \langle k^{out} \rangle = L/N = \langle k \rangle /2$. In certain occasions (see chapter 3), when the set of nodes that acts as links senders or link receivers in a network is a reduced subset of the entire system, it can be convenient to re-define these averages restricting them to the nodes with $k^{in} > 0$ or $k^{iout} > 0$ (i.e. to calculate the average out degree of links senders only or the average in degree of sole link receivers).

The hidden property of complex networks that passed unnoticed in the first analysis by Watts and Strogatz was the high heterogeneous patterns that are found in real systems in what regards the way in which connectivities are distributed among the nodes of a network, that is represented mathematically by the *degree distribution* $P(k)$. A regular graph (as well as a WS network) has a degree distribution which is equal to a δ function:

$$P(k) = \begin{cases} 1 & \text{if } k = K \\ 0 & \text{if } k \neq K \end{cases} \tag{1.5}$$

while an ER graph has a degree distribution defined by a binomial:

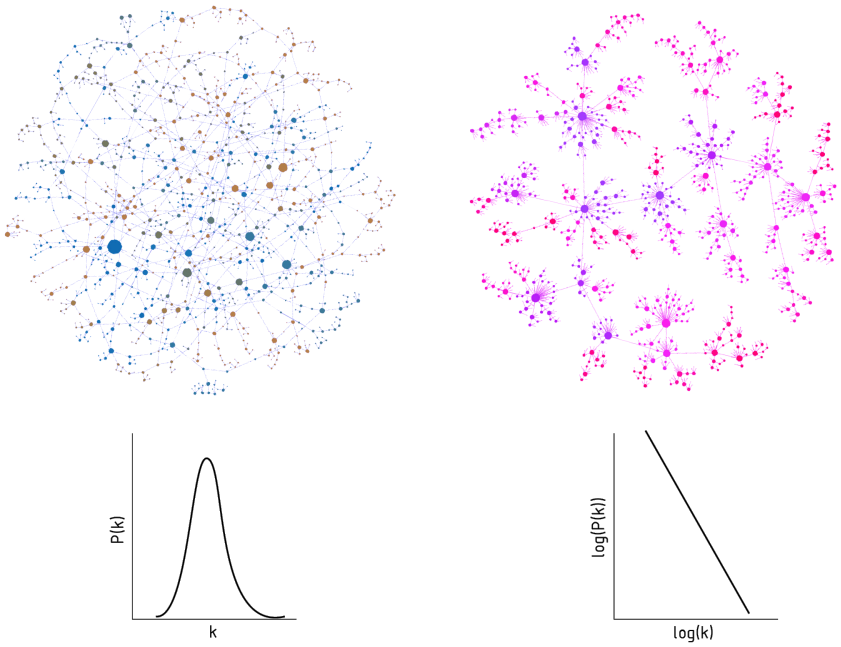


FIGURE 1.2: Left: Erdös Renyi graph. Right: scale free network. The main topological difference between both models is its degree distribution (low).

$$P(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{1.6}$$

where n must be substituted by $n-1$ if self-loops are not allowed.

The behavior of the degree distribution in real networks was found to be radically different from these functions. In the year 1999, Barabasi and Albert found that real systems use to present a connectivity distribution that obeys a power law of negative exponent: $P(k) \simeq k^{-\gamma}$, with $\gamma \in [2,3]$. This finding implied that the real networks analyzed so far had a microscopical structure radically different to those of WS model, random graphs or regular lattices. The key concept is the idea of average degree and the typical deviations of node's degrees around that average: while in network models, degree fluctuations around the average $\langle k \rangle$ are absent (WS and regular graphs) or strongly bound (ER graphs), degree fluctuations in power-law distributed networks may be huge 8see figure 1.2.

For this reason, this kind of systems were named *scale free* networks, as there is no characteristic scale for nodes' degrees (i.e. a typical number of links one random node is expected to have). Remarkably, in this kind of systems, small fractions of nodes -the so called *hubs*– present a very high amount of links when compared with the average, which make them play a central role in many of the different dynamical processes that take place over these systems.

In their seminal work, Barabasi and Albert proposed a model for generating scale free networks based on the mechanism of preferential attachment, according to which,

in a growing system, more connected nodes tend to gain new connections faster than less connected ones. The preferential attachment model (or Barabasi-Albert (BA)) for network generation consists of the following steps: 1) We start with an small network of ($n << N$ nodes in which all are connected to all (which is defined as a *clique*). Then, at each step, we add a new node, which will send $l << L$ links to the already existing network. The probability of sending a link to any of the existing nodes in the network will be proportional to its degree. By proceeding this way, scale free networks of exponent $\gamma = 3$ are produced. Other exponents are possible, if the probabilities scale, instead of with k, with increasing functions of k (typically powers k^a of afine functions $k + A$).

From the works by Watts, Strogatz, Barabasi and Albert, many other network models have been proposed so as to take account of each time more detailed topological features of real networks. However, the works mentioned in this section, the models proposed and the network properties they succeed at describe are probably the most essential in networks theory, both for their future influence on the development of the discipline and the intrinsic importance of the concepts that they develop.

Network motifs in directed systems

Clustering coefficient tells us about the local surroundings of a node in an undirected network. Its definition can be re-formulated by using a terminology based on *triangles* and *connected triplets*, as made with transitivity in equation 1.3: the clustering coefficient of a node in an undirected network is the number of triangles it takes part in divided by the maximum number of triangles it could join, given its number of connections.

When moving to directed networks, the general question of studying the local surroundings of nodes becomes subtler, as many different types of *triangles* are found, as well as an enhanced diversity of *connected triplets*. Leaving apart the task of redefining clustering on directed networks (something that has attained proper attention in the field [35]), the question of studying the different types of connection patterns among reduced groups of nodes (the so called *network motifs*), and their respective abundances in real networks has constituted a prolific research topic in the field.

A motif is a subgraph formed by a low number of nodes -typically 3 or 4- which are connected according to a defined topological pattern. For example, with three nodes we have thirteen possible motifs (or triads) in a directed network.

About motifs, a key question that has been systematically posed during the last years has been: how anormally high –or low– is the number of motifs of each type within a network? Admittedly, the question, as posed, could be reasonably answered by another question: when compared to what? A situation that highlights the importance of having adequate null-models in networks theory to use as meaningful comparison terms in order to test whatever hypothesis.

The more usual null model in networks theory is the configuration model (CM). The configuration model is a model that is built upon an existing network, by preserving the whole degree sequence of it, and randomizing everything else. This means that

any node will have the same connectivity of the original system, but its neighbors will be chosen at random. In the practice, this can be achieved by rewiring links of the existing network, although a CM can also be built from the degree sequence itself, without need of accessing to all the information of the original network (typically encoded in its adjacency matrix). CM's are also widely used for generating scale-free graphs, as these overcome certain "pathological" topological treats of BA networks (e.g. the emergence of degree-degree correlations)

In order to test motifs abundances in real, directed networks, a particular configurational model was used by Alon and co-workers in 2004, which consists in generating CM-versions of the network to study in which, not just the in, and out degrees of each node are preserved, but also the number of feedback loops (bi-directional links) each node has. This can be achieved easily by rewiring separately single-directional links and feedback loops.

In a series of works by Alon and co-workers, they aimed at studying how the number of appearances of each possible triad in different real networks compare with the same measures performed on the corresponding CM-derived null network models. To do that, given a real network, they re-construct a set of CM random versions of the real systems and compare the number of appearances of a triad T on the real system n_T, to the average number of appearances of it within the CM networks of the null ensemble $\langle n_T^{rand} \rangle$, through the following Z score:

$$Z_T = \frac{n_T - \langle n_T^{rand} \rangle}{\sigma_T^{rand}} \tag{1.7}$$

where σ_T^{rand} represents the typical deviation of the appearances of triad T in the networks of the null ensemble, in such a way that positive Z-scores stand for motifs appearing in the real network more than expected by random and viceversa. If one calculates the 13 scores associated to each triad, and normalizes them as if they were the components of a 13-dimensional vector of norm 1, we get the so called *triad significance profiles* (TSPs), which constitute a way to visualize relative trends of a particular network to *choose* certain motifs and to *discard* others.

Technical nuances apart (which will be further discussed in chapters 3 and 5), through these analysis, Alon and co-workers found an striking result: the TSP of different real networks tend to group around four neatly different profiles o superfamilies. This includes social systems, semantyc networks, and two families of biological networks for information processing: one formed by *transcriptional regulatory networks* (TRNs) of unicellular organisms and other one by developmental regulatory networks of pluricellular organisms and synaptic systems.

Community structure in complex networks

Between local metrics (i.e. node degrees or clustering coefficients) and global traits of complex networks (average path lengths, average degrees and clusterings or transitivity) there is a vast range of relevant levels storing meaningful information about the ultimate structure of complex networks. The study of the patterns of aggregation within groups

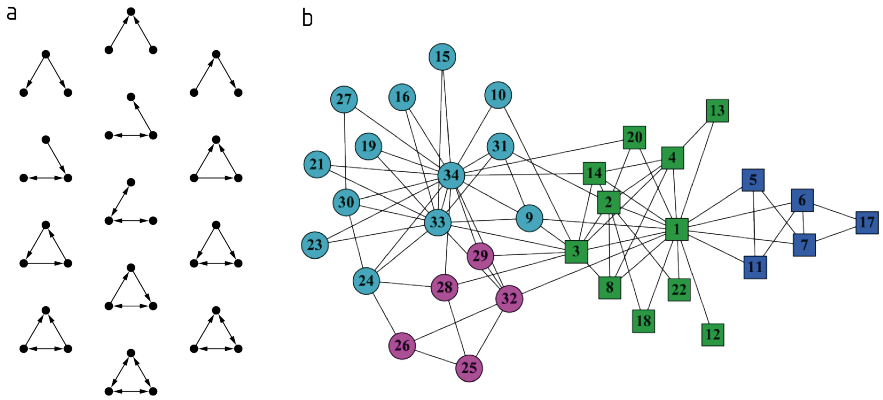


FIGURE 1.3: Community structure and network motifs. A: The thirteen possible connected network motifs in directed networks. B. (Adapted from [37]) Community structure of the social network of Zachary karate club [38]. Nodes shapes and colors represent the community structure of the network at two different scales. Squares and circles represent the community structure resembling the process of social division that took place among the members of Zacchary club. The color-based partition is one particular outcome of the parametric method for modularity optimization proposed in [37].

of nodes (typically, of n nodes, with $1 << n << N$) has attained much attention in the last ten years.

The key question about mesoscale in a complex network can be expressed around two questions: can be identified within a given network subsets of nodes more intensely connected among them *than expected* (taking into account the overall properties of the network)? Can the network be splitted in groups –or *communities*– each one of them presenting more intense internal coherence than external connections with the rest? Generally speaking, we say that a network is modular (or presents a high modularity, or a relevant community structure) if the answers to these questions are affirmative.

Admittedly, the question is complex and subject to ambigüity, as there is no an unique way to define modularity. Notwithstanding that, the first and probably most relevant definition of network modularity is due to Newman [36]. According to Newman definition, we start by defining a partition of the nodes set in groups, assigning each node to one cluster. Let us denote the cluster which node i belong as C_i. Then, we can calculate the modularity associated to that partition as follows:

$$Q = \frac{1}{2L} \sum_{i,j} (A_{i,j} - \frac{k_i k_j}{2E}) \delta(C_i, C_j) \tag{1.8}$$

where δ stands for the Kronecker Delta function. Newman's modularity is a measure of the fraction that links connecting nodes within the same cluster represent, minus what one would expect, for the same fraction, if links were randomly distributed across the network. So, identifying the relevant mesoscale structure of a network means find a partition in clusters maximizing Q.

The problem of modularity maximization has turned into a hot-topic in the field, since modular structure in complex networks have been recurrently found to be related to functional aspects. Since an exhaustive search of all possible partitions is absolutely prohibitive, even for small-size networks, many different computational approaches have been proposed to solve the problem [39, 40, 41, 42], each of which presents advantages and caveats that usually corresponds to an inverse correlation between method's performance and computational requirements and scaling properties. The modularity landscape is usually complex, and characterized by multiple maxima yielding similar values of Q, yet constituting very different partitions. Additionally, modularity maximization is hampered by an intrinsic resolution problem, that impose constraints to the size of partitions accessible to many of the methods of mesoscale analysis. This issue, however, has been treated in a number of works, in which methods for optimizing modularity at different scales are proposed.

1.1.3 Theoretical developments of networks theory

There is yet a long way to walk in networks theory of which the end can hardly be imagined yet. The theoretical development of the discipline goes in parallel to an exponential increase in the quantity and resolution of the data we count with to reconstruct new complex systems as networks. In the so called big-data era, our knowledge of the networked systems that are around us –from human societies to cell interactomes– is each time more thorough and precise.

This explosion of information has allowed researchers to transcend the technical limitations that constrained them to the classical representation of complex systems as static, "simple" complex networks in which nodes and links are frozen objects, immutable in time and intrinsically (i.e. without considering its network properties) identical to each other. In this sense, the study of time evolving networks have emerged as a relevant new branch in the topic, as it allows to study, in a natural way, not only the way in which network's topology influences dynamics of processes taking part over it, but also the way the function of the network modifies the structure of the network itself.

Another relevant emerging thread that can be circumscribed to this context is the development of multi-layer (or *multiplex*) networks theory. In a multiplex network, a set of nodes can interact through different kind of links, typically associated to different layers. Examples of real systems that are naturally modeled as multiplex networks are on-line social networks: although networks like twitter, facebook, Instagram, or even e-mail networks, the blogosphere, etc, can be initially understood as independent systems, actual *nodes of these networks are people*, and no user accounts: the same individual can appear simultaneously in more than one *layer* if, for example she has a facebook account, an e-mail address and a blog.

1.2 Networks theory applications in epidemiology

Epidemic spreading processes constitute a vast field of intense research since nearly a century [43, 44, 45]. The mathematical models developed in this period to describe disease dynamics have become invaluable tools for health authorities. These theoretical tools have been at the root of many of their decisions about strategies of vaccination, prevention and profilaxis [43, 44, 45, 46].

As time goes by, these models have suffered a gradual process of sophistication, very specially during the XXIth century, thanks to the development of networks science and its applications in the field. In this context, one single discovery has supposed a greater impact than any other: contact trees relevant for disease spreading often present the properties of a complex network. From that point on, the influence of these complex properties of contact networks were addressed [47, 13, 14], and, each time, more subtle and precise dynamics [48, 49] considered.

Admittedly, the improvement of computational modeling platforms aimed at providing an increasing accuracy and predictive power necessary implies the gathering of information regarding both population structure and disease specificities. The task of improving the quality and accessibility of the actual information required for an optimal description of epidemic processes is challenging, but in the last few years, the topological characterization of contact networks on top of which epidemics take place has actually stirred up the discipline [50, 51, 52, 49, 53, 54]. This has been feasible thanks to the use of new computational methods, and the incorporation of data coming from such diverse sources as simple statistical surveys [11], mobile phone calls registers [55], bank note mobility patterns [10], global transport networks [53] and geo-demographical distributions of population [56].

Despite of all the progresses made, epidemic modeling has not gained in accuracy for all types of diseases the same way. In fact, as a consequence of its greater -and grateful- mathematical simplicity, SIR (susceptible-infected-recovered) and SIS (susceptible-infected-susceptible) models are the frameworks for which highest levels of accuracy and sophistication have been achieved, both at a theoretical, general level [8, 57, 58, 59] or within more precise, applied scenarios [54, 49, 56]. Moreover, the current degree of modeling sophistication corresponds to short-cycle diseases, whose main feature consists in that individuals become infectious suddenly after becoming infected [45]. These are the cases of diseases that typically transmit from one person to another like respiratory virus and influenzas, two highly topical examples being SARS and H1N1 (influenza A). The fact that the cycle of infection is short-lived allows a most efficient reconstruction of the contact maps between infectious and susceptible individuals and a key theoretical and computational approximation, common to most of current models: the total population size can be assumed to be constant during the outbreak.

Persistent diseases −for which infected individuals can enter into asymptomatic, latent states for not negligible periods before developing clinical symptoms− are the paradigmatic, opposite scheme to short cycle diseases, both regarding the mathematical and computational challenges that their modeling represents, and the relative less

sophisticated models currently available for their description [48, 60, 18]. However, the global impact of persistent diseases -mostly in underdeveloped countries- is everything but a small problem, evidencing the need for a global research effort.

In this Thesis, one of the central objectives is to explore, in the context of persistent diseases like tuberculosis (TB), the influences on the disease spreading process exerted by certain properties of the populations studied, assimilable to cardinal treats of complex networks. In the following lines we will review the models that constitute the basis of our work, as well as the kind of network properties and dynamical behaviors which we aim to study and particularize for the case of TB.

1.2.1 Compartmental models in mathematical epidemiology

Most common theoretical approaches for the spreading of epidemics are based on compartmental models, which offer a description of disease spreading that is based on dividing a population into different subgroups, according to the state of individuals with respect to the disease. Most simple compartmental models are the susceptible-infected-recovered (SIR), firstly introduced by Kermack and McKendrick in the decade of 1920 [61], and, as a variation of it, the susceptible-infected-susceptible (SIS) case.

As a first example, SIR model describes diseases in which susceptible individuals (in class S) can only be infected once (passing to class I), gaining immunity to further re-infection after the first infection (i.e. entering class R). The model, in its classical description, takes into account no more than two parameters: the infectiousness λ, and the recovery rate μ. Each contact between a susceptible individual and an infected one result in a contagion process with a probability λ, i.e.:

$$S + I \overset{\lambda}{\rightarrow} I + I \tag{1.9}$$

On the other hand, for each time stage, the infected individuals have a probability μ to reover:

$$I \overset{\mu}{\rightarrow} R \tag{1.10}$$

At this point, all the individuals in the system are considered to be dynamically equivalent, which is usually designed as the mean field approximation (MF), and supposes that the probability of getting infected is the same for all susceptible individuals, and proportional to I. Furthermore, all individuals are considered to be in contact (well-mixed populations), and the number of contagions per unit time that take place in the system is approximated by the product λSI (where S and I are the number of susceptible and infected individuals, respectively). Under these circumstances, model dynamics is described by the following ordinary differential equations system:

$$\dot{S}(t) = -\lambda S(t)I(t) \tag{1.11}$$
$$\dot{I}(t) = \lambda S(t)I(t) - \mu I(t) \tag{1.12}$$
$$\dot{R}(t) = \mu I(t) \tag{1.13}$$

Nevertheless, for many diseases, an initial episode of disease does not prevent further re-infection. In these cases, spreading dynamics is better described by the SIS model, whereby recovered individuals after infection simply return to the susceptible class: $I \rightarrow S$, also with probability μ. Again, in the approaching of well-mixed populations, the model is described by the following system:

$$\dot{S}(t) = -\lambda S(t)I(t) + \mu I(t) \tag{1.14}$$
$$\dot{I}(t) = \lambda S(t)I(t) - \mu I(t) \tag{1.15}$$

Both models reproduce different dynamics after the introduction of a little seed of infected individuals in a fully susceptible population. In particular, due to the continuous decreasing of the type S, the long term survival of the pathogen within the population is impossible in accordance with classical SIR model. Both formulations are suitable to describe the time course of short cycle diseases, for which an epidemic outbreak is fast enough to justify considering the total volume of the population N as a constant. The modification of the model by allowing new individuals associated with births (or migrations) join the system (and others to leave it) can revert the extinction of susceptible individuals in the SIR case, enabling more naturally the survival of the pathogen within a population even if immunity after infection is attained.

SIS model, though, does not need demographic dynamics to allow asymptotical, stable endemic levels of infection in a population 8see figure 1.4). In this case, the equilibrium between contagion and recovery defines a dynamical equilibrium characterized by the presence of an endemic infection prevalence in the population, that is stable under certain circumstances.

Taking that into account, the metrics reflecting the severity of an epidemic episode according to each model are different. In the case of SIR model, this is the fraction of individuals that have been affected by the disease before the natural extinction of the outbreak (i.e. the number of individuals in class R at time $t = \infty$: R_∞); while, for the SIS case, this is the endemic level of infection prevalence I^*.

So, if, as we say, the above mentioned magnitudes –R_∞ for SIR and I^* for SIS–, constitute the more elementary measures of the severity of a finished epidemic event, one obvious question arises: what will be their values as a function of the model parameters? Or, relatedly, what are the conditions after which, the appearance of an infectious agents will spread yielding the occurrence of epidemic events with R_∞, or I^* greater than zero? In the context of these simple models, the last question can be answered by studying the response of a fully susceptible system ($S(t = 0) = N$), in which we introduce an *infinitesimal* amount of infected individuals $I \rightarrow 0$.

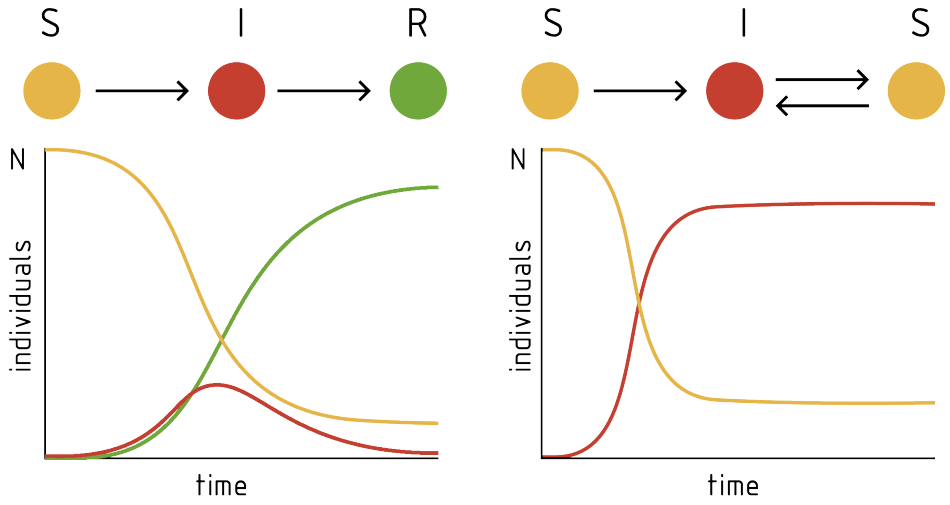


FIGURE 1.4: SIR and SIS compartmental models. Transition schemes (up) and temporal courses of epidemic episodes beyond the thresholds (low).

The question is more easy for the SIS case, because system 1.15 can be rewritten as a one-dimensional system just by substituting $S = N - I$ to get:

$$\dot{I}(t) = (\lambda(N - I(t)) - \mu)\, I(t) \tag{1.16}$$

Which has two solutions: $I = 0$ which correspond to the trivial disease-free fixed point and another, non trivial solution for the endemic equilibrium given by:

$$I^* = N - \frac{\mu}{\lambda} \tag{1.17}$$

which is only physically meaningful (i.e. greater than zero) if:

$$\lambda > \lambda_c = \frac{\mu}{N} \tag{1.18}$$

The critical value λ_c is called the *epidemic threshold*, and stands for the minimum value for the infectiousness needed for the endemic equilibrium to exist. It is easily demonstrable that the condition $\lambda > \lambda_c$ not only yields the existence of the endemic fixed point, but also its stability, as well as the instability of the disease free equilibrium $I = 0$. In other words, when $\lambda > \lambda_c$, the introduction of a perturbation from the disease free state, as a small infection seed in the system, yields the installation of the disease, and the arrival of an endemic equilibrium. In the case of SIR, it can be similarly shown that an equivalent critical infectivity $\lambda_c = \mu/N$ defining an epidemic threshold can be found, and that, in this case $\lambda > \lambda_c$ yields the emergence of transient epidemic outbreaks that, once finished, have affected a macroscopical fraction of the population, represented by $R_\infty > 0$, whereas, in the opposite case ($\lambda < \lambda_c$) R_∞ will remain of the order of the infection seed (i.e. $R_{infty} \simeq I(0) \simeq 0$).

More complex compartmental models are essentially based in SIS and SIR. These contemplate the incorporation of latency periods [48], fluxes associated to births and deaths, different kinds of disease and other more sophisticated phenomenologies. However, existence of epidemic thresholds is a common feature of all of them, and their characterization a central matter in the discipline.

1.2.2 Complex networks and mathematical epidemiology

Compartmental models used during the past century, like the basic formulations of SIR and SIS models presented in the previous section, relied on the well-mixed approximation to describe disease spreading dynamics. Even if it could seem adventured to assume that all the individual within a city are equivalent, and equally able to interact all-to-all at each time step, the approximation had remarkable success at providing a theoretical framework for the description of the time-course of real epidemic events, like the plague outbreak occurred in Bombay in 1906-1907 [61, 62, 63].

The advent of complex networks at the last years of XXth century allowed scientists revisit the works by Kermack and McKendrick with renewed eyes. Key questions for the further development of the discipline were firstly formulated at that time. What can we expect from compartmental models once we abandon the well-mixed scenario? What are the topological properties of actual contact networks through infectious disease spread? Around the answers of these, and other related questions, the foundations of modern mathematical and computational epidemiology have been settled during the last fifteen years.

sexual mating is a complex issue

More than seventy years later the works by Kermack and McKendrick, in 2001, Liljeros and collaborators analized data from a series of social surveys conducted in Sweden [11]. The participants of these surveys were asked for the number of sexual partners they had had in the last year and during their entire lives. Results were crystal clear: the distributions of the number of sexual contacts -both recent and accumulated during the whole life- were strikingly reproduced by power-laws, with slight differences in what regard exponents between men and women. This result means that the network of sexual contacts is also a complex network; more precisely an scale free one.

The relevance of this finding transcended anthropological grounds, causing a big impact in mathematical epidemiology. Sexually transmitted diseases, like the acquired immune deficiency syndrome (AIDS), spread precisely on top of networks of sexual contacts: if its topology can not be reasonably assimilated to a regular graph, classical models of mathematical epidemiology might be substantially wrong. So, the question is obvious: how do degree heterogeneities affect the dynamical predictions of classical compartmental models?

Compartmental models beyond well-mixing hypothesis: the heterogeneous mean field

As we say, using well mixed compartmental models means allowing nodes to interact all-to-all in each time step. The first way to abandon this hypothesis is to restrict the interaction range of each individuals to β neighbors, as in a random regular graph. By this way, we have, $\beta S\frac{I}{N}$ contacts between susceptible and infected individuals per unit time instead of the simple product SI that we had for the well mixed case. In this case, $\beta S\frac{I}{N}$ contacts yield circa $\lambda\beta S\frac{I}{N}$ contagions, the system of equations remains as follows:

$$\dot{S}(t) = -\lambda\beta S(t)\frac{I(t)}{N} \tag{1.19}$$

$$\dot{I}(t) = \lambda\beta S(t)\frac{I(t)}{N} - \mu I(t) \tag{1.20}$$

$$\dot{R}(t) = \mu I(t) \tag{1.21}$$

and it is easy to show that, (both for SIS and SIR), the epidemic threshold reads as:

$$\lambda_c = \frac{\mu}{\beta} \tag{1.22}$$

as now, instead well-mixing, we have just β, and, as it is logical, the less connected is the network the larger the value of the epidemic threshold. Nevertheless, the works of Liljeros and co-workers had demonstrated that connectivity patterns of real contact networks were not assimilable to those of a regular graph, which pointed to the need of abandoning classical mean field approaches in epidemic modeling [11].

As a matter of fact, things are less intuitive when the mean field approximation is discarded. In this context, heterogeneous mean field (HMF) constituted one of the first theoretical approaches to abandon classical mean field in order to consider dynamical differences between individuals [8, 58].

According HMF, individuals are distributed among different classes, each of which associated to a different connectivity. This way, all the individuals within a single degree class not just have the same number of neighbors but also are dynamically equivalent. For this reason, HMF does not fully eliminate mean field, but simply constraint it to the internal domain of each connectivity class.

under the HMF approximation, the SIR model –for example– has no longer three equations, but three times the number of degree classes considered in the system. If we consider a single degree class within which individuals have k neighbors, we call the number of susceptible, infected and recovery individuals in that group as (S_k, I_k, R_k), which will evolve in time according to the following equations:

$$\dot{S}_k(t) = -\lambda\beta S_k(t)\theta(t) \tag{1.23}$$
$$\dot{I}_k(t) = \lambda\beta S(t)\theta(t) - \mu I_k(t) \tag{1.24}$$
$$\dot{R}_k(t) = \mu I_k(t) \tag{1.25}$$

where $\theta(t)$ –that plays the same role of the fraction $I(t)/N$ in the case of regular networks (eq. 1.21)–, is the probability per link of pointing to an infected node, which is responsible of coupling the dynamical state of all connectivity classes:

$$\theta(t) = \frac{\sum_k P(k)kI_k}{\langle k\rangle} \tag{1.26}$$

In 1.26 $P(k)$ is the degree distribution of the system, that defines how individuals are divided in the different connectivity classes, and $\langle k\rangle$ the average degree of the network.

The relevance of considering network topology in compartmental models can be easily understood after deriving the epidemic threshold expression in a HMF model, which was done in a series of works by Vespignani and collaborators [8, 58] more than ten years ago. Straightforward calculations allowed the authors to find that the epidemic threshold –both for SIR or SIS models– results to be proportional to the quotient $\langle k\rangle/\langle k^2\rangle$:

$$\lambda_c = \mu\frac{\langle k\rangle}{\langle k^2\rangle} \tag{1.27}$$

On the other hand, as Liljeros and collaborators found, networks of sexual contacts presented highly heterogeneous degree distributions, best fitted by power laws of the form $P(k) \simeq k^{-\gamma}$ with γ ranging between 2 and 3 (2.54 ± 0.2 in the range $k > 4$ for females, and 2.31 ± 0.2 for males in the range k>5 [11]). Strikingly, the denominator of the epidemic threshold, $\langle k^2\rangle$, when $P(k)$ obeys a a power law distribution with $\gamma < 3$ diverges as the system size increases. This means that, the bigger an scale free like that of sexual contacts is, the closer to zero is the epidemic threshold of a disease spreading over it. The message is really disturbing: the novelly observed network heterogeneity promotes the spreading of diseases, and even the most weak pathogen will be able to cause macroscopic outbreaks if the networks on top of which it transmit are big and heterogeneous enough.

Contact patterns relevant for the spreading of airborne diseases

Strong heterogeneities observed in networks of sexual contacts motivated the appearance of a whole new family of epidemic models on heterogeneous populations which is still growing nowadays. However, other types of diseases, like airborne or zoonotic infections, do not transmit over sexual routes, and so, it would seem that considering complex networks as the most adequate substrate to use for their spreading description may not be a reasonable hypothesis. The key question is whether the contact networks

relevant for the spreading of these kind of diseases are heterogeneous, or if, instead, classical descriptions based on random graphs may account for the contact structures observed.

In what regards airborne transmitted diseases, it is commonly assumed that casual close contact among individuals can result in disease transmission, including, for example, face to face conversations, co-occurrence in confined environments (e.g. public transportation means) or, for some diseases like influenza, physical contact with surfaces previously touched by an infected individual.

The distribution of the number of this kind of social, casual contacts is not as clear as that of sexual encounters. Many aspects exerts an strong influence on the number of social contacts a person is expected to present per unit time: her culture and country, the size of her city and its population density, the number of members living in her household, her occupation etc. This makes the task of estimating degree distributions of social contacts a subtle matter, subject to a multitude of different factors whose multi-variate structure has only been partly characterized in the literature.

Although characterization of such contact distributions is unarguably an open problem, the dependence of numbers of contacts with age has emerged as a central factor exerting a robust influence on the number of contacts of individuals, at least in some countries (see figure 1.5. In 2010, John Edmunds and co-workers conducted a series of surveys in eight different european countries on this matter: the so-called Polymod project [12]. Their findings resulted in the emergence of a robust pattern of age dependence for the number of social contacts per unit time reported by individuals surveyed, essentially preserved across the different countries studied. According to this pattern, individuals tend to interact more strongly with people of the same age, except for the case of strong interactions observed between parents and sons (and, less intensely, grandparents and grandchildren). Additionally, people interacts much more intensely during their first years (i.e. until circa 20 years old), and from that point on, interactions lose frequency and intensity.

Polymod project is not the only work completed so far for characterizing contact patterns relevant for disease spreading. Until a systematic characterization of this kind of systems is available in the different countries (or at least, in the different continents), these systems provide an adequate initial basis for evaluating how emerging heterogeneity in the contact networks affects the spreading dynamics of diseases whose main transmission routes are not sexual; like influenza and, more relevant for us, tuberculosis.

1.2.3 Models for the description of coupled epidemics

In this context, an emergent field of research is the modeling of coupled spreading phenomena, as for example two pathogens or multiple strains of the same disease that propagate concurrently on the same population [64, 65, 66, 67, 68]. Focusing on the two-pathogen scenario, the complexity of the problem increases because now the natural history of one of the diseases is affected by the presence of the second one, typically as a consequence of the modification of the host's immune response after infection − with

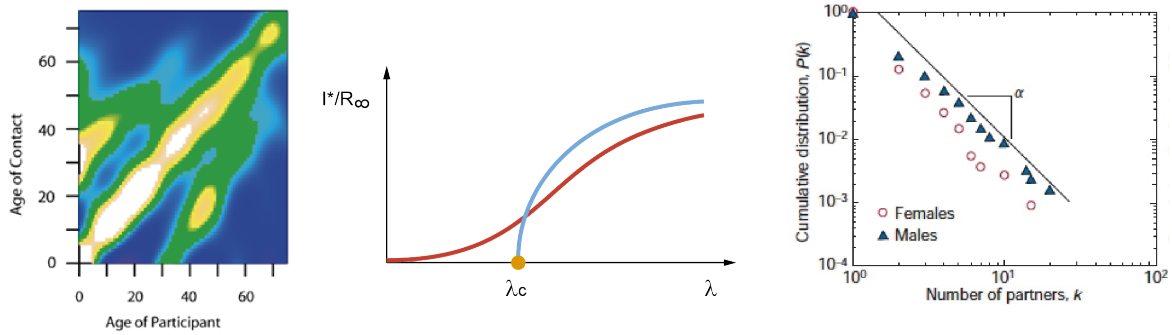


FIGURE 1.5: Heterogeneous patterns in contact networks. Left: Smoothed age-dependent patterns of physical contacts for Deutchland, as reported by Polymod project [12]. Right: distribution of the number of sexual contacts per year corresponding to the surveys performed in Sweden and analyzed by Liljeros et al. [11]. Center: epidemic thresholds in homogeneous (blue) and scale free networks at the thermodynamic limit. As network size increase, the degree fluctuations that characterize SF networks make epidemic thresholds lower.

a plethora of possibilities as given by different interaction schemes −. In addition, the networks of contacts through which the pathogens spread can vary from one disease to the other. Typical examples of these coupled spreading phenomena are given by the interaction between HIV infection and the spreading of certain opportunist pathogens [69, 70], or the strain-strain interactions of a viral pathogen like the flu virus [71].

From a theoretical point of view, one of the first works that considered the above problem from a networked point of view is due to Funk and Jensen in 2010 [66]. In this work, authors address the issue of epidemic thresholds in a SIR (Susceptible-Infected-Removed) model. However, this first work did not consider a framework in which the temporal evolution of the epidemics can be studied. Following a similar approach Mills et al. [72], motivated by the interaction between AIDS and Tuberculosis, studied two coupled SIR models where the diseases can spread in both layers at different rates. Also in this case, due to the *static* nature of the theoretical approach used, the temporal evolution of the diseases can neither be addressed. Recently, Marceau et al. presented a new model [68] aimed at studying the latter aspect, also for the SIR scheme, using "on-the-fly graphs", a network generation model previously introduced by the authors [73]. In this work, although the temporal evolution of the system can be monitored, the modeling approach does not provide information about epidemic thresholds. In addition, the model can not be trivially generalized so as to cover other classical models like SIS (Susceptible-Infected-Susceptible). Other works have considered the problem by studying the spreading of the diseases as a Markovian process. In [74] the authors, using two coupled Markov processes, studied the *competition* of two epidemics and found different regions of the parameters space where the two diseases can coexist. However, the model in [74] only represents the specific scenario of two mutually excluding diseases. It is also worth mentioning that using a similar approach, the so-called microscopic Markov-Chain [75], the specific case in which a disease and

an "awareness" diffusion process coexist has been studied in [76].

In conclusion, the problem of how the interaction two or more diseases affects their spreading processes has only been partly addressed in the literature. Two diseases can interact through very assorted different mechanisms, like modifying each other's transmission parameters, infectiousness times, mortality, latency periods etc.; and it is not clear yet what differences can be expected in the spreading processes as a function of what are the specific driving interaction mechanisms. Furthermore, the interplay between the interaction mechanisms, the particular details of the natural history of the diseases (e.g. if immunity is gained after infection or not), and the topology of the networks of contacts through which they spread (and their eventual correlations) constitute a vast landscape of possibilities that definitely deserves thorough exploration.

1.2.4 Spatial patterns of disease spreading

Compartmental models, as well as mean field approaches (homogeneous or heterogeneous) constitute useful tools for the description of the epidemic spreading on localized populations within which geographical distribution of nodes is neglected.

However, in many occasions, it is important to understand how an epidemic is expected to geographically propagate across an extended area. For answering this kind of questions, many different modeling approaches are available, from reactive-diffusive models to agent based and meta-populations approaches [77, 78].

According reactive-diffusive models, punctual amounts (or densities) of susceptible, infected and recovered individuals are substituted by fields whereby each point in space $\vec{x}$ is assigned with a density of individuals within each class, for example: $(S(\vec{x},t), I(\vec{x},t), R(\vec{x},t))$ in a SIR. Temporal evolution of these densities in a point is subjected to similar terms to those associated to disease dynamics in equations 1.21– (*reactive* terms), but geographical *diffusion* of individuals within all classes is also allowed. This kind of models are able to describe the formation of epidemic fronts and waves, and they are suitable to describe epidemic processes taking place on top of populations distributed across extended areas, in which the mobility patterns of individuals are mainly local (i.e. their movement ranges are lower than geographical distances). That is the case of many zoonotic diseases affecting wild life animals [79], but it is also the case of human populations previous to the industrial revolution. For example, reaction-diffusion models incorporating estimations about geo-demographical distributions at the time have been able to reproduce the complex epidemic front of the black death outbreak that devastated Europe from 1347 to 1351 A.C [45, 80, 81, 82, 83].

Admittedly, human mobility patterns are not any more those of the medieval Europe. The advent of mechanical transportation means after the Industrial revolution first, and after that, the globalization of the air transportation network, have exerted a deep transformation in the way we move, commerce, migrate and live. This transformation has rendered obsolete reaction-diffusion approaches to describe the way in which diseases spread through extended areas, as long-range movements, even at intercontinental scales have become routine. This has motivated the use of novel approaches to describe geographical spreading patterns in epidemic process. Among these, meta-

population models are based on networks in which nodes represent entire populations (i.e. countries, cities, or small geographical areas), and links movements of people from one population to another. The integration of air-transportation networks, census data on the geographical distribution of the populations and compartmental models within each node has allowed researchers to build up detailed meta-population models, accurate enough to reproduce and predict the time course of real epidemic episodes like the influenza $A - H1N1$, from its emergence in 2009 to its worldwide pandemic spreading a few weeks later [15, 16].

The different times needed for the black death and $A - H1N1$ influenza to get a global (or continental) reach highlights the effect of the globalization of human mobility in enhancing the spreading of diseases, and, ultimately, the need of developing accurate models and theoretical tools able to help us to understand epidemic phenomena, with the final objective of achieving preparedness and control of current infectious diseases and future possible threats [47].

1.3 Networks theory applications in cell and systems Biology

During recent years, simulations of biological systems have been spurred by the massive acquisition and availability of data in molecular and cell Biology. It is increasingly becoming evident that simulations can be paired with experiments, and in fact, they are customarily used by computational scientists to understand the quantitative behavior of many complex biological systems. Additionally, *in silico* simulations are also successfully employed in the design of new biomolecular experiments thus driving experimentalists research. Although the gap between *in vivo* and *in silico* biology has been remarkably reduced, there are still many limitations hindering the adoption of computational approaches in everyday bio-molecular research. Filling in this gap will help for a better understanding of mechanisms and operation of cellular processes.

Achieving such a goal is not easy. Experimental data for large biological systems are often incomplete and of non-uniform quality, so that their modeling is often hampered by the lack of complete knowledge of the cellular circuitry of interactions. However, our still limited capability to produce accurate computational models of living systems is however producing simulation tools useful in drawing new principles and laws, both from the topology and the dynamics of the system under consideration, that complement the huge body of experimental work.

In particular, once again, the analysis of cellular interactomes at the light of networks theory has resulted into relevant advances in the field. Systems like metabolic networks of reactions, protein-protein interactomes and gene regulatory systems have been successfully described as networks sharing the same set of ubiquitous complexity footprints: small-world property, high degree heterogeneity, high modularity and exotic motifs profiles, among others. Viewing these systems as complex networks has revealed as a powerful approach that allows elucidating the roles of its components and their dynamical interplay, in order to understand the functioning of the system as a whole.

1.3.1 A classification of biological networks

As it is commonly noticed in the literature, gene regulation is a complex process involving different phases and biochemical phenomenologies [84, 85] that allow the cell to adapt her composition and performance to the characteristics of a changing environment.

Genetic information stored in DNA is first copied into messenger-RNA molecules through a process named transcription. During this process, an enzyme called RNA-polymerase recognize the starting point of a given gene and reads it, gliding over it and generating during the process a transcribed copy of the gene. Then, mRNA diffuses to ribosomes (crossing the nuclear membrane in an eukaryotic cell, or just through simple diffusion, if we talk about prokaryotes) where they are *translated* into proteins, which are the conceptual effectors of the information stored in DNA.

This process is regulated at different stages. First, the affinity of RNAp enzyme for the transcription initiation sites can be altered in the cell, by the presence of transcriptional factors (TFs): regulatory proteins which bind on specific DNA regions, next to the transcription initiation sites of regulated genes, augmenting or diminishing their affinity to RNAp, which makes these target genes be activated or repressed, respectively. In bacteria, another family of transcriptional regulators called sigma factors bind to the RNAp instead -sigma factor itself are considered RNAp subunits–, equally affecting its affinity for different genes. Each sigma factor makes RNAp more prune to transcribe a family of genes, and less prune to transcribe others, being the number of genes regulated, in this case, typically higher than those affected by a a single TF.

Admittedly transcription is not the only phase of biological information processing that admit regulatory mechanisms. Non coding RNA are molecules that are transcribed but not translated into proteins, some of which, like micro-RNA have regulatory function. miRNA are small molecules that bind to the mRNAs of the genes that they regulate. By binding a mRNA, the regulator can inhibit its translation, favoring its degradation [86].

Additionally to these transcriptional and translational regulatory mechanisms, enzymatic proteins are activated and deactivated as a function of the specific needs of the cell. This is typically achieved through a series of post-translational modifications through which the cell is able to perceive an external signal and modify her behavior accordingly. Typically, a sensor protein interacts with a particular environmental signal (i.e. the presence of an small molecule -or ligand-, a nutrient substrate, or a physical condition), which modifies its conformation. Subsequently, the modification in the sensor protein is transmitted over a signalling cascade to the particular proteins that are intended to receive the signal (i.e. the effector proteins), whose differential performance will drive the cellular response to the initial stimulus. Signalling cascades are deeper in eukaryotic systems, while consisting in a few steps in prokaryotic cells.

Finally, a crucial part of the total amount of proteins present in the cell are devoted to enzimatic purpose designed to control the global pipeline of matter and energy constituted by cellular metabolism. The control, homeostasis and environmental adap-

tation of the metabolic networks of chemical reactions in the cell is the final target of its regulatory machinery, and so, it will not be surprising that the anomalous presence or absence of certain key metabolites constitute common signals to which the cells responds.

In the left panel of figure 1.6 the whole process is schematized through a simple example. In this figure, the metabolic network is as simple as a set of three reactions and three metabolites: A,B and C. A would represent, for example, a substrate that can be processed through two different pathways: depending of whether B is more abundant to C or viceversa. A chemical equilibrium between B and C is observed.

In the figure, an active regulatory mechanism is represented through which the cell controls the route for processing of substrate A. On the one hand, in presence of metabolite B, the cell expresses the genes activating the processing of substrate A through the pathway $A + B \rightarrow AB$, and represses the genes needed for the complementary pathway, associated to the reaction $A + C \rightarrow AC$. This is achieved by the activation of the transcriptional factor I through its interaction with metabolite A, that works as a ligand. Once activated, this transcription factor activates the genes needed for final production of proteins V and VI, which forms the protein complex able to catalyze the reaction $A + B \rightarrow AB$, making it feasible. Simultaneously, transcription factor I, once activated, stimulates the transcription of mi-RNA 4, on charge of impairing the translation of protein VII. This ultimately avoid the association of protein V with VII which has two effects: first, the complex $V - VII$ is not formed, and so, the reaction $A + C \rightarrow AC$ does not take place, and second, repressing VII eliminates a competitor of VI for interacting with V, which facilitates the formation of complex $V - VI$, needed for the catalysis of the desired reaction $A + B \rightarrow AB$.

On the other hand, when the cell perceives an excess of metabolite C with respect to B, it becomes desirable to activate the route $A + C \rightarrow AC$ and to inhibit the complementary pathway associated to $A + B \rightarrow AB$. To do so ligand C activates the transcriptional regulator II, which has a role virtually opposite to that of I: it activates the expression of genes 5 and 7, and represses the expression of 6. The final objective is to promote formation of protein complex $V - VII$ instead of complex $V - VI$, so as to favor the kinetics of $A + C \rightarrow AC$ in detriment of that of $A + B \rightarrow AB$.

The example, obviously, is a highly simplified model of a much more complex reality. However, it is illustrative because, besides exemplifying the overall rationale of signaling and regulatory processes in the cell, it includes the principal regulatory mechanisms through which homeostasis and environmental adaptation are achieved. This includes transcriptional regulation, translational silencing mediated by mi-RNA, protein-protein interactions, ligand-protein interactions driving signaling, and metabolism. Remarkably, many of these different mechanisms are customarily represented as networks, as we can see in the right part of figure 1.6, in which the main types of biological networks typically used to describe the cell as an entire system are represented.

First, we have the description of cellular metabolism through bipartite metabolic networks. Under such representation, there exist two different types of nodes: metabolites and reactions: a link connecting a metabolite to a reaction means the participation of the former in the latter. In this sense, link directionallity, as it has been commented

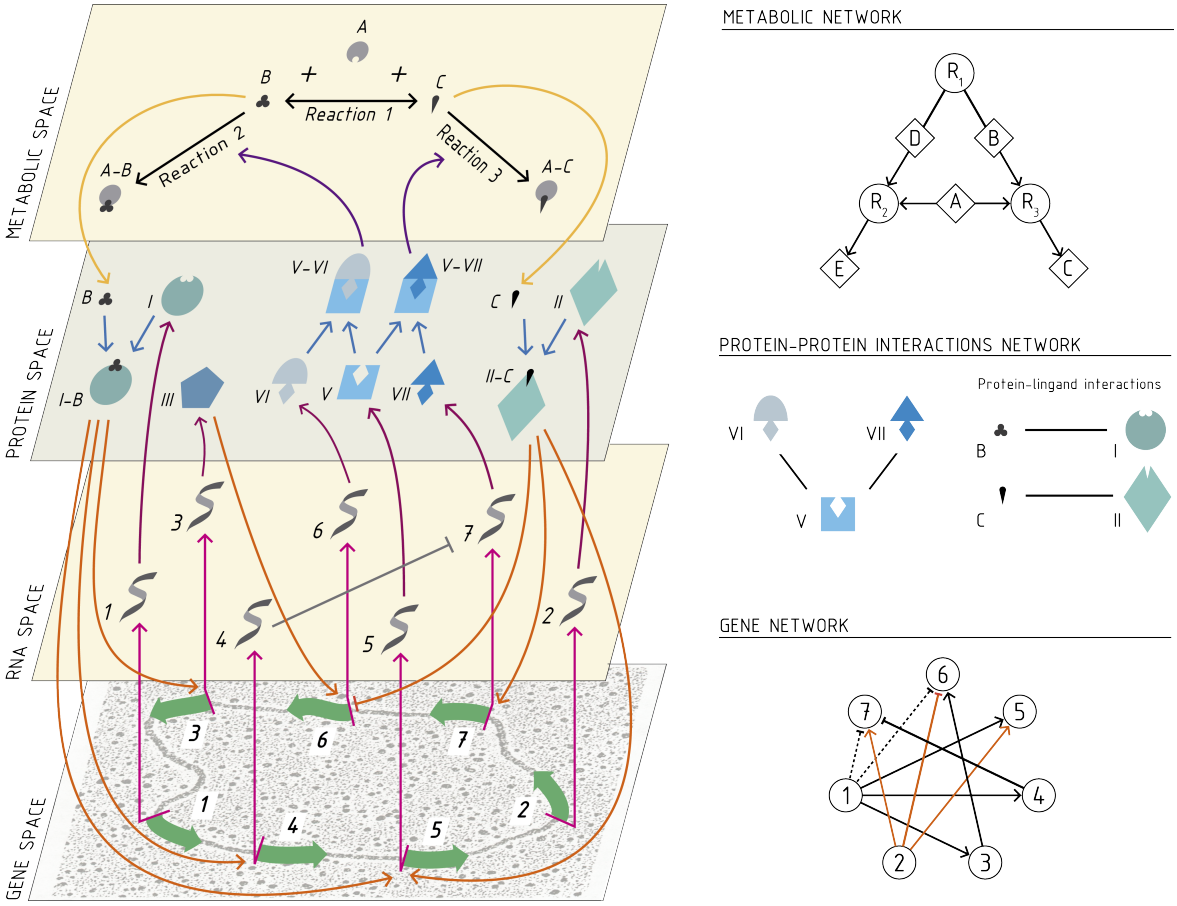


FIGURE 1.6: Left: Regulatory mechanisms involved in gene expression control. The figure represents the main biological mechanisms that are involved in the control of gene expression and protein activity. This includes transcriptional regulation (orange arrows), ncRNAs regulations (grey), protein protein interactions (blue) and enzimatic metabolic control (violet). Magenta arrows from gene space to RNA space represent transcription, and garnet arrows from RNAs to proteins; translation. Right: schematic representations of the different layers coupled in the process.

before, indicates whether the metabolite is being eliminated through a particular reaction or produced (i.e. whether it is a reactive or a product), with bi-directional links representing chemical equilibria.

Other relevant representation of cell system's Biology is the network of protein-protein interactions (PPIN), in which nodes are represented by proteins and links are physical interactions among them. Arguably the most widely experimental technique used to infer interactions between proteins are the so-called two hybrid screen systems. The idea behind the technique is very easy: we start by identifying, for example, a dimeric protein that, once dimerized, presents a particular behavior that allows us to know that it is actually dimerized. This can be a fluorescent protein that only emits light once dimerized, or a dimeric transcription factor that activates the transcription of a reporter gene whose mRNA we are able to detect. Then, we scrutinize the structure of both monomers that form the dimer, and identify the binding domains that allow them to attach and form the dimer. Finally, the DNA sequences corresponding to those domains are deleted in each monomer and substituted by the sequence codifying the two proteins whose interaction we want to address. The result is a couple of modified genes, each of which codifies monomeric "hybrids", that we introduce in a model organism (typically yeast, or easily growing bacteria) in such a way that guarantees that both genes are going to be strongly expressed. In case the proteins we are testing interact, the complex between both hybrids will form, and, as a consequence, the signal that the original protein complex formation caused (i.e. fluorescence or reporter gene transcription) will be observed; otherwise, the dimer doesn't form and no signal is registered.

PPIN topologies have been demonstrated to store useful information about the biological function of the different proteins in a cell. As it is well known, proteins that participate in common processes and interactions tend to be related to similar biological functions. This is of particular usefulness for the development of methods of protein function prediction, especially in organisms for which, as it is the case of *Mycobacterium tuberculosis* (*MTB*), the function of a big fraction of proteins remains unknown.

Finally, in the right, bottom diagram of figure 1.6, the gene regulatory network that would be inferred from the system represented at the left panel of the figure is sketched. Such a gene regulatory network represent the causal flux of events that mediates between regulators activation and gene expression profile shifts in regulated genes. Taking into account that these networks are commonly inferred on the basis of identifying genes causing expression levels modifications at the level of mRNA in other genes, they may contain links that represent different kinds of interactions (e.g. the link $4 - 7$ represents the action of a miRNA, responsible of degradation of mRNA of gene 7, while the link $1 - 5$ represents a transcriptional activation), or even links representing spurious interactions related to second order indirect effects (marked with dashed lines in the figure). When only transcriptional links are contemplated in the network (i.e. links from a transcription, or sigma factor to any target gene), we talk about a transcriptional regulatory network.

Gene regulatory networks constitute a comprehensive representation of cell dynam-

ics at a systemic level. For that reason, their systematic characterization, and coupling with signaling processes has turned into a subject of high scientific interest [87]. Admittedly, as we see in the figure, none of the types of biological networks considered is more than a partial representation of the whole dynamical function of the entire cell. For the achievement of reliable in-silico simulations of the whole cell to be possible, all the pieces of the puzzle will have to be fitted; a scientific goal that researchers of systems and synthetic Biology have just started dreaming at.

1.3.2 Gene regulatory networks inference

Focusing on transcriptional regulatory networks, classical approaches to the problem of network inference have consisted in making single perturbations to a network element (for example, a transcription factor knockout [88]) and registering the differential effects associated to such perturbation. By combining many of these perturbations a network can be reconstructed, [89] whose degree of completeness will depend on the number of perturbations that were performed. Beside the problem of completeness, the experimental challenges underlying the systemic characterization of the TRNs are far from being solved. The quantity and quality of available data on genome-wide transcriptional regulation are significant only for a small set of model organisms. Besides scarcity, the usual problem is related to the heterogeneous quality of the experimental evidences of the regulatory interactions, the building blocks of TRNs, to the difficulty associated to distinguish between correlation and causality to discriminate actual from indirect interactions, and finally, to the fact that the existence of many of these interactions is subject to particular environmental conditions whose characterization is yet very partial.

Despite these problems, the amount of high-quality experimental information about transcriptional regulation at systemic level is growing each day, not only within the context of model prokaryotes

1.3.3 Topological analysis of biological networks

Topological features of TRNs are customarily characterized at all scales using different metrics. At the large scale, genome-wide TRNs are signed and directed networks which present the following features: (i) regulatory proteins –origin of the regulatory interactions of the whole system– represent a small fraction of the total number of nodes; (ii) out-going connectivity patterns are very heterogeneous –a small percentage of global regulators (hubs) send most of the links; and (iii) in-coming link distributions are quite compact: there is a characteristic scale that defines the typical number of regulations each protein receives [90].

Turning to the mesoscale, modularity appears also in TRNs as a key feature to understand the dynamical function of the system. In genome-wide TRNs, each regulator defines its own regulon as the set of nodes directly or indirectly regulated by it. Regulons are then subnetworks, that can be sometimes hierarchically organized; in other

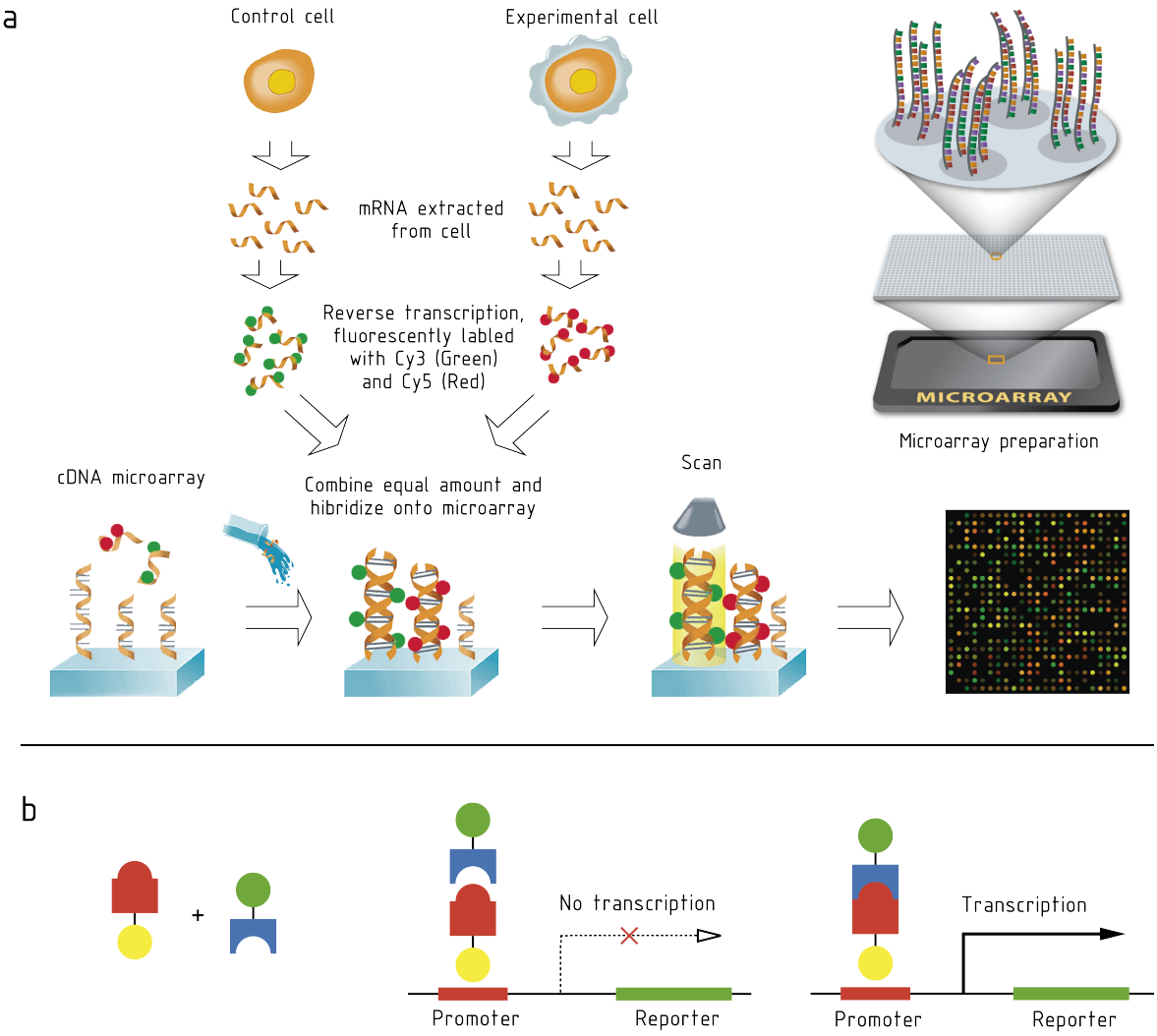


FIGURE 1.7: Experimental techniques used in biological networks inference. A) Micro-array experiment scheme for measuring gene expression profiles. A micro-array consists on a physical matrix formed by spots in each of which single-stranded nucleic acid segments corresponding to a specific gene of an organism is attached (typically, to each spot corresponds one gene). Once the array is available, an amount of cells whose gene expression is going to be measured is lysed, along with an equivalent amount of control cells. Then, the mRNA corresponding to each type of cells is marked with a fluorescent label of a different color and dropped on the array, where mRNA hybridizes with the array's strands. From the color adopted by each spot, the ratio between the expression levels in the experimental and control cells is estimated. For example, by comparing in a microarray wild type cells (control) to mutants in which a specific transcription factor has been deleted (experimental), genes regulated by the deleted TF can be inferred in a first approximation. B) Two hybrid assay. In this kind of experiment, the starting point is a dimeric protein that -for example- is able to activate the transcription of a reporter gene. Then, the binding domains that the monomers used to conjoin are substituted by the sequences of the two proteins whose interaction is going to be tested (in red and blue in the figure). In case the tested proteins interact, the reporter is transcribed.

occasions, regulons partially overlap in non-trivial ways. Thus, the identification of groups of regulons –or parts of them– interconnected through atypical, dense patterns is expected to store information about the biological role of the proteins within them [91]. The underlying idea is that community structure in biological networks might contribute to unveil functional modularity, as it has been recurrently demonstrated in metabolic and protein-protein interaction networks.

Chapter 2

Biological portrait of tuberculosis infection

Tuberculosis is an ancient disease whose presence in human populations is documented from the Neolithic period, although its actual origin is thought to have happened much earlier, circa 70.000 years ago [92]. Across the History of mankind, its principal causative agent –the bacillus *Mycobacterium tuberculosis*, discovered by Robert Koch in 1882– has constituted one of the deadliest pathogenic agents ever. However, in the last decades, the increasing availability of antibiotics and the development of global strategies for diagnosis and treatment optimization have made the disease burden to decay worldwide [93]; even allowing the public health community to dream at the disease eradication within the present century [94].

Nevertheless, many tragical problems remain open in tuberculosis epidemiology. That is the case of the increasing emergence of multi drug and extreme drug resistant strains [95], or the enhancement of the disease burden that appears in certain areas as a consequence of the syndemics between tuberculosis and VIH [412, 96, 19]. For these and other reasons, for tuberculosis eradication to be possible, novel epidemiological interventions will have to be implemented, and relevant decisions about how to do it will have to be made in the following years [94].

Surely, the most relevant tool the public health authorities count with to fight tuberculosis in the next decades is the development of a new vaccine [97] aimed either at substituting current BCG vaccine, or at least at enhancing its immunogenic action, almost null in adult individuals [93].

In fact, nowadays more than fifteen novel vaccine candidates are being developed by different research teams, and it is expected that, at least some of them will have finished the complex process of development of the new drugs within the next decade. By the end of this period, clinical trials will be finished and evaluated, the immunogenic efficacy of the new vaccines will be finally known and their respective impacts will be forecasted and compared [429, 430]. Unraveling the biochemical mechanisms behind *M.tb* infection is crucial for the development of new drugs and vaccines aimed

at eradicating the disease.

2.1 *Mycobacterium tuberculosis*: the captain of all these men of death

Human tuberculosis is caused by the bacillus *Mycobacterium tuberculosis* (*MTB*), a human parasite unable to infect animals or to survive for prolonged periods outside its host. *MTB* is characterized by a slow growth *in vitro*, airborne transmission between humans and a life cycle characterized by causing long, asymptomatic latency periods of infection in its host followed by eventual reactivation of bacterial proliferation causing disease; alternated with cases of prompt disease progression after exposure. Such a sophisticated life cycle is the product of a complex host-pathogen interaction process. In the following lines, we will review the principal characteristics of the pathogen and its parasitic life-style.

2.1.1 The life cycle of *Mycobacterium tuberculosis*

As we say, the infection cycle of tuberculosis is remarkably complex, and its final outcome depends on a series of events and interactions between pathogen and host's immune system that take place over a vast range of temporal scales.

MTB infection starts when bacilli are inhaled as droplet nuclei. After inhalation, viable bacteria reach the lungs, and they are phagocytosed by dendritic cells and macrophages in lung alveoli [100, 101]. Through phagocytosis, immune cells engulf pathogenic bacteria in a closed compartment called phagosome, which is separated from the cytoplasm by a lipid bi-layer, where the immune cell will try to isolate and kill the intruder.

The role of DCs is crucial at these initial stages of the infection process, and consists of two phases: first, as we say, DCs work as phagocytic units that engulf and kill small amounts of bacteria. Then, after phagocytosis, take the main antigens from the remains of the bacterium killed and place them into their membranes, so becoming antigen-presenting cells. At this point, they migrate from the lungs to the draining lymph nodes, where they present antigens to awaiting lymphocytes, mainly T helper 1 cells (THC1), and activate them by the production of signaling interleukin (IL)-12 and IL-18; which yield TH1 activation after which they will accumulate in lungs and orchestrate the acquired immune response against *MTB*. This process has been demonstrated to last several weeks, much more time than for other diseases like influenza.

The performance of DC as TH1 activators and the ultimate role of these as primary drivers of the acquired immune response at the infection site is crucial for the containment of infection. During the first weeks after infection, before acquired immune response is deployed, naive phagocytic cells, like macrophages, monocytes, and neutrophils constitute the main *shock troops* against the pathogen [102, 101], although their sole performance is hardly able to contain bacterial growth. At this phase, bacilli reproduce quickly, serially infecting and killing more and more macrophages. After en-

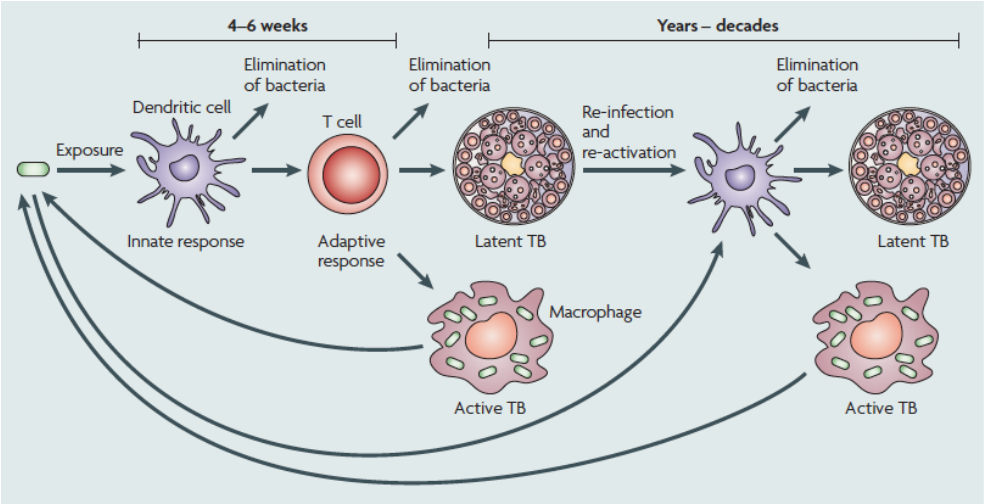


FIGURE 2.1: Infection cycle of *MTB*. Viable bacteria are inhaled, and then, phagocyted by macrophages and dendritic cells, which activate T-cells as a fundamental step of the acquired immune response. Depending the adequacy of the immune response to the infection, bacteria can be totally cleared, contained within granuloma (latency) or proliferate causing active disease. In turn, from latency, even decades after infection, bacterial growth can be reactivated, generating endogenous re-activation of disease. When the focus of infection in active disease is located in the lungs, bacteria reach the sputum and make the host infectious, so restarting the cycle. Adapted from [2].

counter with bacteria, alveolar macrophages invade the subtending epithelial layer and secrete chemokines and cytokines that, in turn, yield the recruitment and activation of inflammatory cells from neighbouring blood vessels. As more cells of these and another types are recruited, they progressively become infected by an increasing population of bacteria that tends to expand.

This process of bacterial containment around primary infection foci has different consequences at disparate levels. At the organismic level, the patient may suffer transient disease symptoms caused by the local inflammatory process, including fever and an atypical skin rash termed *erytrhema nodosum*. At a tissular level, the successive recruitment of phagocytic and inflammatory cells results in the formation of *granuloma*, which are a particular type of lung lesions characteristic of the disease consisting of concentric layers of different types of immune cells trying to contain *MTB* growth; primarily infected macrophages, epitheliod cells, foamy Mfs, and multinucleated giant cells surrounded by a mantle of activated T lymphocytes [103, 104, 105, 106].

Granulomatous containment of bacterial proliferation results in different outcomes, mostly depending on the delay and adequacy of the acquired immune response, deployed by T-lymphocytes activated by DCs. In case this acquired immune response is strong, T-cells will have success in totally clearing the infectious bacteria. This will be achieved through the reproduction of interferon gamma (IFNγ) and other relevant cytokines like tumor necrosis factor alpha (TNFα). The presence of these cytokines promotes a process of macrophage activation, according to which the phagosomal environment of activated macrophages becomes much more hostile to *MTB*, eventually making impossible its survival and further reproduction [107, 108].

In case the acquired immune response is not the adequate, bacterial containment within granuloma become unlikely. This cause bacilli to grow, and to cause extended cavities in host's lungs where bacilli proliferate extra-cellularly causing *pulmonary tuberculosis*; a medical condition that causes death in a high fraction of individuals if not treated with antibiotics. In certain cases, bacteria can invade further host organs, like his nervous system, digestive tract, bones or kidneys, where it will grow instead in lungs; which is termed *extra-pulmonary tuberculosis* [109]. In some cases, it can even invade simultaneously a number of different organs, where it proliferates under no control, causing a rapid damage in host's tissues. This last medical condition is termed *milliar tuberculosis*, and, even under antibiotic treatment, presents a highly uncertain medical prognostic. Individuals not fully immunocompetent (like newborns or immuno-depressed people) are more prune to develop this form of the disease upon infection, which is, paradoxically, associated to much lower transmission rates, as in this case, bacteria is not dispersed in the air. Whatever the type of disease developed, if this takes place right after infection (i.e. within a period lower than a year), we talk about TB *primo-infection*.

Last, but not least, the most common infection outcome is neither fully bacterial clearance nor prompt progression to disease. In circa 90% of cases, granuloma get closed by the successive recruitment of immune cells, and bacteria get contained within it, but not totally killed, entering in a dormant state (or a regime of limited growth) about whose specific nature much has been discussed [110]. In this state, pathogenic load is

low and sustained, as the granuloma interior is highly hypoxic and poor in nutrients needed for the bacteria to grow. Additionally, antibiotics show a reduced efficacy to kill dormant *MTB*.

Dormant state of bacteria within granuloma translates into an asymptomatic state in the host, who neither experiment any clinical symptom, nor is infectious. Then, a certain time after infection, spanning from a few months to several decades, the equilibrium between host and pathogen can be broken, sometimes after an episode of immuno-supression, (for example due to infection with HIV or cancer), and sometimes without an apparent reason, in individuals fully immunocompetent. If infection reactivation happens, the bacteria may break its confinement and reproduce outside granuloma causing disease; if bacterial growth take place in the lungs, bacteria will ultimately reach the upper respiratory tract and will appear in the sputum, making its host infectious and restarting the cycle.

Probably the most striking feature of the life cycle of *MTB* is the extremely long average latency period, which has been estimated to be longer than five centuries (inverse of latency times ranging from circa $0.7{\cdot}10^{-3}$ years [22] to $2{\cdot}10^{-3}$) [60]. This huge time scale means that, once granuloma are sealed, most infected individuals will not have any relapse episode in their entire lives, but will have success in containing bacteria within them for the rest of their lives. It is worth remarking that granuloma structure is very heterogeneous at distinct phases of its evolution. Strikingly, different types of granuloma can be identified simultaneously in a single host, evidencing a complex interplay between organismic mechanisms of immune response and local differences in the process of host-pathogen interactions (or cross talk) at the level of single granuloma, or even at lower scales.

2.1.2 Phagosomal environment

As we say, *MTB* is an intracellular parasite which finds in the phagosomal environment, under certain circumstances, an adequate medium to reproduce. For phagosomal survival and reproduction to be achieved, the pathogen must avoid the disparate means by which macrophages modifies phagosomal environment with the purpose of killing bacteria. This process is termed phagosome maturation, and mainly includes three different mechanisms: acidification, release of oxygen-reactive species and fusion with the lysosome. *MTB* has the ability of, at least in a first instance, inhibit it.

On the one hand, *MTB*, Ph levels in inmature containing phagosomes have been estimated to be 6.4 [111]. Although this is more acidic than extracellular media, mature phagosomes reach much lower Ph levels between 4.5 and 5.0 [112], which have highly deleterious effects for pathogen survival. Additionally, through phagosome maturation, the inducible nitric oxide synthase (iNOS) enzyme generates NO, which is a strong source of oxydative stress that, associated to acid Ph, has a strong bactericidal activity [113]. Finally, the process of phagosome maturation is complete by the phagosome-lysosome fusion, which releases lysosome contents into phagosome medium: that consists of several enzymes with strong hydrolase activities, causing degradation of *MTB* proteins, lipids and nucleic acids. Both iNOS activity and phagolysosome fusion

are inhibited by *MTB* in naive macrophages, a situation that is reversed (or mitigated) upon macrophage activation. $IFN\gamma$ mediated active macrophages are more hostile to mycobacteria, and phagosomes of active macrophages have been found to be more acidic and to present stronger iNOS and hydrolitic activities [114, 115, 116].

Beside phagosome maturation arrest, one of the most relevant *tasks* that *MTB* accomplishes upon its entrance on human macrophages is a global metabolic adaptation to the phagosome medium. This includes a global shift from a carbo-hidrates based metabolism to an state in which the pathogen takes advantage of host's lipids, and uses them as a main source of carbon and energy. This includes utilization of fatty acids and host's cholesterol [117], both for obtaining energy and for synthesizing the complex lipids of the bacterial cell wall as well as other storage lipids.

This observation has been revealed by the different composition of the lipid fraction of the cell wall that is observed after intracellular infection with respect to *in vitro* cultures [118, 119]. Additionally, mutants lacking certain genes needed for lipid metabolism show reduced virulence [120, 121, 122].

In this context, the role of the mycobacterial protein isocitrate lyase (icl1) has been revealed to be necessary for bacterial viability in activated macrophages. Icl1 is the key enzyme of the glyoxilate cycle, which is needed for *MTB* to retain carbon when growing relying on fatty acids as a carbon source. But beyond that question, probably the most relevant issue that *MTB* is committed to solve when adapting to a lipid-based metabolism is the accumulation in its cytoplasm of propionil-coA, a subproduct of the metabolism of certain host's lipids (like methyl-branched or odd-chain-length fatty acids as well as cholesterol) which is highly toxic for the bacterium unless it is metabolized [123]. Remarkably, icl1 plays also a relevant role in propionil-coA detoxification, through its participation in 2-methylcitrate acid cycle, that yields propionil-coA elimination [124].

Another relevant metabolic pathways used by the bacterium to reduce the pool of cytoplasmic propionil-coA is its usage in anabolic routes to build up complex poliketydes and methyl-branched fatty acids related to virulence. According to this mechanism, proposed by Singh [119], the bacterium is able to respond to two concurrent challenges by one single coupled mechanism, i.e. the need of synthesizing virulence lipids that allow it to arrest phagosomal maturation and evade immune response and the need to restore redox unbalance by the accumulation of propionil-coA, which is a consequence of the usage of host's lipids as energy source. Some of the best characterized cell wall complex lipids whose synthesis is coupled to propionil-coA homeostasis are phthiocerol dimycocerosates (PDIM) [125], sulfolipids (SL), and acyltrehalose-derived lipids.

Besides re-directing cell metabolism from carbohydrates to lipids, and solve the redox homeostatic complications (e.g. propionil-coA homeostasis) that are derived from that,and from direct host's attacks; *MTB* is also able to orchestrate pathogenic *counter-attack* mechanisms whereby phagocytic host cells are damaged, and granuloma are modified by the pathogen to favor bacterial growth and transmission.

Secreted proteins constitute a relevant pathogenic resource for the bacteria. The genome of *MTB* contains diverse secretion systems formed by sets of membrane proteins whose biological function consists of recognizing specific proteins and secreting them

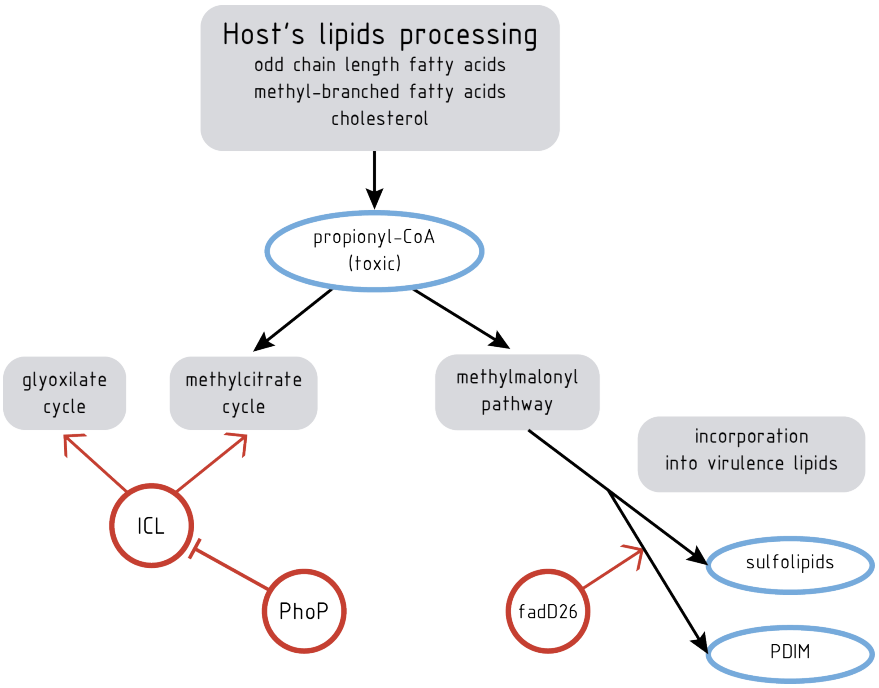


FIGURE 2.2: Propionyl-coA detoxification in *MTB*. Metabolism of certain host lipids, like methyl-branched fatty acids or cholesterol, generate propionyl-coA, which is highly toxic for the bacterium, and it is detoxified through two different routes: the 2-methylcitrate cycle, and the methylmalonyl pathway. Furthermore, the bacterium is able to *re-cycle* the products of methylmalonyl pathway and incorporate them into virulence lipids, such as sulfolipids or phthiocerol dimycocerosates (PDIM). The bacterial proteins *icl*,*PhoP* and *fadD26* regulate these processes at different stages, as well as other relevant pathways, like the glyoxilate cycle, needed for retaining carbon when lipids are its main source.

to the extracellular medium (e.g., the host's phagosome), allowing their traffic through the highly hydrophobic cell wall which is a cardinal property of all mycobacteria. One of the most relevant secretory systems in *MTB* is ESX-1; a type VII secretory system [126] that presents as its principal function the secretion of the dimeric complex formed by two small proteins: the early secreted antigenic target of 6kDa (ESAT6) and the culture filtrate proten of 10 kDa (cfp10) [245]. These two small proteins form dimers in the bacterial cytoplasm, which allows their correct secretion through the ESX1 secretory system [245]. Once secreted, cfp10 –whose principal role is thought to be related to ESAT-6 stabilization– splits, and ESAT-6 develops a strong cytolitic activity, opening pores in host's lipidic membranes [246]. As an example of its central role in immunogenicity, one of the principal features of the vaccine candidate *MTB*VAC is that, through disruption of the global regulator PhoP, the proteins of the secretion system of ESAT-6 become insufficiently expressed, which turns into a *MTB* strain that, yet expressing ESAT-6 (which provides immunogenicity), is not able to secrete it into the medium (with reduces its pathogenicity) [129].

Secretion of extra-cellular compounds is a double-edged weapon for pathogens though. While secreted proteins, like ESAT6, can be indispensable for deploying certain pathogenic mechanisms, these are often recognized as major antigens by immune cells, enhancing host's immunogenicity capabilities. Remarkably, ESAT6 itself is not only involved in pathogenesis, but also constitutes one of the most relevant antigens used by the host to recognize *MTB* and deploy adequate acquired immune response to the infection [244]. That is also the case, for example, of the major antigen complex Ag85, formed by three small proteins (30-32 kDa) whose introduction in mice have been demonstrated to engage a strong T-cell specific immune response in mice [131], a result that has motivated the development of several new candidates for novel anti-TB vaccines that are based in these effects. Ag85 antigens, as other proteins like Rv2525 are secreted through the twin-arginine-translocation (TAT) secretory system [132]

2.1.3 Gene diversity in *Mycobacterium tuberculosis* and its host

Beyond *MTB*, *M.africanum* is a closely related species that (to a lower extent) can cause human tuberculosis as well. Both are human obligate parasites with no zoonotic reservoir and very reduced survival outside their host. Along with other bacteria causing disease in animals (and occasionally in humans), like *M.bovis* (responsible of bovine TB) and *M.microti*, the so called *vole bacillus*, they form the *Mycobacterium tuberculosis complex* (*MTBC*). In turn, *MTBC* belongs to Mycobacteria genus, that, beyond these pathogenic species contains different non-pathogenic sole saprophytes with very different lifestyles from that of an intracellular parasite like *MTB*.

Notwithstanding that, most mycobacteria share a number of characteristic features, such as complex, lipid-rich cell walls and highly conserved families of genes [134, 135]. Noteworthy enough, some of the major immunodominant antigens are conserved among mycobacteria, which causes cross reactivity [136, 137, 138], a phenomenon with crucial implications for tuberculosis vaccine development, as we will discuss later. For example, the antigenic complex *Ag85* is one of most widely conserved antigens within the genus

[139], and even ESAT-6 have found to be present in some subtypes of *M. kansasii*, *M. szulgai*, *M. marinum*, and *M. riyadhense* [140].

In what regards genetic diversity within *MTBC*, it is very reduced, with a 99.9% of identity at nucleotides level and identical 16S rRNA sequences [141]. Similarly, there is a very constant gene content and no sign of ongoing recombination among strains [142, 143]. Among the genes more conserved within *MTBC* genomes noteworthy lie epitopes related to T-cell recognition [144], genes which precisely are those that show greater variation in populations of pathogenic organisms which base their pathogenic fitness in passing unnoticed to the human immune system [145], suggesting that tuberculosis has exactly an opposite behavior. Indeed, this finding has been related to other experimental evidences pointing to T-cell specific responses having benefits for the pathogen. Mainly, it has been observed that the existence of strong T-cell mediated responses is a requisite for the formation of the characteristic cavitary lesions in host's lungs, which are the main requisite for efficient airborne transmission [146]. Relatedly, individuals with reduced immunological competence, like HIV patients and infants, present, at populations levels, more tendency to develop extra-pulmonary forms of disease, that, even if commonly associated to increased morbidity and mortality, generate more reduced numbers of secondary cases [146].

The origin of parasitic association between *MTBC* and humans is dated circa 70.000 years ago, when a common ancestor would have jumped –more likely from a saprophytic lifestyle [92]– to become one of the first documented parasites of mankind, along its close relative *M.leprae*, causing lepra [147]. From that moment on, coevolution between *MTB* and humans is the story of a *cheek to cheek* dancing. Recent phylogenetic studies conducted on large samples of *MTBC* representative strains and humans of all ethnic profiles have stablished that the evolutionary history of *MTBC* have tightly followed the demographical evolution of mankind and its geographical distribution; from its african origins to its worldwide expansion followed by the neolithic demographic transition [175]. This result has been complemented with the observation of that *MTb* strains belonging to different evolutionary clades tend to associate with hosts of ethnic profiles associated to the same region, when epidemiological studies of multi-cultural communities are performed [174].

As a relevant conclusion of these studies, the Neolithic demographical transition, circa 10.000 years ago, caused a major expansion event in *MTBC* global population. The reason, rather than greater exposure to zoonotic reservoires, or even net augments in human population –which, in other phases of History did not influence the overall *MTBC* population–, is related to the emergence, for the first time in History, of large human crowds around the first agricultural settlements, much more densely populated than previous hunter-gatherers groups; with a subsequent impact in favoring *MTBC* transmission.

This is unarguably a disturbing conclusion to reach for us, inhabitants of a hyper populated world in which the fraction of people living in urban settings outnumbered that of rural ares in the last decade for the first time in History, and continues growing.

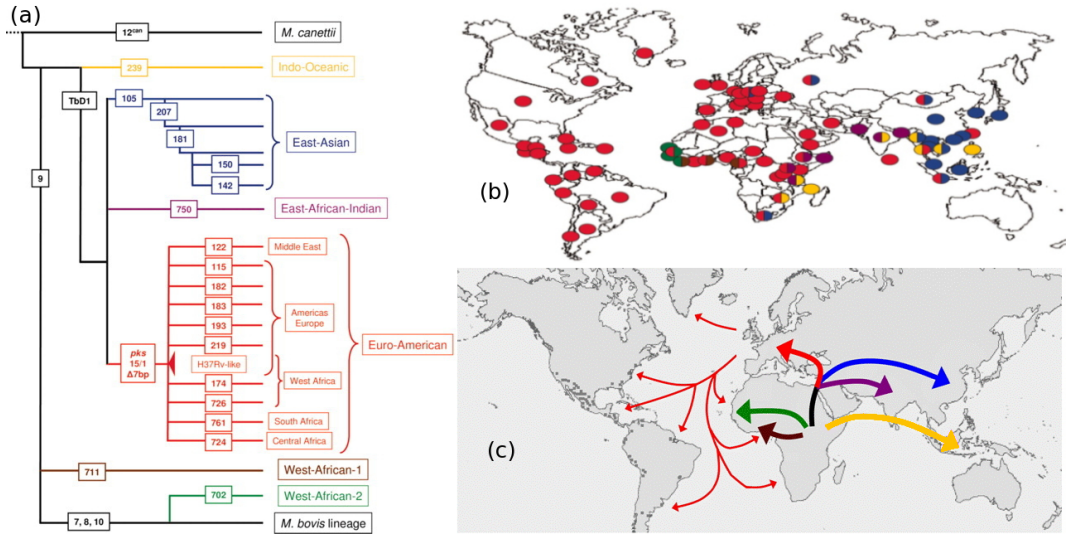


FIGURE 2.3: Host-pathogen coevolution and geographical distribution of TB. TB origin is dated circa 70.000 years ago. Indeed, from that beginning, members of *MTBC* have evolved into six main phylogenetic clades (panel a), whose geographical distribution (panel b) follows human migration patterns (panel c), from their ancient african exodus to the rest of continents (tick arrows) to the migration of europeans to America and Africa during colonialism and the industrial era (red thin arrows). (figure adapted from [175]).

2.2 Tuberculosis epidemiology: stories of the white plague

Mycobacterium tuberculosis is an extraordinarily successful pathogen that is usually considered as on the deadliest killers of the History of mankind. Indeed, TB is the disease responsible for a greater number of deaths in the past 200 centuries, with as many as one billion casualties estimated to be attributable to the so-called *white plague*. If we carry counts back two centuries more, we find that, between the seventeenth and nineteenth centuries, TB was the responsible of 1 over each 5 deaths in Europe and North America [148].

But even if historical records tell us that it has been one of the deadliest diseases ever, the popular thought extended in western countries according to which TB is a virtually eradicated disease, could hardly be more wrong. Indeed, TB is -with HIV and malaria- one of the three deadliest diseases nowadays; specially in the underdeveloped world. In the Global Tuberculosis report 2013, WHO estimates that in the precedent year there were 8.6 new million TB cases all around the world; and 1.3 million deaths caused by TB [1]. Current estimations tell us that 1 over 3 people in the world is infected with *MTB*; and the association between TB and HIV, as well as the emergence of antibiotic resistant strains constitute critical concerns for Health Authorities.

In the following lines we review the principal characteristics of TB epidemiology:

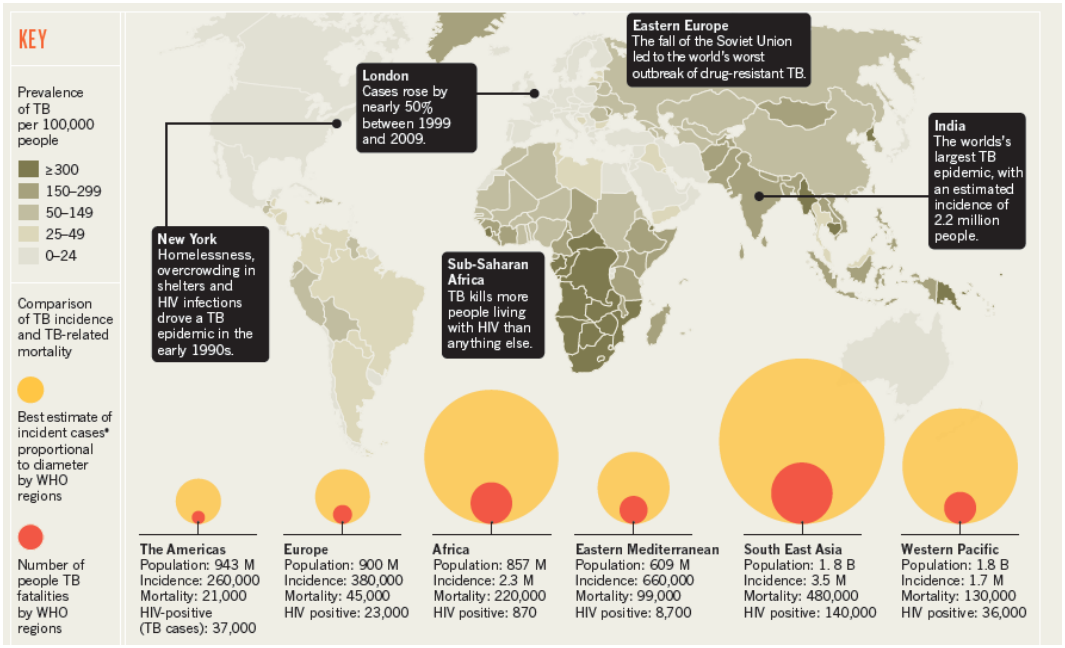


FIGURE 2.4: Geographic distribution of TB burden in 2011 (image extracted from [149]. According to Paulson, in 2011 there was 1.4 million deaths because of TB, and almost 9 million new TB cases (the 84% of which in Africa and Asia).

the issues related to its surveillance, the natural History of the disease and the state of the art of TB spreading mathematical modeling.

2.2.1 Latent TB infection: a huge hidden variable

As we have discussed, tuberculosis infection cycle is characterized by the pathogen's ability of infecting huge amounts of hosts, causing in them an asymptomatic state of latency that is frequently unnoticed. In this sense, probably the main hindrance in the study of tuberculosis epidemiology is the difficulty of addressing chronic infection with sufficient reliability.

Tuberculin skin test (TST) has been until recent the only tool at hand for assessing TB infection; and it remains so in the clinical practice in many countries in the world. TST consists in the intra-dermal injection of TB antigens, that cause a local inflammation whose size serves as a proxy of previous infection with *MTB*. However, several drawbacks make the outcome of TST not always perfectly interpretable. On the one hand the purified-protein derivative (PPD) that is injected contains several antigens, some of which are also present in other mycobacteria different from *MTB*; which makes TST prune to generate false positives. On the other hand, false negatives often take place, mainly in immunocompromised individuals and young children. Last, but not least, previous vaccination with BCG often yields a positive TST response, and, if the test is repeated more than once in the same person, the results are less reliable, as the

PPD exposure of the first test (eve if its result was negative) is able to cause a positive response if the test is repeated.

These drawbacks has been solved, at least in research grounds, by the introduction of Interferon-gamma-release assays (IGRA), which are performed on blood samples extracted from the patients. In this test, the blood sampled is exposed to ESAT-6 (or other antigens specific to *MTB*), and the release of $IFN\gamma$ by T-cells in the sample is then registered. $IFN\gamma$ production after exposure to *MTB* antigens is a highly specific sign of the existence of an acquired immune response to *MTB*, something only possible after previous exposure to the bacillus. The specificity of the antigens used guarantees a much lower rates of false positives, both caused by sensitization to environmental mycobacteria foreign to *MTB* and by BCG vaccination.

Admittedly, IGRA tests are not available in the everyday clinics in most countries with highest TB burden, when TST is yet the main tool used to assessing TB infection, a diagnosis tool that was firstly proposed as many as 124 years ago, by Robert Koch himself [150]. In addition, as we said before, *MTB* infection is strikingly general in human populations, with an estimated prevalence of circa 1/3, huge numbers which illustrate up to what point *MTB* infection is able to pass unnoticed at the level of human populations. This ultimately compromise the accuracy of the estimations about the global volume of *MTB* reservoir, and our knowledge about the Biology of latent infection [110].

Notwithstanding that, tuberculosis is nowadays a notifiable disease (i.e. one of compulsory declaration and surveillance) in most countries, which has allowed international authorities to build up an ever-increasing and more precise system for global TB surveillance that is allowing epidemiologists to identify geographical trends and caveats related to TB burden distributions worldwide. Additionally, the generalization of molecular epidemiological tools have allowed the genetic foot-printing of clinical isolates, which has supposed a great advance for backtracking epidemic outbreaks, distinguising recent transmission from relapses from latency and, definitely, allowing deeper understanding of the circulating strains and the main risk factors associated to the disease.

These and other advances have allowed nowadays epidemiologists to reach an overall knowledge of TB natural history and epidemiological characteristics of the disease more complete than ever in the History of Medicine.

2.2.2 Types of tuberculous disease

Typically, different types of TB are distinguised depending on the organs where the infection locates: pulmonary and extra-pulmonary TB. As we say, the main target of TB infection is host's lungs. Pulmonary tuberculosis is usually detected through X-ray imaging of tuberculous cavities and/or molecular culturing of sputum. Patients with positive results in sputum cultures (termed *smear positive*) are more infectious; though *smear negative* tuberculosis is associated to lower, still non negligible infectiousness.

In turn, extra-pulmonary and miliar TB are different from pulmonary TB in many aspects, being the main divergence the virtually null infectivity of TB when the infec-

tion proliferates in organ foreign the lungs, as airborne transmission is the principal contagion route for TB.

Beyond lungs, *MTB* is known to be able to colonize virtually all human organs [151], and, as a function of the infection site, disease manifestations and diagnosis proofs can be very assorted, and often requires invasive procedures [152]. This results in much higher drawbacks for reliable diagnosis of extra-pulmonary TB, even in high burden areas [153, 154, 155]. This situation often cases delay in the treatment start, which, along with the fact that depending on the infection focus much more damage can be achieved by lower pathogenic loads, make extra-pulmonary TB cause higher morbidity and mortality rates than the pulmonary forms of the disease.

2.2.3 TB burden: global trends and prospects

Since the discoveries by Robert Koch at the end of XIXth century, burden levels associated to TB have been consistently decreasing during decades in Europe and America. The reasons for this situation have been the introduction of the first anti-tuberculosis vaccine (Bacillus Celmette-Guerin, or BCG) in the decade of 1920, and, even more important, the advent of anti tuberculous antibiotics in the fifties. Additionally, the deployment of specific public health plans against the disease and the intense socio-economic improvements that western societies have progressively enjoyed during the last century contributed to sustain the decline of a disease typically associated to poverty, malnutrition and insalubrity. Nowadays, the main risk factors associated to TB consists of poor living conditions that favors disease transmission and factors associated to the inability of deploying an adequate immune response against disease, like malnutrition, diabetes, drug abuse, alcoholism and HIV infection [156]. Furthermore, young adults constitute the age segments typically more affected by the disease, mostly in what regards pulmonary TB.

The decline of the disease during the last century had the effect of generating relaxation in all grounds related to anti TB fight, ranging from public health authorities to researchers and clinicians. However, this situation reversed in the decade of 1980 due to the emergence of the new human immunodeficiency virus (HIV), after whose irruption TB burden rates turned to scale worldwide. TB re-emergence associated to HIV [157] constitute one of most worrying phenomena in current TB epidemiology, and it has motivated renewed research efforts in anti tuberculosis fight; from the development of new drugs and vaccines to the general awareness of the need to improve diagnosis protocols and techniques for reducing infectious periods and disease misreporting.

This global relapse of tuberculosis burden has not hit uniformly all countries in the world. As Lauzardo and Ashkin write [160]: "*HIV, the scourge of the late 20th century, has conspired with poor distribution of resources and poverty to fuel the fire of the TB epidemic. More people have died during the last decade from TB than perhaps any other decade in history.*" As a matter of fact, under-developed countries, now much more populated than the western world in the XIXth century, bear similar incidence rates to those that devastated Europe at that time, which constitute a dramatic public health situation that made the WHO declare tuberculosis a global emergency in

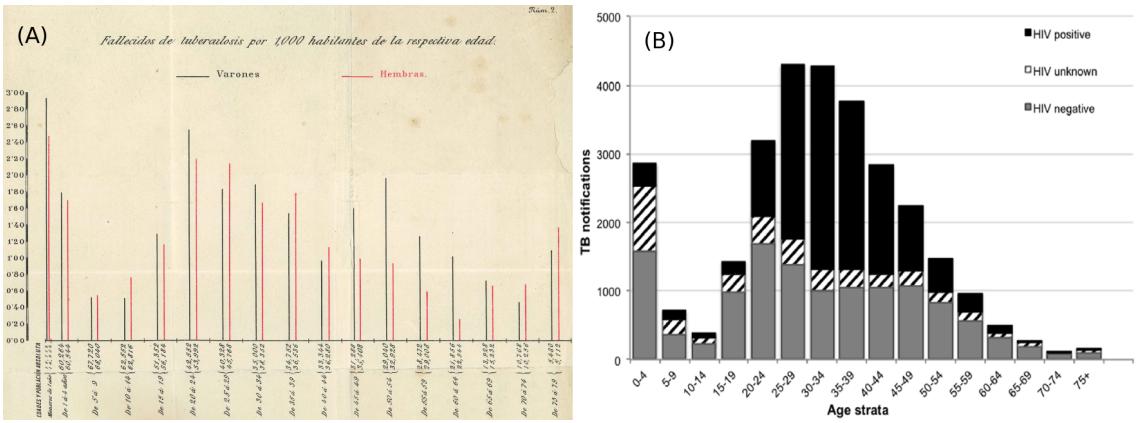


FIGURE 2.5: TB burden by ages. A: number of deaths caused by TB per 1000 inhabitants of the same age in the spanish province of Soria, from 1900 to 1907, as reported by M. Íñiguez and M. Hercilla in 1909 [158]. B: Number of TB cases notified in Cape Town, Republic of South Africa during 2009, according to [159], stratified by age and HIV status. Even if the measurements of the two graphs are not fully comparable, they constitute graphical examples of how TB burden is distributed among the different age groups: with higher affection in infants younger than five years old and young adults from 20 to 50 years old. Remarkably, the first dataset corresponds to a period in which antibiotics or vaccines were not available, and HIV did not constitute a thread for human health yet. Even if that is not the case of panel B, the age distribution of both burden measures is highly similar.

1993. Nowadays, South and east Asia constitute the main geographical focus of TB in number of cases, followed by sub-saharian Africa, where HIV-TB syndemic association constitutes the main driving effect of TB epidemics.

Eradication of TB within 2050 was settled as the initial objective of the "STOP-TB" global strategy conceived by the world health organization in 2000. This global campaign was based on the creation of an international partnership including international, nongovernmental and governmental organizations and patient groups all over the world. Their principal activities and objectives consist of the promotion of social awareness and funding for TB prophylaxis and prevention and the development of new treatments, enhanced diagnosis tools and vaccines. 14 years later the partnership was conceived, the final goal of eradicating TB at the equator of the XXIthe century seems unlikely; very specially unless new vaccines are not urgently deployed [94].

2.2.4 TB spreading models

Tuberculosis has attracted the attention of mathematical epidemiology, at least, from 1962 [161, 162, 163, 164]. The dynamics of its natural history has been the matter of an extensive and thorough bibliographical corpus within the discipline [165]. Ordinary differential equations, partial differential equations, finite difference equations, integro-differential equations and Markov chain models are mathematical methods that have been recurrently used to model TB dynamics in the literature [166].

The particularities and complexity of TB infection cycle are usually taken into account in the models mentioned, to the extent of the degree of detail of the models themselves. This way, aspects like long latency periods and primo-infection, age-dependence of the susceptibility to disease, different types of diseases, effects of treatment and vaccination and other phenomenologies are usually taken into account.

Beyond the mathematical implementation of these modeling tools, and the specific details of the natural History of the diseases being described, an important feature of all spreading models is their range of application and their data feedback. This way, on the one hand, some spreading models are built up and studied from an eminent theoretical viewpoint, with the main objective of exploring the dynamical features of the system in generic regions of the parameter spaces, like epidemic thresholds or predicted endemic levels. The development of epidemic models of this kind, from the works by Kermack and McKendrick to our days, has allowed mathematical epidemiologists to develop increasing insight of the dynamics of the spreading processes under different conditions, which has constituted an essential step towards the development of modeling tools of greater usability. Admittedly, the task of epidemic modeling does not (and must not) stop at this point, and, upon the basis of a general knowledge of the spreading processes along with the estimation of dynamical parameters through epidemiological studies, mathematical epidemiology have moved forward by substituting *parameters* by *numbers*, as a first step towards the development of operative platforms for epidemiological surveillance, forecast and control. In that phase, a thorough assessment of outcome's uncertainties and sensitivities to modeling hypothesis constitute a fundamental issue that is not always easy to accomplish. Finally, a relevant characteristic of this family of *data-driven* models is their generality and range of application, that can range to specific epidemic settings (based, for example, on high-resolution data extracted from exhaustive field studies on reduced areas during reduced periods) to the entire world over decades.

In this context, diverse models for the quantitative description of worldwide TB burden levels have been developed during the last two decades [422, 167, 22, 21], which incorporate data from the transnational TB database curated by the WHO, and nowadays publicly available [24], with the aim of providing suitable platforms for quantitative evaluation of TB burden levels worldwide. These models explicitly consider a detailed description of the Natural History of the disease, including age-dependencies of different dynamical parameters, as well as regional variations for some of these, in some cases in which homogeneous country specific information relevant for parameter estimation is available [22]. However, different conceptual issues, mostly –but not exclusively– related to the lack of appropriate data to feed them, impose limitations to these models.

First, TB spreading dynamics is usually presented uncoupled from demographical evolution, through different algorithmic procedures [422, 22] in an attempt of isolating TB intrinsic dynamics from the influence of demographic variations. The suitability of this classical hypothesis in epidemic modeling may be argued in the context of a disease whose dynamics is strongly age-dependent, like TB, due to the fact that, in such a case, disregarding demographical variations might result in the introduction of systematic

bias in the forecasts, as it has been recently suggested [168]. Additionally, contact patterns are classically disregarded in these kinds of models, in which homogeneous mixing for all the individuals within all age groups is considered. Instead, the influence of the contact patterns is known to be crucial for disease spreading, as it has also been shown in TB, where the impact of preventive and therapeutical measures foreseen by spreading models is known to strongly depend on the structure of the contact networks considered [169]. Even if these dynamical aspects have been predicted to play a relevant role in TB spreading dynamics, [168, 169], it remains pendent the quantification of their effects on overall burden rates at a global scale.

Additionally, the integration of "fine grain" phenomenologies on data-driven TB spreading models constitute a major challenge for mathematical epidemiologists. Admittedly, the dynamical description of the emergence of antibiotic resistences [170], the explicit consideration of HIV-TB co-occurring epidemics [171, 172, 169], the addressing of the influence of human mobility patterns on the spreading of the disease [173] and even the consideration of genetic heterogeneities, both of pathogens and hosts [60, 174, 175, 176, 177] have been all considered, and, for some cases, thoroughly analyzed under the light of specificly devised spreading models. Nevertheless, the context of these studies is always subject to the particular phenomenology being analyzed, the degree of detail of the spreading dynamics is lower (e.g. without age-structure) and not even a tentative of quantifying the influences of these effects on global TB burden levels is provided by most of these works. Once again, integration of all these phenomenologies around global data-driven models able to offer simultaneous forecasts about all the relevant aspects of TB epidemiology suppose the assembling of a huge scientific puzzle, which, though, could hardly be better motivated.

2.3 The fight against tuberculosis

Upon the global re-emergence of TB promoted by HIV irruption, huge investment efforts have been made in TB pharmacology, focused at the development of new drugs and vaccines against TB aimed at substituting those currently on use. The discovery and commercialization of improved, novel pharmacological tools, both for disease prevention and treatment is nowadays considered an urgent need if TB wants to be eradicated in the following decades [94]. In the following lines, we contextualize the state of current TB pharmacology and the perspectives in the field.

2.3.1 Current anti-TB antibiotics: the story of an obsolete warfare

Along with the dramatic consequences of HIV-TB interactions, the emergence of drug resistances constitutes a global concern for public health authorities [178]. TB is a particularly *stubborn* microorganism, against which regular treatments consists of the administration of a series of different antibiotics (typically rifampizin, isoniazid, pyrazinamide and ethambutol) during several months. Other anti-TB drugs complementarily

used are fluoroquinolones and several injectable drugs like kanamycin. Multi-drug resistent (MDR) [179, 180] and extensively drug resistant (XDR) [181]strains [182] are also emerging, and their spreading threatens to carry back a dramatic epidemiological situation in which TB would constitute again a hardly remediable condition. For a strain to be considered MDR, it must present resistance at least to rifampizin and isoniazide, while XDR strains show additional resistances to any fluoroquinolone, and one of the three injectable drugs, capreomycin, kanamycin, and amikacin [178]. Recently, there have been detected strains that show no susceptibility to any anti-tuberculous drug known [183].

The reason for the emergence of these resistance strains is two-fold. On the one hand, long periods of current antibiotic treatment protocols, –that typically last 6 months [184]–, make treatment adhesion highly challenging, mostly in underdeveloped areas with fragile health care systems. Treatment default is indeed a major correlate for further relapse and emergence of drug resistances, and for this reason, WHO comes developing DOTS strategy (directly observed treatment, short course), which consists in a treatment follow-up protocol for physicians that has allowed a relevant increase in both treatment adherence and curation rates worlwide [185]. Beyond the problems associated to long treatment duration and therapeutic default, the emergence of antibiotic resistances is favored by the fact that first-line treatments against the disease have remained essentially unaltered during the last 50 years; being the *youngest* among the four first-line antibiotics most widely used (rifampicin) discovered in 1963. This highlights the need of developing new anti-TB antibiotics to be administrated during shorter time courses to reduce the threaten of drug resistant strains.

2.3.2 Anti-TB vaccines: past, present and future

These are exciting days in tuberculosis research. As we talk, more than fifteen independent teams all over the world are developing as many different vaccine candidates aimed either at substituting current, deficient BCG vaccine or at enhancing its effects [186]. This situation has motivated the creation of two independent agencies for the promotion and development of new TB vaccines: Aeras TB vaccine foundation, in the US, and the european TB vaccine initiative (TBVI) [187, 188]. Similarly, a precise development protocol have been defined, according to which each vaccine candidate needs to go through a six-phases pathway of testing, from its discovery (1) to final market licensure (6), including, as intermediate steps, preclinical development (2) , Phase I (3) and Phase II/IIb (4) o proof-of concept trials and Phase III large clinical trials (5), before commercialization. In each of these phases, each vaccine candidate must fulfill a series of stringent requirements so as to pass to the next step, according to a stage-gating approach designed for optimizing limited resources [187]. Among these new candidates, there are novel vaccines of different profiles that base their action on disparate immunological principles and mechanisms; all of which holding the promise of offering better levels of protection than BCG, whose global administration for almost a century has not been enough for eradicating the disease. As Hellen McShane wrote in a recent editorial comment, our understanding of current BCG impact on reducing TB

burden levels wordwide is still limited; and a better understanding on the performance of the old vaccine constitute *the key to improve it* [189].

BCG: the willful vaccine against TB

The current TB vaccine was developed in the decade of 1920 by Albert León Charles Calmette and Camile Guérin, based on a series of repeated sub-cultures of a strain of *Mycobacterium bovis* isolated from a cow, that resulted on a live attenuated strain named BCG (bacillus Calmette-Guérin). The new strain, because of reasons that would not be fully understood until decades later, presented an avirulent phenotype when inoculated on humans, yet generating an immune response in the host that offers partial protection against further infection with *MTB*. The principal genetic treats responsible of such phenotype are the deletion of a genetic region –called RD1: region of difference 1– during the process of culturing [190], containing the genes Rv3874 and Rv3875, which codify the major *MTB* antigens ESAT-6 and cfp10, central for mycobacterial virulence [191]; combined with the production of other relevant antigens, like the *Ag85* complex, which is considered the main cause for its immunogenic effect.

BCG offers consistently protection against the disease in children, specially to the worst types of tuberculous meningitis and miliar TB. Despite constituting a great advance in the fight against TB, BCG's importance for the overall trend of TB decay observed worldwide during the XXth century has been largely argued [192, 193], as it fails at providing consistent protection to the pulmonary forms of the disease, specially in adults, with relevant variances among the studies performed to test it. Currently, BCG induced protection is estimated to last between 10 and 20 years after newborn vaccination [214, 203, 215].

Epidemiological assays aimed at determining BCG efficacy started in the decade of 1930, but it was not until the decade of 1950 that the scientific community was conscious of the high disparities found among the different studies [197]. The causes underlying the observed variability of BCG efficacy in different settings still constitute a matter of open debate [194, 195], and they include strain variation in BCG preparations [196], genetic, epi-genetic or socio-economical differences between populations, study quality, parasitic co-infections, etc [186].

However, multi-variate meta-analysis of BCG efficacy determination studies consistently determine that latitude is a variable showing a great –direct– correlation with BCG performance [197, 198, 199, 194], pointing to the existence of latitude-driven mechanisms influencing BCG performance. Among these possible mechanisms, different levels of exposure to UV light, due to its mycobacterial killing effect is commonly mentioned, although the hypothesis that arouses greater consensus points to exposure to environmental non-tuberculous mycobacteria (NTM), whose presence is greater in tropical and sub-tropical areas [200, 201, 202], as the most likely cause underlying BCG efficacy variability. According this hypothesis, greater levels of exposition to these sources of immunological sensitization prior to the moment of vaccination would cause reduced protection by the vaccine, which would explain BCG's reduced performance as trials get closer to the equator. NTM share many relevant antigens with

MTB, which causes cross-reactivity [136, 137] able to generate cell mediated immune responses that have been measured *in-vitro* [200, 202]

It is worth mentioning that, beyond NTM, BCG vaccine itself, as well as the very reservoir of latent *MTB* infection constitute relevant relevant sources of exposition to mycobacterial antigens at populations level. Although individuals infected with *MTB* or vaccinated with BCG use to be discriminated in clinical trials (at least, in some of them, with enough reliability) their implications as environmental sources of antigenic sensitization should be also considered, as we will discuss later.

Diverse epidemiological observations backup the hypothesis of sensitization with environmental mycobacteria being the source of BCG efficacy variability. First, it has been observed that BCG is effective in trials from which tuberculin-skin-test-positive individuals have been stringently excluded [194, 203]. Second, neonatal vaccination -and therefore applied on individuals that have not been exposed to mycobacteria- is successful against TB [204, 205], and BCG efficacy is lower the older the individuals are at the moment of vaccination [195]. This last observation is coherent with the fact that older individuals have been exposed during more time to the mycobacterial agents, and so, are more strongly sensitized.

Two different mechanisms have been theorized on how this exposition to mycobacteria would affect the behavior of the host to a vaccine like BCG [139]. The masking hypothesis postulates that environmental sensitization confers a significant protection against TB in such a way that a vaccine can barely offer an additional level of protection [206]. Another possible effect is that the exposition to mycobacteria may trigger a immune response capable to block the *assimilation* of the vaccine by the host. This is known as the blocking hypothesis [207].

These two effects have the potential to explain a considerable fraction of the variability observed in the trials performed, which explains both the dependence of BCG efficacy with age at the time of vaccination -as an individual gets older its exposition to mycobacteria increases- and its geographical variations. The crucial implications of these effects in TB vaccine development are twofold. On the one hand, understanding the range and causes behind the variations of BCG efficacy is essential [189], as the efficacy of any novel vaccine will be measured on counterposition to BCG. On the other hand, depending the type of vaccine, where it is applied and how old are the target populations, new vaccines could also be affected by masking or blocking effects.

Types of preventive vaccines

As we say, several research teams are currently developing new candidates for novel anti-TB vaccines aimed at substituting or complementing current BCG. Depending on the specific objective of the novel candidates, we can distinguish two main groups of vaccines, preventive vaccines –aimed at reducing risk of infection–, and therapeutic vaccines, whose objective is to reinforce antibiotic therapy during active disease, shorten the time needed for therapy completion and, in general, augment the proportion of treatment success. Along this thesis, we will primarily focus on preventive vaccines, which, in turn, are divided in two groups.

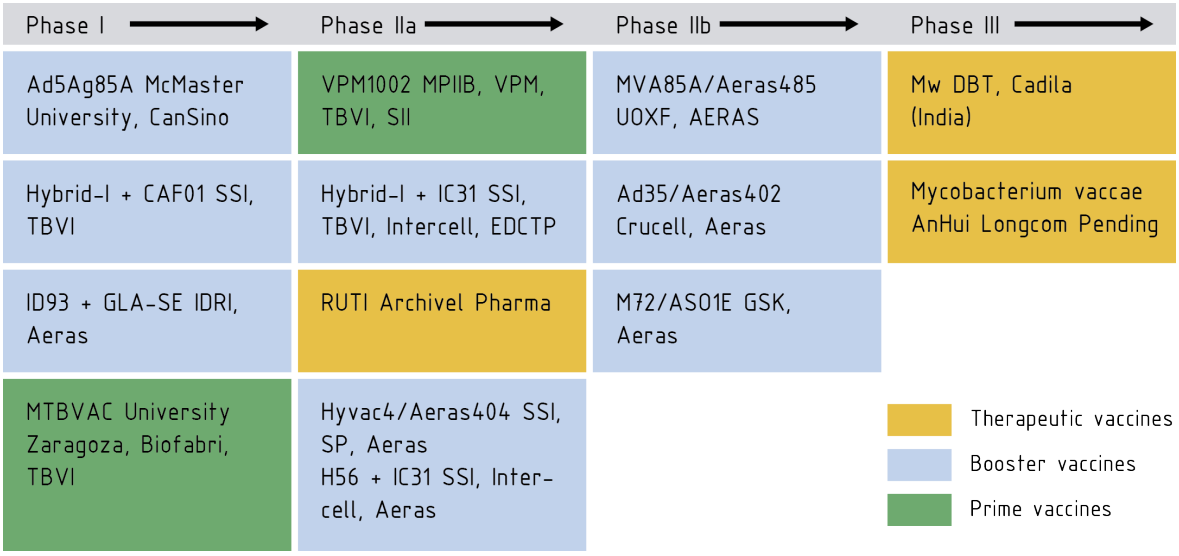


FIGURE 2.6: TB vaccine candidates in clinical development. The pipeline is organized in three main phases: phase I trials are devoted to assess safety and immunogenicity in reference populations, usually healthy adults. Instead, Phase II trials consist on similar tests, now conducted in the age target populations. Phase III constitute a final step in which trials previous to marketing authorization are conducted at a greater scale. Image adapted from [186].

The first of these groups is that of BCG booster vaccines, designed to be applied on individuals previously immunized with BCG, and more advanced, in general, across the development pipeline. The immunogenic principles of these vaccines are based on providing an additional exposure to one or several relevant antigens from BCG itself or from *MTB*, in their standard molecular form or, instead, as protein hybrids of two or more antigens. The delivery of these antigens in host's organisms is achieved through the usage of viral vectors or synthetic adjuvants.

The second group of vaccines is constituted by prime, vaccines consist of attenuated live strains of *M.bovis* or *MTB*, genetically modified so as to guarantee safe, avirulent but yet immunogenic phenotypes. Primer vaccines, instead of boosters, are intended to substitute BCG, although, in principle, they could be used on individuals previously vaccinated with BCG. The most advanced of these vaccines is the BCG-recombinant VPM1002, currently into PhaseII clinical trials [208, 209].

Caveats in the process of vaccine development

Current knowledge of tuberculosis immunology is only partial. The complexity of the infection cycle of the disease, and the disparity of the immunological processes that articulate host's response make difficult to find robust immunological correlates of infection or protection, something that constitutes a fundamental drawback for the development of novel vaccines [186]. For example, $IFN\gamma$ production by T-cells, even

if it has been found to be significantly associated to the onset of TB disease, may not be so clearly correlated to vaccine-induced protection [210]. In this line, the adequacy of using the presence of multifunctional T cells as a principal readout for predicting vaccine efficacy is currently being questioned, after a trial conducted on South African infants in which no correlation was found between BCG induced protection and multifunctional T-cells ten weeks after vaccination [211]. Furthermore, other readouts for vaccine protection, like lung histopathology and replication reduction, offer also highly arguable outcomes [212], that very often result to be poorly correlated to actual vaccine protection.

In this context, the lack of an adequate protection correlate shifts the focus to clinical trials of efficacy determination as the main tool we count with to reliably evaluate different vaccine candidates. However, these kind of trials must be conducted in high burden areas, and need the recruitment of numerous cohorts of individuals and observe them during long periods of time in order to generate statistically relevant results; which makes them extremely costly, slow and hard to design. For example, the most advanced vaccine in the development pipeline is the booster MVA85A, who presented in January 2014 the results of a phase IIb trials conducted in Worcester, South Africa, for the testing of safety and efficacy of the novel vaccine [213]. This trial, designed by the South Africa Tuberculosis Vaccine Initiative –SATVI–, constituted the first Phase IIb clinical trial conducted since the approval of BCG; and needed the recruitment of almost 3000 BCG-vaccinated infants (6 months old).

Although the safety assessment of MVA85A in the trial of Worcester was positive, the disappointing results regarding vaccine efficacy have made the whole research community enter in a state of alarm. The failure of MVA85A at providing consistent protection, either to infection or to disease progression, has made evident the need of improving the procedures to compare and discriminate possible candidates at their early stages of development. According to both funders and researchers, a cardinal priority in the field is to limit, as much as possible, the risks of that any other vaccine candidate comes to reach reach advanced phases of the pipeline to finally get there a negative result.

That is arguably a fundamental goal, but it is obviously pointless unless any vaccine able to offer adequate protection levels when properly tested and administered can confidently rely on not being prematurely discarded during the development process. In this context, the development of new, better protection correlates for TB is an urgent necessity.

Whom do vaccinate first?

Until immunology advances allow the identification of general protection correlates of usefulness on all the vaccine candidates, it is indeed the refinement of the design of clinical trials the only tools available for the community to ensure that, if not positive, at least meaningful outcomes are obtained from the trials themselves. Among the aspects that have been more strongly questioned after the publications of the results of the trials of MVA85A; a central place is reserved for the age group of the populations

on which the efficacy of the vaccines must be tested, and eventually, applied.

On the one hand, the results of ref. [213] suggest that MVA85A is not able to offer a better performance than BCG on infants. However, when talking about adults, its immunogenic role could be different because that –recalling that the waning of BCG induced immunity is thought to take place after 10-20 years of vaccination [214, 203, 215]–, at that age, the baseline BCG reduced performance could be much easier to overcome –or repair– by the administration of a booster vaccine. Additionally, temporal patterns of efficacy waning found in BCG could also be observed again in novel vaccines –no matter if these are boosters or primes– , which would advise to apply them on adolescents or young adults, so as to take advantage of the first, *fresh* years of the new vaccine to cover these age groups, which are those bearing more intense incidence and prevalence levels, as it has been discussed before.

This situation has made emerge a general consensus towards the idea of that new vaccines should be targeted on adolescents, in order to provide a rapid and more intense protection to these age groups more affected by the disease [216, 217].

Notwithstanding that, the application of tuberculosis vaccines on individuals of that age may not be equally suitable for all new vaccine candidates. In the case of live attenuated vaccines, waiting until adolescence make target populations be consistently sensitized to NTM at the moment of vaccination in many parts of the world. That might impair proper reproduction of any novel live vaccine, preventing adequate vaccine "take" the same way it is thought to impair that of BCG when applied on individuals of this age [194, 195] (blocking hypothesis), or, alternatively, vaccine mediated protection could be impossible to observe above protection levels induced by individuals's exposition to mycobacterial antigens (masking hypothesis). Noteworthy, not only live vaccines might suffer efficacy loss due to these effects, but also boosters [139]. For example, for booster vaccines based on viral vectors, the existence of viral-antibodies in the vaccinated populations (for instance, from the smallpox vaccination campaign) could equally block viral booster vaccines [139], and protection by environmental sources of mycobacterial antigens could mask their effect too.

Furthermore, it is worth noting that, in case the vaccine is intended to have any additional effect beyond diminishing the infection probability of immunized individuals (paradigmatically, reducing the progression rates from latency to disease), the difference between vaccinating people before or after they could have been infected with *MTB* has to be considered, as *MTB* itself may act, in these cases, as a source of antigenic sensitization that could affect vaccine performance in what regards these additional effects, in the large population fractions latently infected. For this type of vaccines, no matter the vaccine type, depending on whether a vaccine applied on already infected individuals with *MTB* retains its properties or not, the optimal vaccination strategies could differ.

In conclusion, defining the precise age group on which new vaccines should be administrated (and previously tested) to achieve greater impacts is a crucial task affected by different effects. The quantification of the the effects of prior sensitization to mycobacterial antigens, and their variation with age is, indeed, one of the main objectives of this Thesis.

MTBVAC: learning to tame bacteria

After VPM1002, the only alternative live attenuated vaccine currently on clinical trials is *MTB*VAC, developed by the group of mycobacterium genetics of the University of Zaragoza [129]. Furthermore, *MTB*VAC is the only candidate that instead on being based on *BCG* modification or boosting, consists in the attenuation of a virulent strain of *MTB*, something that, once it has been demonstrated as safe as BCG in humans, constitutes a promising, conceptual advantage with respect to other candidates. The reason is that its similarity with the causative agent of the disease is, by definition, greater than that of any other vaccine, and, particularly, it shares a more complete antigen collection with wild type *MTB*, overcoming this way the principal limitation of boosters, which base their activity on the presence of a very limited number of antigens [186], or recombinant-BCG candidates, which are genetically further to *MTB* [218].

*MTB*VAC is built upon the deletion of two key genes that regulate different virulence related mechanisms: PhoP and fad26. The vaccine represents a second version of a previous vaccine, called SO2, based in the sole deletion of PhoP. Although SO2 passed through successive phases of security and stability testing [219, 220, 221, 222], it did not fulfill the Geneva consensus, that stablished –among other stringent requisites– that novel *MTB* based vaccine candidates must present, at least, two stable gene deletions within the genome to be considered safe enough to enter clinical trials [223, 224], which rendered SO2 unsuitable for further development. The incorporation to the second gene deletion of fadD26 solved the problem, generating a vaccine fulfilling all the required safety requisites, yet functionally similar to SO2.

PhoP is the sensor protein of a two component signaling system that acts as a global regulator [88]. Among the genes controlled by PhoP there are several key pieces for *MTB* virulence. First, it activates transcription of proteins of the ESX1 secretion system, that allow proper secretion of the key antigens ESAT6-cfp10 [88]. Second, it activates a novel non-coding RNA recently discovered, named *mcr7*, responsible of repressing tatC, a protein of TAT secretion system, in charge of secretion of Ag85 complex antigens [132, 225]. Thus, deletion of PhoP translates on a phenotype whereby ESAT-6/cfp10 are expressed but not secreted, and secretion levels of Ag85 antigens are, in turn, augmented with respect to wild-type strains. This ultimately translates into reduced pathogenicity (because of *MTB*VAC inability to secrete ESAT6) concurrent with immunogenicity levels improved (due to augmented exposure to Ag85 and intracellular presence of ESAT6, which BCG lacks)[129].

On the other hand, fadD26 gene is needed for the synthesis of phthiocerol dimycocerosates (PDIM), complex virulence lipids of the cell wall [226], and so, its knockout also contributes to reduce virulence in *MTB*VAC. Remarkably, if we recall section 2.1.2, we have that synthesis of PDIMs is a resource of *MTB* to detoxify the pool of cytoplasmic propionyl-coA, through the methylmalonyl pathway, a metabolic route that can not be completed after the knockout of *fadD26* (see figure2.2). As a consequence, the bacterium is still able to transform propionyl-coA into methylmalonyl-coA; but further synthesis of PDIM from that point is impaired because of the deletion of *fadD26* [227, 228]. This situation could ultimately compromise the ability of *MTB*VAC to per-

sist with a lipid based metabolism, because of a deficient detoxification of propionyl-coA products from the methylmalonyl cycle. Fortunately, it is not the case, as *MTB*VAC is able to persist within the host. Indeed, the reason for this is thought to be related to the isocitrate-lyase (*icl*) which indeed constitutes one of the few genes repressed by *PhoP*. In this way, PhoP deletion in *MTB*VAC causes an enhanced expression for *icl*, which offers a robust alternative pathway to detoxify propionyl-coA through the 2-methylcitrate acid cycle, and maintain an adequate REDOX balance when lipids constitute the principal metabolic substrate, even if *MTB*VAC is not able to synthesize PDIMs. In conclussion, simultaneous deletion of *PhoP* and *fadD26* reduces *MTB*VAC virulence without compromising its *in vivo* survival [88, 129].

The combination of these three features –i.e. low pathogenicity, enhanced immunogenicity and persistence– makes *MTB*VAC a conceptually optimal vaccine candidate. Nowadays, the novel vaccine is finishing Phase I clinical trials, the performance of the novel vaccine, as well as the level of protection it is able to offer will be eventually measured in the following years, if *MTB*VAC progresses to the final phases of the production pipeline satisfactorily.

Part II

Networks of bio-molecular interactions in *M.tuberculosis*

–Pray of what disease did Mr. Badman die, for now I perceive we are come up to his death?

–I cannot so properly say that he died of one disease, for there were many that had consented, and laid their heads together to bring him to his end. He was dropsical, he was consumptive, he was surfeited, was gouty, and, as some say, he had a tang of the Pox in his bowels. Yet the Captain of all these men of death that came against him to take him away, was the Consumption, for 'twas that that brought him down to the grave.

John Bunyan
Life and Death of Mr. Badman, 1680.

Chapter 3

The transcriptional regulatory network of M.tb

3.1 Introduction.

A thorough understanding of the biochemical mechanisms behind *MTB* infection is an unavoidable step for the development of new drugs and vaccines aimed at eradicating TB. Such a challenging goal would imply to comprehend both the transcriptional control of the signaling system and the interaction of the bacillus with the immune system of the host. As a first step, it is necessary to study the backbone of this complex system, i.e., the circuitry behind the biochemical processes operating at the gene expression and signaling levels. In this chapter, we report the most complete transcriptional regulatory network (TRN) of *MTB* to the date this work [3] was published, in July 2011. Capitalizing on previous attempts to build up the TRN of *MTB*, we have been able to assemble a network that links together the many isolated interactions reported in the literature.

Thinking on a wider perspective than the strictly biomedical interest of this work, our study also addresses important open questions in the field of network science. To the best of our knowledge, the network here assembled is the first intracellular pathogen whose TRN is characterized to a reasonable level of accuracy and completeness. In this paper, we calculate the main macro-scale features of the *MTB* TRN: connectivity distributions and mean values and clustering coefficients as well as average path lengths [50]. Of further interest is the analysis of small scale features as given by the abundances and significances of network motifs. As it has been thoroughly discussed in the introduction to this thesis, motif abundance profiles can be interpreted as topological markers carrying relevant information about the effective tasks for which the network under study is designed or evolved [261, 262]. To illustrate this point in this context, we also present a comparative topological analysis between the TRN of an intracellular parasite like *MTB* and other already available analogous system, *E.coli*, with diametrically opposite life styles.

3.2 Results.

3.2.1 Construction of the TRN of *MTB*

Our starting point is the TRN proposed by Balázsi and colleagues a few years ago [89], which is the largest *MTB* transcriptional network to date (783 nodes, 50 regulators and 936 links). Based on this TRN, we have performed a considerable expansion by using publicly available sources, most of which appeared after Balazsi's compilation,

expanding the network up to 1624 nodes (which supposses, roughly, a 40% of the genome) and 3212 interactions sent from 83 transcription and sigma factors.

For such an expansion, the links we are reporting have been reported upon the outcomes of experiments belonging to two different groups of methodologies. Within the first family of experimental procedures, we have considered techniques that are based on detecting significant changes of target-gene expression levels caused by disrupting, over-expressing, or inducing a certain regulator, compared with wild type reference expression levels. These techniques include microarray (MA) analysis (genome-wide, poorly specific), or quantitative real time PCR analysis (qRT-PCR , that provides higher accuracy and reliability), as well as fusion in target promoters of sequences coding reporters like green fluorescent protein (GFP) or *lacZ*. On the other hand, the second family of methodologies covers procedures that are based on the identification of the binding sites of DNA and transcription factors (TFs), and, eventually, the characterization of the physical protein-DNA interaction. Electrophoretic mobility shift assays, one hybrid reporter systems and ChiP-on-chip assays are examples of these methodologies. Moreover, once the new information coming from experimental sources and computational inference is compiled, we have further enlarged the network by operon-based expansion as done in [89], using the operon map predicted in [263]. See *Materials and Methods* for more details. Figure 3.1 shows the resulting TRN of *MTB*. We next analyze its main topological properties at different resolution levels.

3.2.2 Global topological properties of the network.

We have measured several topological properties of the assembled network. It consists of $N = 1624$ nodes and $E = 3212$ edges, with a giant connected component of order N. As the network is directed, one can compute the total degree of a node, k, as the sum of the incoming links (meaning that the node is a target of a regulation) and the outgoing links (meaning that the node regulates another node), i.e., $k = k_{in} + k_{out}$. Figure 3.2 represents the cumulative degree distributions for the TRN of *MTB*, either for total, in and out connectivities. As can be seen from the figure, the cumulative degree distribution reasonably follows a power law $P(k\rangle k^*) \sim k^{-\gamma}$, (best fit $\gamma = 0.99 \pm 0.07$ with $R^2 = 0.974$). In other words, the TRN of *MTB* shows the same highly heterogeneous degree distribution found for other biological networks not only at the cellular level, but also at larger scales [50]. This means that the vast majority of genes only interact with a few other genes, while there is a small but statistically significant number of genes that interact with hundreds of genes.

Concerning the directionality of the regulations we find, as usual for these kind of system that links senders constitute a small fraction of all the genes in the network: 83 over 1624. This is coherent with the fact that the larger contribution to degree heterogeneity is caused by out-coming links, sent by this reduced set of 83 regulators formed by transcription and sigma factors. In-degree heterogeneities are much more reduced, as, even is most of genes in the network receive some regulation (1598 over 1624), the gene receiving more links gets just 12 regulations.

This can be appreciated already in table 3.1, where we have summarized several

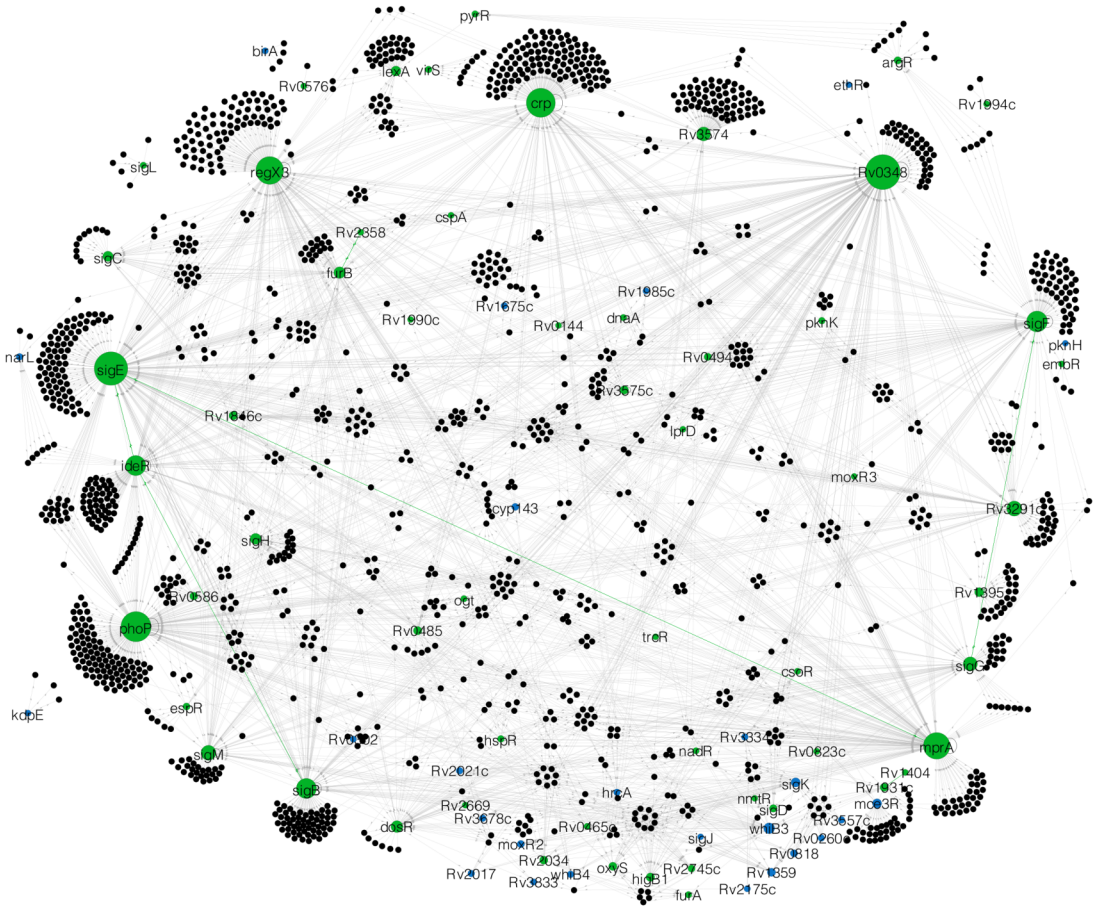


FIGURE 3.1: TRN of *MTB*. Blue nodes represent regulatory genes that are not regulated by other nodes, while green ones are transit nodes, both regulating the activity of other targets and receiving regulation from other TFs. Self-regulations are represented by black arcs, while feedback loops (FBL) of mutual regulations are represented in green, thick lines.

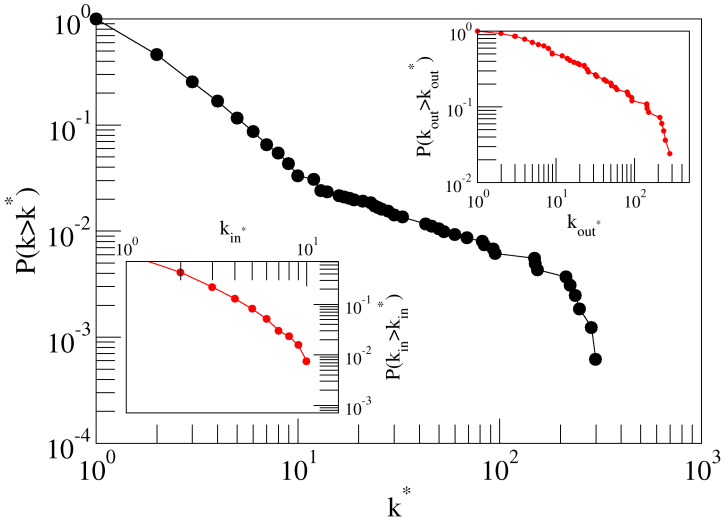


FIGURE 3.2: Total, in and out degree distributions of the TRN of *MTB*. The figure shows the cumulative degree distributions, i.e., the probability to find a node whose connectivity is larger than or equal to k^*. For the in and out distributions –in the insets– only genes that receives (1598) or send (83) links has been considered in each case so as to normalize the distribution. Hence, the lower inset represent the probability $P(k_{in} > k^*_{in})$ for any of the 1598 regulated genes to receive more than $k_{in}*$ links; while, for the upper inset, we represent the probability $P(k_{out} > k^*_{out})$ for each of the 83 regulators found to send more than $k_{out}*$ links.

Property	*MTB* TRN
genes	1624
Regulators	83
Links	3212
Self-loops	43
2 nodes FBLs	6
Mean connectivity	3.96
Mean in-connectivity	2.01
Mean out-connectivity	38.70
Directed average path length (Giant Component)	2.07

TABLE 3.1: Topological properties of TRN of *MTB*. We report some global metrics of the network such as its mean connectivity and directed average path length. For the definition of this quantities, see [50]. Note that the mean in and out degrees are calculated with respect to the number of target genes and regulators, so they are not normalized in the same way.

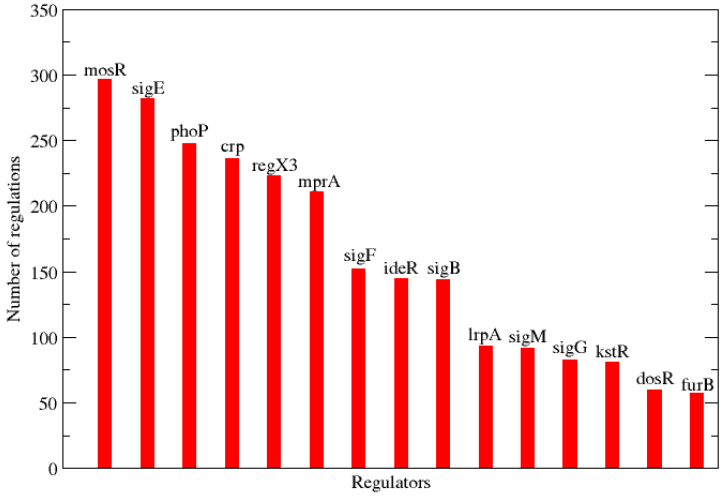


FIGURE 3.3: Most connected regulatory hubs in the *MTB* TRN.

topological properties. As a matter of fact, the average out-degree of TFs is much larger than the average in-degree of genes that have at least one regulator, indicating that the networks retains this typical characteristics of this kind of networks (i.e. a reduced set of links senders distribute all the links across the system, and so, when summing up incoming and outcoming degrees, most hubs are link senders). At this point it worths mentioning that average in and out degrees (as well as the degree distributions represented in the insets of figure 3.2) are not calculated in the usual way, -i.e. averaging over N-, otherwise $\langle k_{\text{in}} \rangle = \langle k_{\text{out}} \rangle = \langle k \rangle / 2$), but averaging only over the number of link senders and over the number of link receivers for $\langle k_{\text{out}} \rangle$ and $\langle k_{\text{in}} \rangle$, respectively. Concerning other topological features, we see that they are within the typical range of values for other biological (and, in general, real complex) networks [50].

We have also inspected what are the most connected TFs, see figure 3.3. As one can see, *phoP*, whose deletion –aside to that of fadD26– is at the root of the new live vaccine *MTB*VAC [129], appears as the third most connected hub of the network, only following to *mosR* and *sigmaE* regulators. Indeed, it is known that *PhoP* regulates key functions required for the intracellular survival and persistence of *MTB*. Remarkably, as it has discussed in the introduction, inactivation of *phoP* results in down-regulation of genes needed to successfully grow within macrophages and consequently in *MTB* virulence attenuation. Summarizing, TFs and SigFs highlighted in the degree rank as principal network hubs in our system are those for which a role of global regulators -i.e. a large number of regulations targeting a relevant fraction of the bacterial genome– has been already characterized in the literature analyzed to build the network. Admittedly, taken that into account, it is not surprising that key roles in cell cycle regulation, metabolic control and virulence, similar to those mentioned for PhoP, had been reported precisely for these genes, provided that the very existence of such studies is the ultimate reason

for the privileged topological situation that these nodes have in our network.

However, even that the analysis summarized in figure 3.3 is in that sense quite straightforward once the network is built up, integration of transcriptomic data around a single dataset is a meaningful way to compare the role and relevance of network regulators that have been similarly studied in literature.

The situation is somehow different when checking what are the genes receiving a higher number of regulations in the system, as their identification among all the genes of the network is a genuine result derived from the assembling of the network. The reason is that, unless what happens with global regulators (i.e. one can understand the their global action by obtaining a knockout mutant and conducting a genome wide transcriptomic comparison WT-mutant, with no particular necessity of a network based context), the task of identifying genes receiving more transcriptional inputs than others is not so clear.

Although the heterogeneity in the number of links received per gene is much more reduced than for the case of link senders, genes receiving a higher number of regulatory inputs are reasonably expected to be related to relevant biological functions that must can be controlled either by a different set of inputs (or particular combinations of them). Remarkably, among the only 12 genes that receive more than 10 links in our network we have the gene cluster Rv2930-Rv2939 (receiving 12 links each) and the genes icl and fadB2 (receiving 10 links each). The gene group that spans from fadD26 (Rv2930) to Rv2939 is known to be involved, as it has been discussed in the Introduction, in the biosynthesis of the complex lipids phthiocerol dimycocerosates (PDIM), some of the more relevant mycobacterial virulence factors. In what regards to icl and fadB2; they are both part of the same operon, and they are related to the metabolic shift from carbo-hidrates to lipids that follows intracellular infection [88, 264], as it was also issued in the introduction to this thesis. Very remarkable is the role of icl, whose expression has been related to the pahogen's ability to persist within macrophage.

Relevantly, the final genetic profile of the *MTB* based, live attenuated vaccine *MTB*VAC modifies the expression profiles of all these highly regulated genes. On the one hand, fadD26 (Rv2930) is directly deleted in *MTB*VAC, which prevents the bacterium to synthesize PDIMs [129]. On the other hand, the couple icl-fadB2 has been identified as one of the few genes negatively regulated by PhoP, which has been related to the more accused tendency of *MTB*VAC to show a dormant, persistent phenotype [88], a definitely desired property for a live vaccine to present.

All these results, taken together, suppose an examplo of how trying to establish links between genes biological relevance and their position and importance within a regulatory network often is a meaningful way to give clues about their potential as putative drug targets.

3.2.3 Small-scale properties of the network: motifs.

Beyond the topological characterization of the macroscopic traits of the network summarized in table 3.1, mesoscopic or small scale topological features are key to understand the dynamics and function of biological networks like the one we are studying

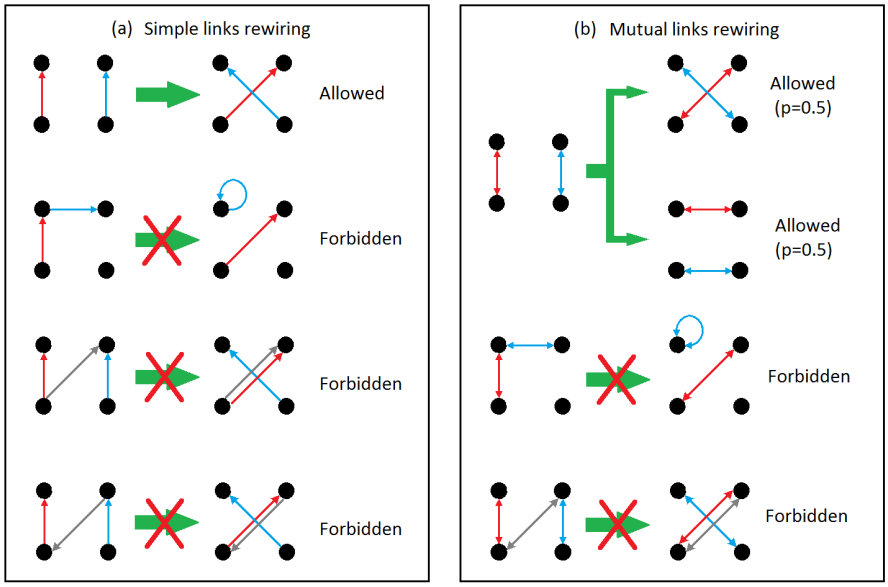


FIGURE 3.4: The figure represents the set of allowed and forbidden rewiring steps for the randomization of the TRN. The left panel corresponds to the situation in which simple links are being rewired whereas the right panel represent the cases considered when mutual links are being rewired. More details are provided in 3.5.

in this chapter. For instance, as we have discussed in the introduction to this thesis, communities are often identified with functional modules, and motifs statistics are thought to store much relevant information about the structural design principles of the complex systems in which appear.

Using an approach similar to that introduced by Alon and coworkers [261, 262], we have registered the number of appearances for each of the subgraphs of size 3 and 4 in the TRN of *MTB* and compared them to what is found in a null ensemble of networks built up from the original system after an adequate random rewiring procedure schematized in figure 3.4. In this case, the null ensemble essentially consists in a collection of randomized networks, in each of which every node preserves its in and out degree -and, remarkably, its number of –bi-directional links–, but its neighbors are randomized.

The way in which the null ensemble can be constructed for extracting meaningful conclussions from NMs analysis has been the subject of intense research in the last years, and besides the method used in this paper, there are other alternative randomization schemes [262, 265, 266].

Once the ensemble of randomized networks is generated, we calculate the mean values and the typical deviations of the number of appearances of each of the possible motifs of a given size in all the randomized networks. The significance of each motif h is determined by the following Z-Score:

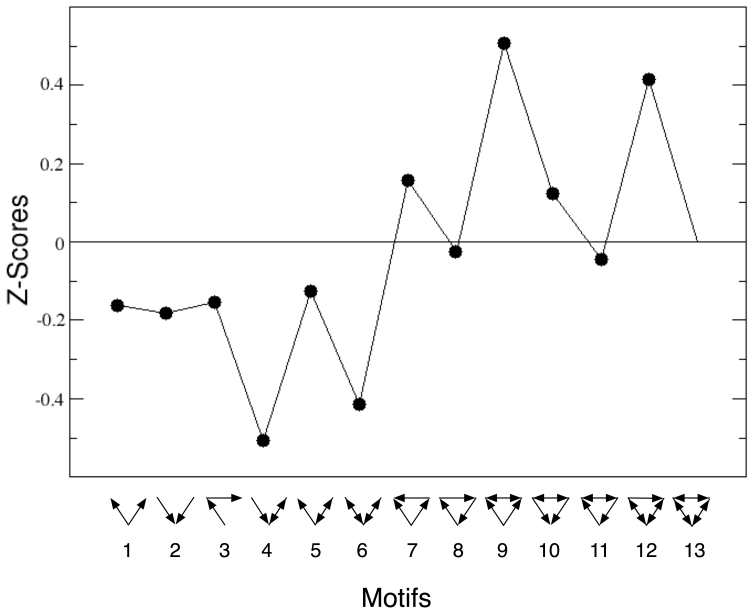


FIGURE 3.5: Triads significance profile of the *MTB* TRN. We represent the values of the Z-scores as defined in Eq. (3.1) for each of the 13 possible 3-nodes directed motifs, which are depicted in the x-axis. Only one of them cannot be defined in the *MTB* network. See the main text for further details as well as the section 3.5.

$$Z - Score_h = \frac{n_h - \langle n_{rand,h} \rangle}{\sigma_{rand,h}} \tag{3.1}$$

At this point, it is worth noticing that, if a given motif h, is absent in any of the randomized networks of the null model, then $\sigma_{rand,h}$ is not defined and so is $Z - Score_h$. Moreover, all the defined Z-scores are normalized considering each of them to be a component of a unitary vector. This normalization allows to compare motifs significance profiles of networks of very different sizes.

We have calculated the significance profiles for the newly assembled TRN of *MTB* focusing on 3 and 4 nodes' motifs. Figure 3.5 shows the Z-scores obtained for triads. Note that as our network is directed, so are the triads. Out of 13 possible motifs of size 3, only triad 13 is not present in the Z-scores representation. Additionally, the first six motifs, all of which correspond to open structures (i.e., loopless motifs) are underrepresented in the *MTB* TRN, while those that are found more frequently than in the random version of the TRN are closed motifs.

3.3 Tetrads statistics of *MTB* and *E.coli.*

The kind of statistics of motifs as represented in figure 3.5 for our case, have been commonly used as a way to deepen our understanding of the relation between the

structure and function the biological networks analyzed. Specifically, as we have discussed in the introduction to this thesis, previous works have focussed on directed triads as a means to categorize networks around superfamilies which, roughly, share the same functionality even when the networks belong to very different fields [267]. Two distinct superfamilies were identified for informational-processing networks [267]. In one of these superfamilies, we can find networks for which the response of the whole system to an stimulus cannot last much more than the response time of one of its interactions. These so-called *rate-limited* networks include TRNs of unicellular microorganisms, covering both prokaryotic and eukaryotic organisms. The second informational processing superfamily groups networks that work in a less immediate way, being able to perform slower responses with characteristic times that can be even several orders of magnitude greater than the response times of its single interactions. Synaptic and developmental networks in multicellular organisms are typical cases of this kind of *not rate-limited* behaviors.

The question is then how the TRN of an intracellular parasite like the *MTB* integrates into the above framework, i.e., which of the two informational processing superfamilies the *MTB* TRN belongs to. It is worth mentioning that *a priori* the question is not trivial. All unicellular microorganisms have presented up to now the characteristic profiles of rate-limited networks [267], so that one may be tempted to ascribe to this family the network of *MTB*. However, the bacillus spent most of its vital cycle within the macrophage, so that the characteristic response times of its TRN should be slower. Several findings support this hypothesis. For instance, it has been reported [89] that after hypoxia stimulus, the dynamics stabilizes in a time as long as eighty days, suggesting that the TRN of *MTB* could better adjust to the not rate-limited superfamily. As a matter of fact, after a direct comparison of the triads profile of *MTB* with those reported in [267] would seem that this is indeed the case. The interpretation of this result is beyond the scope of this section, and will be addressed in chapter 5.

Beyond the question of network superfamilies, it is also of interest to compare the TRNs of *E.coli* and *MTB* as two examples of prokaryotic, unicellular bacteria. The aim is to identify whether there are relevant divergences in their motifs profiles that could be related to their different vital cycles: that of an intracellular parasite in the case of *MTB* and that of an extracellular bacterium in the case of *E.coli*.

Taking into account the above remarks, we have compared the Z-scores significance profiles for the networks of *E.coli* and *MTB* looking for common (i.e., present in both organisms) structures that exhibit the larger differences in their Z-scores. In our analysis, we only consider FBL-free tetrads that appear in both TRNs at least a thousand times. This pruning is made so as to restrict our analysis on motifs for which we are sure that the analysis is robust enough- This excludes scarce structures and motifs that, even being relatively numerous, only appear grouped around punctual structures in the system: i.e. FBLs, which are really scarce in our systems (only 10 in the TRN of *E.coli* and 5 in the TRN of *MTB*). Within this subset, of confidently well represented motifs, we have selected the structures with a differential behavior in both systems (i.e. tetrads being over-represented ($Z_{score} > 0$) in one system and under-represented in the other ($Z_{score} < 0$). The reason for choosing tetrads instead of triads for this analysis is

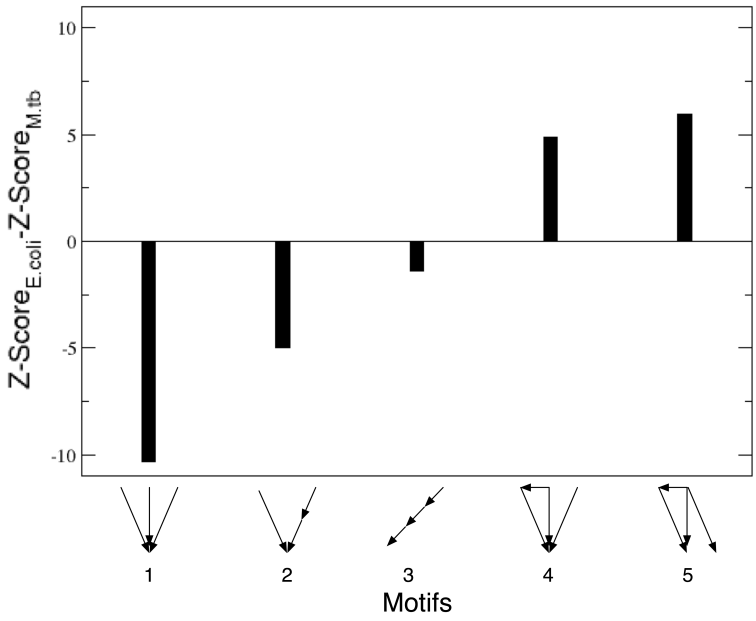


FIGURE 3.6: Differentially significant motifs present in *MTB* and *E.coli* TRNs. Feedback-free tetrads present more than 1000 times in both networks were divided into three groups according to their Z-scores: we consider as overrepresented those tetrads with Z-Score> 1 and as underrepresented those with Z-Score< 1, being the third group joined by tetrads for which $|\text{Z-score}| < 1$. We have only looked for motifs that belong to different groups in *E.coli* and in *MTB*, and we have sorted them according to the absolute difference ΔZ=Z-Score$_{E.Coli}$-Z-Score$_{M.tb}$. After this filtering process, the tetrads in the figure are those with $|\Delta Z| > 1$. The resulting tetrads with highest $|\Delta Z|$ are the same when one takes the normalized Z-score.

that, the more numerous statistics associated to them allows us to find some examples of such divergent motifs between the networks being analyzed: indeed, Z-scores of triads have always the same sign in both TRNs, so, the kind of differential behaviors we want to look at are not present at the –more coarse grained– level of triads.

The results are depicted in figure 3.6, where we have plotted the cases, within this group, for which the differences in Z-scores between tetrads of both networks are greater than unity. As one can see in the figure, *E.coli* presents stronger trends for two of the simplest parallel combinations of feedforward loops (FFLs) with single regulations (tetrads 4 and 5). On the other hand, the three structures with highest differences in favor of *MTB* are single output modules (SOM, tetrad 1), cascades, (tetrad 3) and the combination of them (tetrad 2). The preference for different motifs in *E.coli* and *MTB* TRNs supports the evolutionary pressure hypothesis put forward in [261, 262, 267]. At a glance, given the homology between the two networks − both organisms are bacteria −, one might think that their TRNs come from common underlying mechanisms that regulate the way they grow and are assembled. However, the TRNs should have been shaped in relation to the vital needs of both bacteria, which are known to be radically different. Therefore, over and under represented subgraphs

with different significances in these two organisms would be, under this perspective, a consequence of an evolutionary process related to the adaptation of both bacteria to their respective life-styles.

From a dynamical point of view, the previous results also make sense. The dynamics associated to most of the motifs that appear to be under or over represented in the TRNs of *MTB* or *E.coli* have been well characterized previously [261, 268, 269, 270, 271, 272, 273, 274]. For instance, FFLs have been described as a discriminator between persistent and transient signals [261, 268] or as a pulse-generator to accelerate signal responses [269]. The performance of the motif as persistence discriminator or accelerator depends, essentially, on the signs of the regulations and the logical scheme [275], but nevertheless, both functions are useful for the organism when it is subjected to a highly dynamic or unpredictable environment. This is the case of *E.coli* whose environment offers a very rich amount of nutrients to the bacterium, but also a considerable concentration of harmful substances and other threats. This highly variable environment requires a valuable mechanism to filter the very noisy signals that are being sensed continuously. Therefore, it is to be expected that FFLs and closely related structures be over represented in the *E.coli* TRN, as they are.

On the other hand, as noted before, the *MTB* bacillus expends, on average, more than 90% of its vital cycle in a dormant state, surviving under a latency regime within the human immune system macrophages, which principal function is, essentially to kill the pathogen. In this sense, it is difficult to conceive a bacterial environment more hostile than that of *MTB*. However, hostility does not mean variability, so that the environment of any bacterium could be hardly less volatile or more predictable than the intracellular environment to which *MTB* has specifically evolved to live within. These considerations are in qualitative agreement with our observations of a less marked bent for FFLs in *MTB*.

3.4 Conclusion

In this chapter, we have assembled the more complete TRN of *MTB*. to the date this work was published. The network has been built up by exhaustively including publicly available bibliographical information relative to MA essays, protein-promoter binding sites determination experiments and synthetic biology techniques coming from as many as 31 different works. The importance of gathering all this information into a common frame with tractable format is twofold. First, it is a valuable database that can be directly used for research purposes. And second, the new network constitutes an important tool for the application and development of computationally inspired models and methods that may be able to guide future in-vitro and in-vivo experiments. The latter includes using the network to develop new tools for tasks such as the identification of spurious links and missing interactions [276], prediction of unknown functions of the proteins, the generation of more accurate operon maps predictions [263] or the dynamical modeling of network operations as a sensory system [277]. Concerning the role of *phoP*, our analysis shows that it appears as the third-most connected hub of

the network (see figure 3.3), thus, the relevance of altering its functionality is easily understandable and expected from a network viewpoint. Analyzing the most regulated genes in our network remarkably we recover genes like icl or fadD26, whose expression profiles are altered in the new vaccine *MTB*VAC, resulting in improved persistence and reduced virulence, in part because of the altered performance of the very genes mentioned [88, 129].

Additionally, we have performed a detailed analysis of the topology of the TRN. The results show that our system shares the main macro-scale features of other similar systems such as the small-world property or a fat-tailed degree distribution [278]. The statistical characterization of the relative abundance of different motifs has also shown interesting results.

First, our new system could initially be incorporated into the differentiation scheme firstly proposed in [267], apparently joining the family of not rate-limited family proposed by Alon et al. [267]. This preliminary result is intriguing, as, if confirmed, would suppose the first example of an unicellular organism resembling, at the level of NMs statistics, the structure of pluricellular networks of biological information processing.

Secondly, we have performed a comparative analysis between the significance levels of the most relevant tetrads in the TRNs of *E.coli* and *MTB*. The reason for choosing tetrads instead triads is due to the fact that, given the more reduced combinations that we have for triads, there exist no triad of different Z_{score} sign of size 3. Our results show that the relative abundance of the different subgraphs is reasonably related to the dynamical response of each subgraph [261, 268, 269, 270, 271, 272, 273, 274] and to the life styles (in relation to their environments) of both bacteria. Besides, given that the TRNs correspond to two bacteria, the differences between subgraph significances of *E.coli* and *MTB* can only be a consequence of divergent evolutive pathways.

In summary, the expanded TRN will be useful to provide an overview of multiple functional aspects of *MTB*, and to suggest new experiments. It is also important to note (especially for future studies) that the network obtained neither distinguish the relative strength of the different interactions reported (i.e. link weights) nor the specific environmental conditions or signals needed for a given regulation to actually take place under in-vivo conditions.

About interaction signs, although they were not reported in the original article in which we published this research [3], these were collected whenever possible and reported in a second release firstly referenced in [6].

Besides, the methodology used to expand the network and the different kinds of resources used in the process, which have been treated case by case, individually for each TF, allow for a quick and easy review and extension of the network. Finally, the development of a functional sense upon the extended TRN of *MTB* that encompasses the whole network and the integrative action of the signaling system, would be one of the essential objectives to achieve in the future.

3.5 Materials and methods

3.5.1 Bibliographical revision and datasets of the TRN of *MTB*

We have updated the network presented in [89], using new and copious experimental information available from as many as 31 different works dated in the last ten years, (see table 3.2). To assemble and expand the network, we have also used the predicted operon map proposed in [263] assuming that if a TF A regulates a gene B belonging to an operon BCD, then, also the interactions A-C and A-D are present.

All the information extracted is available in the supplementary information of [3], including all the links, works in which they had been reported (also including those already present in Balázsi's compilation) and experimental methodologies used to infer them.

Finally, we filtered all repeated information to build our network. The final system contains more than two times the number of genes or links than the previously available dataset [89]. The expanded network is freely available on-line [279], also with all the interaction signs, whenever available.

3.5.2 *E.coli* TRN

We have used the TRN of *E.coli* [314] updated as of August 2010 (release 6.8), which contains experimental information until the ultimate large-scale revision published in 2008 [311]. Dimeric TFs and toxin/antitoxin systems are taken as single nodes of the network. The network can be found in the supplementary material of [3].

Chapter 4

Context-specific networks of protein-protein interactions in M.tb.

Reference	Strain	Methodologies	Regulator(s)	Links
[118]	H37Rv	MA	*phoP*	78
[280]	H37Rv	MA, qRT-PCR	*phoP*	114
[281]	H37Rv	MA	*RegX3*	98
[282]	H37Rv	MA, EMSA, CS	*dosR*	49*
[283]	H37Rv	LacZ,EMSA	*mce2R*	14
[284]	H37Rv	MA, qRT-PCR, EMSA	*mosR*	173
[285]	H37Rv	MA, qRT-PCR, CS, EMSA	*FurB*	32
[286]	H37Rv	OH	**	325
[287]	H37Rv	MA, qRT-PCR, EMSA	*MprA*	135
[288]	H37Rv	MA, qRT-PCR	*sigE*	16
[289]	H37Rv	MA, CS	*sigE*	116***
[290]	CD1551	MA, qRT-PCR, CS, IVTA	*sigF*	74
[291]	H37Rv	MA, CS	*sigM*	43
[292]	CD1551	MA, qRT-PCR, CS, IVTA	*sigG*	43
[293]	CD1551	MA, qRT-PCR, CS, IVTA	*sigB, sigF*	78
[294]	H37Rv	CS	*crp*	44
[295]	H37Rv	MA, qRT-PCR, EMSA	*RamB*	4
[296]	Erdman	MA, qRT-PCR, EMSA	*EspR*	12
[297]	H37Rv	EMSA, CS	*LrpA*	15
[298]	BCG,H37Rv	MA, qRT-PCR, PS, EMSA	*cmr*	5
[299]	CD1551	MA, qRT-PCR, CS	*sigM*	25
[300]	H37Rv	MA	*Rv0576*	2
[301]	H37Rv	MA	*IdeR*	71
[302]	H37Rv	qRT-PCR, EMSA	*CsoR*	4****
[303]	H37Rv	PSBI	*pyrR*	6
[304]	H37Rv	qRT-PCR, EMSA, CS	*mprA*	3
[305]	Erdman	LacZ, GFP	*VirS*	8
[306]	H37Rv	MA, qRT-PCR, PS, CS	*Mce3R*	27
[307]	H37Rv	ORT, EMSA, PS, CS	*KstR*	77
[308]	H37Rv	MA, qRT-PCR	*Rv0485*	13
[309]	H37Rv	qRT-PCR, EMSA, ChiP-on-chip, CS	*BlaI*	16

TABLE 3.2: The table contains the references used to build up the TRN reported in the main text as well as the TFs studied. *, *** and ****: These are works already cited in [89], nevertheless, not all the regulations included in these works was considered in the compilation of Balazsi et al. ** reports regulations coming from the following 31 TFs: *oxyS, Rv0260c, sigK, regX3, Rv0818, Rv0823c, mprA, sigE, Rv1359, Rv1931c, higB1, Rv1990c, Rv2017, Rv2021c, Rv2034, Rv2175c, Rv2669, sigB, Rv2745c, dosR, moxR3, sigJ, Rv3334, sigD, whiB3, Rv3557c, Rv3678c, whiB4, moxR2, nmtR, Rv3833.* Abreviations used: MA: microarrays. qRT-PCR: quantitative real time polimerase chain reaction. EMSA: electrophoretic mobility shift assay. CS: identification of consensus sequences. LacZ: LacZ-promoter fusion. OH: one hybrid reporter system. IVTA: in-vitro transcription assay. PS: proteomic studies. PSBI: protein structure based inference. GFP: green fluorescent protein promoter fusion. ORT: orthologies with *M.smegmatis*.

4.1 Introduction

The general purpose of this chapter is the study of the protein-protein interaction network (PPIN) of *MTB*. To do so, we capitalize on a previous dataset described by Wang et. al. [4], in which all the physical interactions between couples of proteins observable in a two-hybrid in-vitro assay are listed.

More specifically, we aim to characterize which of these interactions are strengthened under disparate environmental conditions, as a consequence of gene expression variation of the genes codifying the respective proteins involved in the interactions. In the end, at a systemic level, this means to characterize the topological transformations that the PPIN suffers as a consequence of the gene expression shifts that follow exposure to different environments, either *in vivo* or *in vitro*. To achieve that goal, we have analyzed all the information available in the on-line database gene-expression omnibus (GEO) about microarray (MA) experiments conducted in *MTB* [5].

The analysis of how the PPIN of the pathogen adapts to disparate conditions will help us to discuss some open questions regarding the life cycle of the pathogen. For example, by comparing the overall topologies of the PPIN corresponding to different in-vitro models of phagosomal infection based on exposure to different stresses may help us to address which of these models more closely resemble pathogen's response within the phagosome.

Moreover, the kind of analysis proposed in this chapter could be applied in other organisms to uncover general tendencies that could help to understand transcriptional responses to similar stresses, as well as the relationships between these responses and the organisms' life styles, among other relevant biological questions.

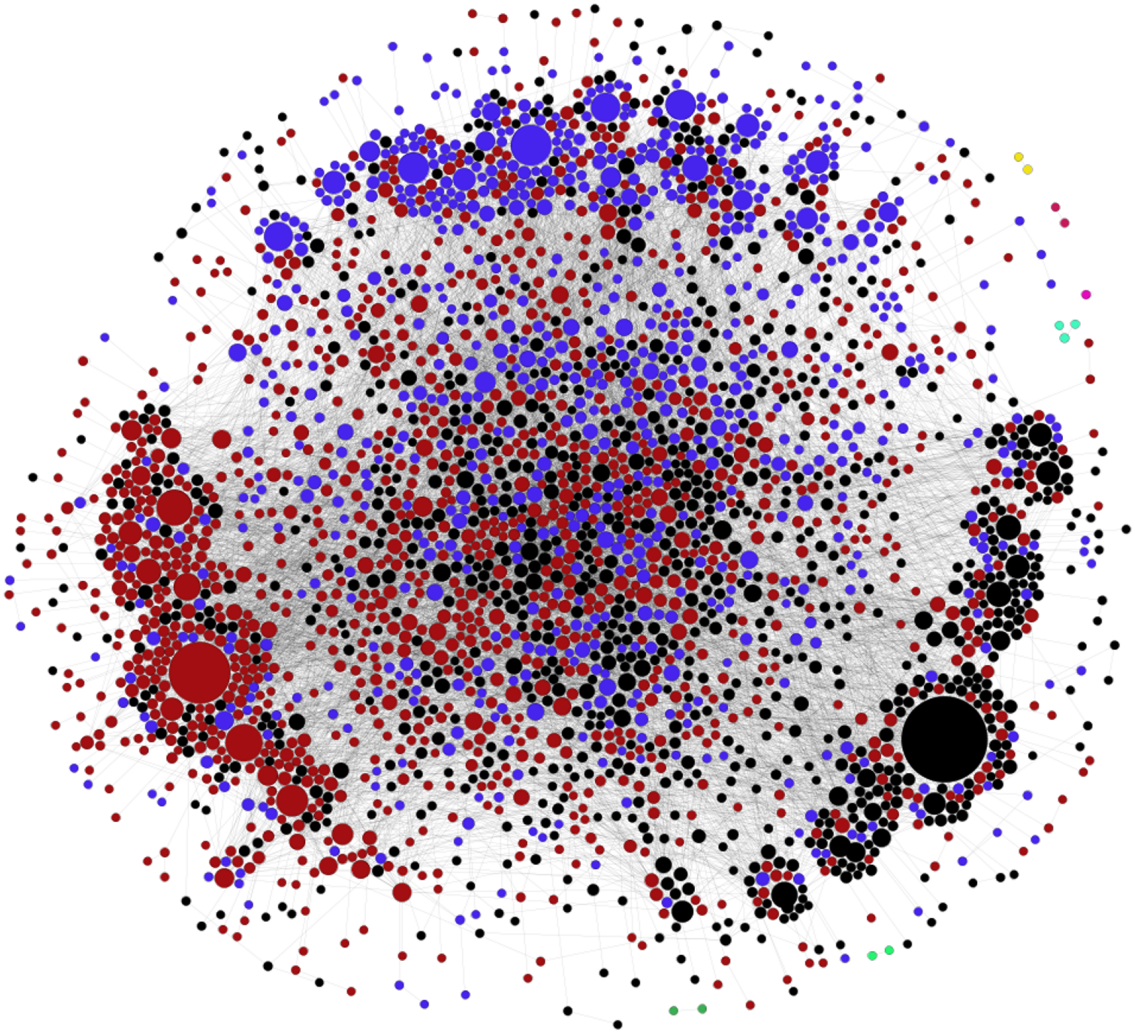


FIGURE 4.1: PPIN of *MTB* compiled in [4]. In this network each node represents a protein, and a link between two proteins is registered if they physically interact, according to the bacterial two hybrid screen performed by Wang *et al.* The network has no weights, and nodes' sizes indicate proteins connectivity in the system. Nodes' colors are the outcome of a simple community structure analysis (performed using the software Gephi for network representation and analysis [229]). The network has 2907 nodes and 8042 links

4.2 Methods

4.2.1 Multi-layer network construction

In order to evaluate the changes of the PPIN of *MTB* followed by exposure to different environmental conditions, we have followed a procedure whereby we reconstruct a different version of the original PPIN [4] for each environment analyzed. By proceeding that way, we obtain a multi-layer system [230] in which each layer corresponds to each of these versions, and contains only the genes and interactions that, in the correspondent environmental contexts, are over-expressed with respect to a basal, stress-free state.

The idea behind this rationale is certainly simple. Wang et al. dataset describes all the possible physical interactions between couples of proteins that might be observed in vivo *provided that the proteins involved in one of these interactions are present in the cell at minimum concentrations*. So, under a given environment, some interactions are expected to gain relevance -as a consequence of that the proteins involved are over-expressed- and some others will directly vanish, as a consequence of gene repression. Therefore, we aim to reconstruct a layer for each environmental condition studied, containing only genes and interactions which are over-expressed with respect to a common basal state. The topological characterization of the multi-layer system [230, 231] resulting from that process constitutes our principal scope.

In order to perform that analysis, microarray data were extracted form Gene Expression Omnibus [5] on March, 2014. GEO data is structured in GEO-series, each of them containing a set of GEO-samples. Each GEO-sample contains the information of a single microarray experiment in which, typically, two conditions are compared, andfolding rates (FR) associated to expression variation between both conditions for each gene are reported. The first of these conditions (channel 1 of the MA) typically corresponds to an untreated *MTB* culture exponentially growing *in vitro* under no stress, while the second channel corresponds to the experimental setup being studied. The FR of a gene is the relation between its expression in the two channels of the microarray, typically channel 2 divided by channel 1 (when the sample description indicated the contrary; FR's were inverted). By this way, if FR=1 for a particular gene, it means that that gene is expressed at the same level both in the environment studied and basal conditions. By contrast, a FR > 1 corresponds to an over expression of that gene in this situation and FR < 1 is shown by repressed genes.

A GEO-series contains samples of related experiments (for example, samples corresponding to exposure to a particular stimulus), with minor differences among samples regarding, for example, to exposure times or stress agents concentration levels. In addition, two different GEO samples often correspond to technical or biological replicates of the same experiment, and, similarly, in a single GEO sample, more than one replicate can be reported (for example, if more than one copy of the same gene has been seeded in a single microarray) or, for some genes, no report may be available.

As an example of GEO-sample we can mention the sample GSM886273, contained within the GEO-series GSE8839, in which the results of a genome-wide microarray expression profiling experiment are reported, for an experiment consisting on comparing

gene expression profile of a sample of *MTB*'s strain 1254 after exposure to hypoxia during two hours (channel 2) to the gene expression profile corresponding to the situation of exponential growth prior to exposure of the same strain to stress (channel 1). The series within which this sample is reported is GSE8839, that contains 131 samples corresponding to as many different experiments, all of them related to hypoxic conditions and/or exposure to oxidative stress.

Taken that into account, let us denote $FR(i,j,l)$ the FR corresponding to the l^{th} lecture associated to gene i in sample j. In the most common case, the only possible value for l is 1, although, in some cases, l will take more values (or none, in which case, we have gene i not being reported in sample j). So, we will name $N(i,j)$ the number of times that expression of gene i is reported in sample j (and therefore $l \in [1, N(i,j)] \leftrightarrow N(i,j) > 0$) The first task to accomplish is to group samples corresponding to repetitions (either technical or biological) of the same experiment, so as to define the average FR of a gene i in a experiment k $\langle FR \rangle_k(i)$ as follows:

$$\langle FR \rangle_k(i) = \frac{\sum_{j \in \{exp(k)\}} \sum_l^{N(i,j)} FR(i,j,l)}{N_k(i)} \tag{4.1}$$

where $\{exp(k)\}$ represents the set of samples associated to experiment k (ie. repetitions of the same experiment) and $N_k(i)$ is the total number of reports of gene i FR within such set:

$$N_k(i) = \sum_{j \in \{exp(k)\}} N(i,j) \tag{4.2}$$

Besides the average expression level of gene i in experiment k, we can also calculate its standard deviation:

$$\sigma_k(i) = \sqrt{\frac{\sum_{j \in \{exp(k)\}} \sum_l^{N(i,j)} (FR(i,j,l) - \langle FR \rangle_k(i))^2}{N_k(i)}} \tag{4.3}$$

Now, let us focus on a couple of proteins a and b between which an interaction is reported in the PPIN [4]. Making use of the average expression level of the genes codifying proteins a and b in experiment k we can define the "expression level" of the interaction $a \leftrightarrow b$ in that experiment, as follows:

$$\langle FR \rangle_k(a \leftrightarrow b) = \langle FR \rangle_k(a) \cdot \langle FR \rangle_k(b) \tag{4.4}$$

The biochemical meaning of that magnitude is related to the kinetic rate ρ at which the chemical reaction $A + B \rightarrow AB$ takes place:

$$\rho_{a-b} \simeq [a][b] \tag{4.5}$$

where we assume that the reaction between a and b to form the complex ab takes place at a stoichoimetry $1:1$, and that it occurs in a unique step, which is a reasonable assumption, provided that two-hybrid screens' positive results are primarily constituted by this kind of interactions [4].

It is straightforward to show that $\langle FR\rangle_k(a \leftrightarrow b)$ represents the FR associated to ρ_{a-b}, and so $\langle FR\rangle_k(a \leftrightarrow b) > 1$ means that the interaction $a \leftrightarrow b$ is taking place in the studied condition faster than in the basal state (and viceversa, for the opposite case).

Similarly, we can also define the associated typical deviation for these interaction expression levels as follows:

$$\sigma_k(a \leftrightarrow b) = \sqrt{(\langle FR\rangle_k(a) \cdot \sigma_k(b))^2 + (\sigma_k(a) \cdot \langle FR\rangle_k(b))^2} \tag{4.6}$$

And we can use equations 4.4 and 4.6 to define a criterion of statistical significance useful for discriminating whether a given link is present in a layer or not. This criterion is that the average FR of each interaction must lie at least one sigma over unity; i.e., the Z-score of the link FR:

$$Z_k(a \leftrightarrow b) = \frac{\langle FR\rangle_k(a \leftrightarrow b) - 1}{\sigma_k(a \leftrightarrow b)} \tag{4.7}$$

must be greater than one for the interaction to be considered within the layer corresponding to experiment k: $Z_k(a \leftrightarrow b) > 1$. in other words, and since that what we pursue is a weighted representation of the system, we assimilate the weight of a given link $a \leftrightarrow b$ within layer k to the interaction expression level $\langle FR\rangle_k(a \leftrightarrow b)$, provided that the interaction has passed the significance test $Z_k(a \leftrightarrow b) > 1$. More formally, we have:

$$w_k(a,b) = \begin{cases} \langle FR\rangle_k(a \leftrightarrow b) & \text{if } Z_k(a \leftrightarrow b) > 1 \\ 0 & \text{if } Z_k(a \leftrightarrow b) \leq 1 \end{cases} \tag{4.8}$$

The final result of this simple thresholding procedure is a multilayer network of 147 layers each of them containing data from an independent experiment. These experimental conditions can be grouped around 8 major groups: respiration arrest, cell wall damage, ion deprivation, oxidative stresses, DNA damage, dormancy models, macrophage ($M\phi$) infections and exposure to cholesterol) each of them, in turn, containing experiments grouped into sub-types. The original size of the network compiled by Wang et al. is 2907 nodes and 8042 interactions. Among the 147 layers of our system, we have, over the same nodes' set, as many as 388464 interactions, which makes an average of 2643 over-expressed links per layer.

Data necessary for that reconstruction come from 22 series and 609 samples within them. In addition to these sources, data from 27 additional series reporting MA experiments for *MTB* were analyzed, but no profitable information could be extracted from these. The reasons for that were two. On the one hand, we are only interested on the analysis of MA experiments reporting gene expression response of any experimental condition in relation to a basal, stress-free state consisting in bacterial cultures growing exponentially at the initial phases of a culturing procedure performed on standard culture media. Taken that into account, we must discard experiments reporting variations between two conditions none of which is assimilable to basal exponential growth. For similar reasons, as we aim to address stress response of circulating and reference strains of *MTB*, we discard samples corresponding to mutant strains (i.e. knockout

mutants). Finally, experiments related with exposure to antibiotics were also excluded from the analysis, as our aim is that of studying stimuli that *MTB* faces (or may face) in a natural way during its lifecycle.

Additionally, data stored in 6 additional series lacked the sufficient statistics so as to allow any interaction to pass the significance criterium. This occurs when all the samples in the series report un-repeated experiments, and so, there is, at most, one expression report per gene and experiment. This prevents us to define typical deviations and z-scores. Remarkably, this kind of problem prevented us to build any layer corresponding to acid Ph experiments, one of the major traits of phagosomal environment [232]. The reason for this problem is arguably related to the difficulties associated to make cultures grow under acid Ph [234].

Detailed information about the identity of the specific samples and series used in this study, types of stresses corresponding to each sample and experiment, etc. is available in the appendix (section 4.5).

4.2.2 Consensus multi-layer of stress response

Among the 147 layers of our original multi-layer system, very assorted experiments are represented. This includes *in vivo* infection assays and *in vitro* experiments, addressing response to single or multiple stresses or to the presence of certain signaling molecules thought to generate consistent and relevant transcriptional responses *in vivo*.

In this section, we will focus on addressing *MTB* defense to individual environmental stresses *in vitro*, more specifically, to those that have been associated to the principal characteristics of phagosomal environment. This includes oxidative stresses, nutrient deprivation, hypoxia and cell wall damage, beside others that we will not be able to analyze because of lack of relevant data (i.e. acid Ph). With this purpose in mind, we have first discarded layers corresponding to *in vivo* experiments, *in vitro* multi-stress dormancy models (as we aim to address the isolated response to single, not combined, stimuli), and *in vitro* exposure to signaling intermediates not causing cell-stress (e.g. cholesterol).

Once isolated the layers corresponding to *in vitro* exposure to environmental stresses, we have grouped them around six biologically meaningful categories:

- Hypoxia. This contains layers originally grouped under the label of "respiration arrest" in which the reason for such arrest is oxygen deprivation. It contains layers associated to different hypoxia models ("regular" hypoxia and Wayne's model of hypoxic growth), reareation experiments (channel 2 corresponds to initial hypoxia and channel 1 to final re-aireated sample).

- General oxidative stress. This corresponds to exposure to agents generating oxidative stress, like diamide and hydrogen peroxide H_2O_2.

- Exposure to NO. Nitric oxide is the principal damaging agent causing oxidative stress that *MTB* faces in the phagosomal environment. Nevertheless, it is known to interfere in the respiration pathways, behaving like an intermediate signal that

shares certain pathways with hypoxic conditions [233]. For these reasons, it is of high interest to consider NO exposure as an independent layer.

- Exposure cell wall damaging environments. These environments consist of exposure to surfactants that attack the integrity of lipids in the outer layers of the cell envelope, and exposure to extreme conditions of osmotic pressure.

- Ion deprivation. Corresponding to experimental setups in which *MTB* access to key nutrients (oligo-element ions) is impaired. This can be achieved through culturing *MTB* in ion deprived media (iron or phosphate, for example), or through introducing, for the case of iron, "iron scavengers", which are siderophores that sequestrate iron ions which can not be up-taken by *MTB*.

- Starvation: here we group layers corresponding to respiration arrest in which respiration is disrupted by nutrient deprivation. This can be achieved after culturing *MTB* in minimal, or deprived media or by leaving cultures reach the stationary phase, in which it is the very growth of the culture the ultimate reason for nutrient deprivation.

Once we have defined our main stress types, we aim at grouping all the layers of our previous dataset that correspond to each of these categories into single consensus layers. The resulting consensus multilayer system will thus contain just six layers, one for each of these major groups. Our purpose is to represent in any of these layers, only those interactions significantly represented in most of the experiments associated to that stress category. Taken that into account, we denote the consensus expression level of the interaction $a \leftrightarrow b$ in the category c as $\langle \bar{FR} \rangle_c(a \leftrightarrow b)$, and we have:

$$\langle \bar{FR} \rangle_c(a \leftrightarrow b) = \frac{\sum_{k \in \{cat(c)\}} \langle FR \rangle_k(a \leftrightarrow b)}{\bar{N}_c} \tag{4.9}$$

where $\{cat(c)\}$ represents the sets of layers of the original multi-layer system associated to the stress category c, and $\bar{N}_c$ is precisely the number of such layers. For now on, we will use bars to denote magnitudes referring to the reduced consensus multilayer. Additionally, it is worth noticing that, when an interaction has been discarded in a certain layer k of the original system we have that it does not contribute to the average, as we consider $\langle FR \rangle_k(a \leftrightarrow b) = 0$ for that case.

Additionally, we perform a similar average for the typical deviations:

$$\bar{\sigma_c}(a \leftrightarrow b) = \frac{\sum_{k \in \{cat(c)\}} \sigma_k(a \leftrightarrow b)}{\bar{N}_c} \tag{4.10}$$

and finally we define the corresponding Z-score for each interaction within each of the layers of the consensus multi-layer system as follows:

$$\bar{Z}_c(a \leftrightarrow b) = \frac{\langle \bar{FR} \rangle_c(a \leftrightarrow b) - 1}{\bar{\sigma}_c(a \leftrightarrow b)} \tag{4.11}$$

It is worth mentioning that it is not the only way to perform such filtering procedure, but there exist other alternatives, like averaging Z-scores or only admitting links which are present in a fraction of layers of the original system greater than a certain threshold. We decided to use this one for two reasons. On the one hand, it favors a moderate enrichment of consensus genes and interactions, (i.e. those genes whose presence is recurrent in most the experiments within a category). Another interesting feature is that it provides with the same statistical relevance to each of the experiments within a category, regardless the number of samples corresponding to each experiment.

Once again, we will associate interaction expression levels to link weights only when surpassing the overexpression threshold ($\langle \bar{FR} \rangle_c(a \leftrightarrow b) > 1$) with a significance of at least one sigma ($\bar{Z}_c(a \leftrightarrow b) > 1$). So we have that the interaction weight of the interaction $a \leftrightarrow b$ within layer c in the consensus network $\bar{w}_c(a, b)$ is defined as follows:

$$\bar{w}_c(a,b) = \begin{cases} \langle \bar{FR} \rangle_c(a \leftrightarrow b) & \text{if } \bar{Z}_c(a \leftrightarrow b) > 1 \\ 0 & \text{if } \bar{Z}_c(a \leftrightarrow b) \leq 1 \end{cases} \tag{4.12}$$

The result was a reduced six-layer system containing 1593 interactions, which is represented in figure 4.3.

4.2.3 Layer similarities in the complete multi-layer system

As we said before, among the 147 layers that constitute our original multi-layer system, some correspond to *in-vitro* experiments, while others are associated to *in-vivo MTB* infection trials performed on different types of $M\phi$s (human or animal). In addition, there are experiments performed using different *MTB* strains, and, of course, layers can be also grouped according to the general type of environmental stimuli, as it has been discussed previously.

A first question to address around this system is up to what extent the topologies of each of the layers are similar to each other, and, more relevantly, how these similarities correlate with the principal attributes of the correspondent experiments listed before.

So as to elucidate this question, we have selected two measurements of similarity among couples of layers: the coherence between the nodes' strength ranks of the layers, and the conditioned probability for a random link to exist in one layer provided it does exist in the other. Then, we linearly transform these similarity measurements so as to get normalized distances, i.e. measures within the interval $[0, 1]$, in which 0 means maximum similarity and 1 maximum dissimilarity. After transforming them we perform a hierarchical clustering analysis in order to get from there an overview of how topological similitude among layers is related to the different attributes of the associated experiments.

Strength rank coherences

In order to compute the first of these measures, we recall the definition of node strength in a given layer α as the sum of the weights of all the interactions it is involved in within that precise layer:

$$s_\alpha(i) = \sum_j w_\alpha(i,j) \tag{4.13}$$

From that definition, we can build a ranking for each layer, by decreasingly ordering the nodes according to their strengths, and placing absent nodes (i.e. those for which the strength is zero) at the end of the rank. Once done so, we are able to compare any couple of layers according the coherence of their strengths ranks, using the Kendall coefficient $r(\alpha, \beta)$:

$$r(\alpha,\beta) = \frac{2(c(\alpha,\beta) - d(\alpha,\beta))}{n(n-1)} \tag{4.14}$$

where $c(\alpha, \beta)$ is the number of pairs of nodes which are concordant between the strength ranks of layers α and β and $d(\alpha, \beta)$ the number of discordant pairs. Couples of nodes are concordant if they share the same hierarchy in the ranks of both layers (i.e., if we talk about nodes i and j, they constitute a concordant pair if $s_\alpha(i) > s_\alpha(j)$ and $s_\beta(i) > s_\beta(j)$ or if $s_\alpha(i) < s_\alpha(j)$ and $s_\beta(i) < s_\beta(j)$), and they are discordant in the strictly opposite case (i.e., i and j constitute a discordant pair if $s_\alpha(i) > s_\alpha(j)$ and $s_\beta(i) < s_\beta(j)$ or if $s_\alpha(i) < s_\alpha(j)$ and $s_\beta(i) > s_\beta(j)$). In case two nodes are equally ranked in some of the layers ($s_\alpha(i) = s_\alpha(j)$ or $s_\beta(i) = s_\beta(j)$) they are nor concordant neither discordant. This metrics takes values in the interval $[-1, 1]$, corresponding -1 to perfectly inverted ranks and 1 to identical ranks. The case $r(\alpha, \beta) = 0$ indicates that strength ranks of layers α and β are statistically independent. In order to transform Kendall coefficients in a distance-like metrics whose image interval is $[0, 1]$, we perform he following transformation:

$$R(\alpha,\beta) = \frac{(1 - r(\alpha,\beta))}{2} \tag{4.15}$$

which is equal to zero when $r(\alpha, \beta) = 1$ (identical layers, minimum distance) and one when $r(\alpha, \beta) = -1$ (opposite layers, maximum distance).

Conditioned probabilities

Let $L(\alpha)$ be the number of links with weight greater than zero in layer α. Furthermore, let us denote $\alpha \cap \beta$ as the intersection between layers α and β (i.e. the network that we recover by considering links and nodes present only in both layers, disregarding link weights). Then, the conditioned probability for a random link $a \leftrightarrow b$ to exist in layer β provided that it exists in layer α is defined as follows:

$$P(w_\beta(a,b) > 0 | w_\alpha(a,b) > 0) = \frac{L(\alpha \cap \beta)}{L(\alpha)} \tag{4.16}$$

So as to recover a distance-like symmetric measure from that conditioned probability, we define:

$$\Pi(\alpha,\beta) = \begin{cases} 1 - P(w_\beta(a,b) > 0 | w_\alpha(a,b) > 0) & \text{if } L(\alpha) \leq L(\beta) \\ 1 - P(w_\alpha(a,b) > 0 | w_\beta(a,b) > 0) & \text{if } L(\alpha) > L(\beta) \end{cases} \tag{4.17}$$

which i) guarantees that $\Pi(\alpha, \beta) = \Pi(\beta, \alpha)$ for all (α, β) and ii) verifies $\Pi(\alpha, \beta) = 0$ when the smaller layer is totally included in the bigger one (minimum distance) and $\Pi(\alpha, \beta) = 1$ when $\alpha \cap \beta = 0$ (maximum distance).

Hierarchical clustering analysis

From the pairwise metrics defined in equations 4.14 and 4.16, we get two distance matrixes of inter-layer topological dissimilarity in 4.15 and 4.17, what we analyze using hierarchical clustering. The essential objective of clustering analysis algorithms is to identify groups of elements (in our case, layers) that are at lower distances than from elements belonging to other groups. In order to do so, and based on the distance matrix, we start by identifying the couple of elements being at lowest distance among all pairwise distances in the matrix. These two elements are then associated to a single one (i.e. a cluster), after which the process is iterated until all elements have been analyzed.

From the second iteration on, the algorithm must calculate distances between a cluster and a node (or directly, between two clusters). As a cluster is an "extended", not punctual object, there exist different meaningful ways to compute distance among clusters, which yield different algorithms for clustering analysis. In this Thesis we have used average linkage algorithm (also called weighted pair-group method). According to this algorithm, we define the distance between two clusters as the average distance between each of the members. This method does not provide very large or very small clusters and it is largely used for this type of representations [235].

4.2.4 Genes' roles in stress-response

Once the methods used to disentangle relations among layers have been presented, we aim to analyze the role of single nodes (genes) on our system. In order to do so, we will follow the methodology proposed by Battiston et al. [231], directly inspired by the previous work by Guimerá and Amaral [236]. This implies the characterization, in the stress-response consensus multi-layer network, of two topological metrics for each node, the node overlap, and, for example, the participation coefficient.

The overlap of a gene in the reduced multi-layer network is defined as the sum of the node's strengths in all the layers:

$$\bar{o}(i) = \sum_{c} \bar{s}_c(i) \tag{4.18}$$

This quantity measures the total sum of all the weights of every interaction the node take part in within any of the layers of the consensus system. Keeping in mind that in that system each layer illustrates the PPIN after exposure to a different type of stress condition, all of which are thought to be cardinal characteristics of the phagosomal environment, the overlap of a gene can be interpreted as a measure of its global relevance in mediating generic stress responses in *MTB*.

Admittedly, this general measure of node's importance can be complemented with the participation coefficient $p(i)$, defined as follows:

$$\bar{p}(i) = \frac{\mathcal{L}}{\mathcal{L}-1}\left(1 - \sum_{c=1}^{\mathcal{L}}\left(\frac{\bar{s}_c(i)}{\bar{o}(i)}\right)^2\right) \tag{4.19}$$

where $\mathcal{L} = 6$ is the number of layers of the consensus system. It is easy to show that $p(i) = 0$ implies that the node is only present in one of the layers, while a value of $p(i) = 1$ means that the presence of a gene –no matter how weak or strong– is homogeneously distributed among all the layers in the system.

In our context, the participation coefficient can be interpreted as a measure of how "specialist" a gene is in what regards the type of stresses that it primarily responds to: high values of p will correspond to generic-stress response genes, and low values of p will correspond to genes responding only after a particular type of stress.

4.3 Results

4.3.1 Layer clustering analysis: insights on TB infection models

The outcome of the hierarchical clustering analysis of the original multi-layer system is captured in dendrograms depicted in figures 4.1 and 4.2. These dendrograms constitute a valuable graphical resource to analyze the clustered structure of the datasets, and to distinguish layers which share common topological profiles from those that present more divergent topologies. Figure 4.1 represent the dendrogram corresponding to nodes' strength ranks coherences and figure 4.2 represents the outcome of the clustering analysis based upon conditioned probabilities of links co-existence.

Below the dendrograms, four color bars are plotted, which represents experiments classifications according to different criteria. By comparing the color bars with the structure of the dendrograms, we can get a global understanding of what aspects of the experiments exert deeper influences on PPIN topologies.

First of all, it worths noticing that both figures show strong similarities in what regards the overall structure. On the one hand, layers corresponding to *in-vivo* assays of murine derived $M\phi$s infection appear consistently grouped in both figures. At the same time, most in-vitro experiments tend also to form coherent clusters. Finally, both dendrograms present an outlier region corresponding to control and low stress exposure in vitro experiments (second color-bar, in black). Average distances among these control-like experiments are highest, as well as the distances from them to the rest of the layers. This is not surprising, as these layers correspond to experiments in which the stress condition is absent, or too mild to generate substantial transcriptional responses, which translates into high levels of noise and randomness. Remarkably, most experiments associated to low availability of some key ions (i.e. phosphate and Fe^{2+}, GEO series GSE14840, GSE1642 and GSE8732) lie in this area in both figures (yellow stripes, fourth color bar), suggesting that, at least at the doses applied in these experiments, reduced ion availability generates a modest transcriptional response at

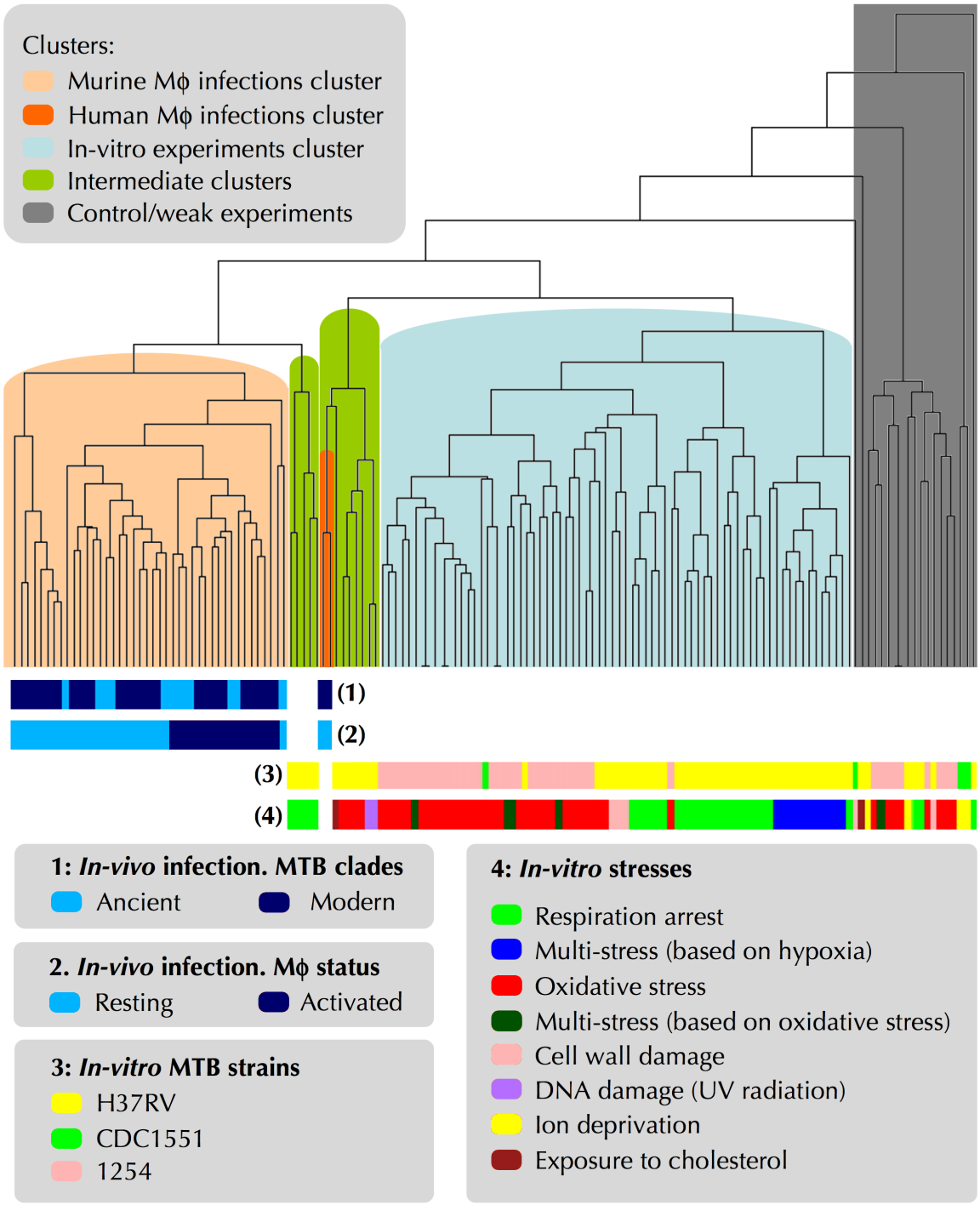


FIGURE 4.2: Multilayer PPIN: layers' clustering dendrogram according to strength ranks. Each branch in the dendrogram represents a version of the original PPIN [4], transformed using the transcriptional response associated to a different experiment obtained from GEO database [5]. Different experiment attributes are indicated through different color codes. Dendrograms contructed using software R (packet dendextend)

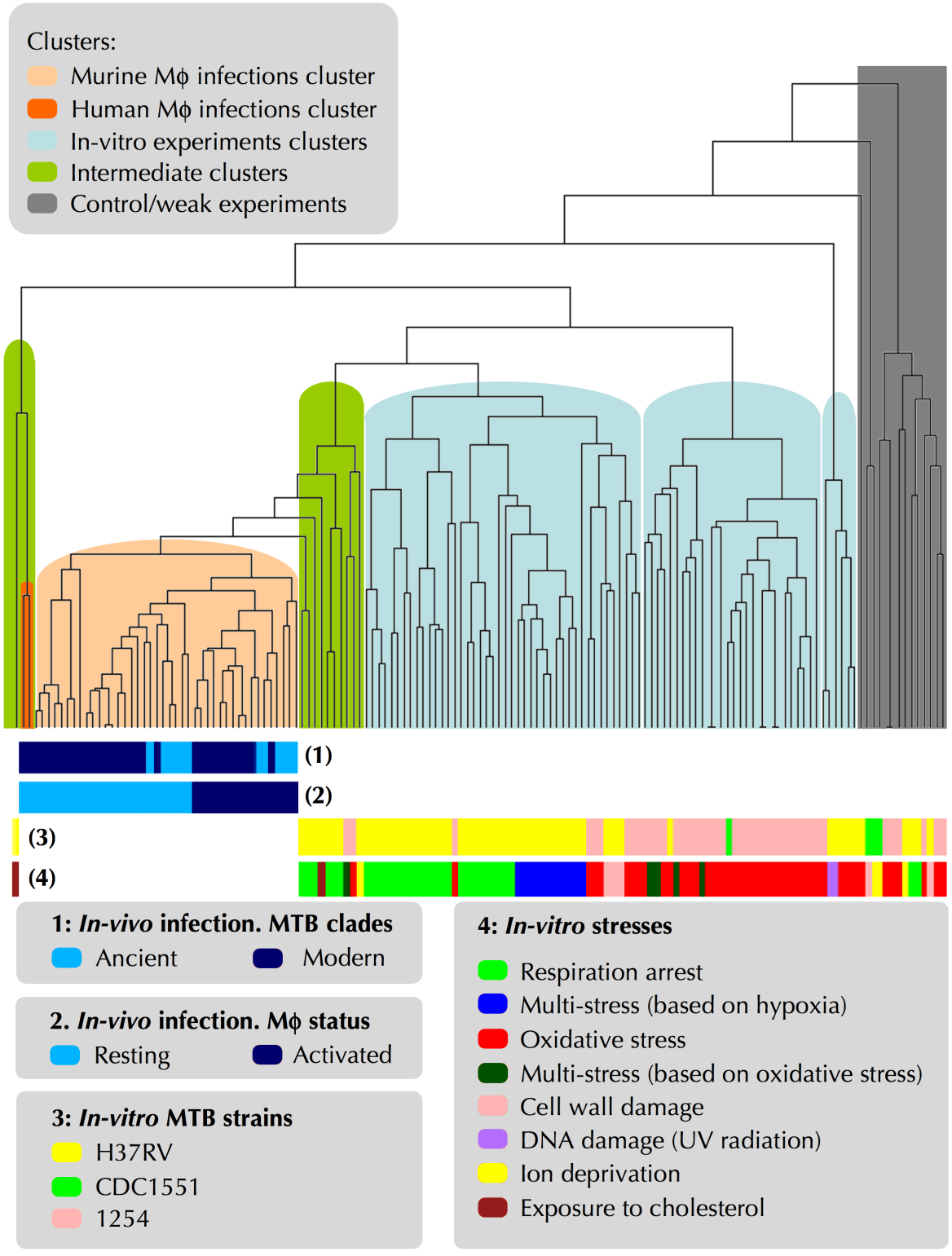


FIGURE 4.3: Multilayer PPIN: layers' clustering dendrogram according to conditional probabilities of links coexistence between layers. Each branch in the dendrogram represents a version of the original PPIN [4], transformed using the transcriptional response associated to a different experiment obtained from GEO database [5]. Different experiment attributes are indicated through different color codes.

genome wide level.

However, additional nuances to the above general outline arise after a deeper analysis of the dendrograms. On the one hand, it is important to note that, if murine-derive $M\phi$ infection experiments tend to group together, infection assays performed on human $M\phi$s (only two) split from them. Indeed, for some layers corresponding to *in-vitro* experiments ("intermediate clusters" on green background), phylogenetic distances to either murine-derived or –more remarkably– human $M\phi$s infection layers are lower than the distance between the clusters corresponding to both types of *in-vivo* infection models. In this line, we have that most *in-vitro* layers are at least as similar to human $M\phi$s layers as murine $M\phi$s infection layers, and, in addition, the closest layer to human $M\phi$s infection experiments is, in both dendrograms, an *in-vitro* experiment corresponding to 24 hours of exposure to cholesterol (taken from GSE13978). Similarly, several *in-vitro* layers appear in the figures as the closest parents of murine $M\phi$s infection assays, of which, re-aireation after hypoxia experiments appears in both dendrograms.

The relevance of these results, from a biological perspective, is twofold. First, this indicates that, for what respects the reach of our observations, murine infection assays does not offer a particularly better proxy of human $M\phi$ infection than certain in-vitro setups. Furthermore, among all the layers corresponding to *in-vitro* experiments, those corresponding to re-aireation experiments and exposure to cholesterol emerge from our analysis as the most promising candidates to constitute more adequate *in-vitro* models of intracellular *MTB* infection. This is of particular relevance in our case, as *in-vitro* modeling of phagosomal environment has constituted an intense subject of debate during the last decade [232]. Remarkably, the appearance of cholesterol as a putative key signal mediator able to generate a global transcriptional response highly similar to that followed by *MTB* infection in human $M\phi$s results particularly suggestive at the light of some recent works that have proposed a relevant role for this molecule and its metabolic pathways in the adaptation of *MTB* to the intracellular environment [117, 227, 119].

Beyond these principal results, further analysis of how the layers regarding *in-vitro* and *in vivo* experiments organize deserve attention.

In-vitro layers tend to group according the type of stress (color bar 4), with two (or three) major clusters of experiments standing out, corresponding to the two stress types more represented in GEO database: respiration arrest experiments (in green in color bar 4) and oxidative stress (in red). Associated to these stresses, multi-stress dormancy models primarily based on oxidative stress (dark green) or hypoxia (dark blue) group together to each of the two main stresses mentioned. Other less represented kinds of stress tend to consistently group too, as cell wall damage (pink) and DNA damage (violet). However, the structure of the data does not allow us to assure that the reason of the emergence of this structure was the type of stress. Indeed, if we focus on the two more represented types of stresses in the dendrogram we can see how almost all the layers corresponding to respiration arrest experiments (green, fourth color bar) have been conducted using H37RV strain (yellow in the third bar), while, for the case of oxidative stress experiments (red in the fourth bar) we have an overwhelming majority of experiments conducted on *MTB* strain 1254 (pink in the third bar). This means

that the differentiation between the mentioned two clusters can be either a consequence of the stresses associated to each group of experiments or to the strains on which the experiments were performed.

Regarding the inner structure of the cluster of *in vivo* experiments performed on murine-derived $M\phi$s, the situation is much clearer. First, experiments on activated $M\phi$s (from GEO series GEO21113) and experiments on resting $M\phi$ (from GEO series GSE35362 and GEO21112) split almost perfectly, suggesting that phage activation is a more relevant trait of the systemic response of *MTB* after infection than the specific clinical isolate (*MTB* strain) within this group of layers. Additionally, the experiments on resting $M\phi$s also split in two major clusters in each graph: one for experiments coming from GSE35362 (8 first layers of the orange cluster of murine $M\phi$ infections, not marked in the dendrograms), and other for experiments from GSE21113. This is likely caused by different experimental protocols followed in both series (for example, infection times are different in both the series mentioned). Finally, within the division according to $M\phi$ activation, the evolutionary clades which the strains used in each of the infection experiments belong to can be distinguished, suggesting that our approach is also sensitive to differences in gene expression attributable to a genetic origin as a second (or third) order effect.

4.3.2 *MTB* genes roles in response to environmental stress

The consensus multi-layer system representing PPINs after different environmental stresses contains 1593 interactions distributed among six layers, which corresponds to 265 interactions per layer. In this consensus system only 606 genes of the 2907 from Wang's original network take part. The six layers of this system are represented in figure 4.4, where nodes are colored according to functional ontologies extracted from tuberculist database [237], and node sizes correspond to genes' strengths.

Looking at figure 4.4 two relevant features of the system are evident at a glance. First, if we compare it to the original network by Wang et al., we have that hubs in that network (i.e. nodes with more connections in the not-weighted network represented in 4.1) does not correspond with principal hubs in any of the layers of our stress-response consensus system. The interpretation for this is simple: the activity of proteins with the ability of interacting with large amounts of different proteins is not favored under stress conditions. Instead of that, proteins showing greatest strengths in our stress-related layers (i.e. those whose activity is most enhanced because of stress exposure) present intermedium and low degrees in the original dataset. Additionally, we see how the overall structure is highly similar in all the layers, with a reduced set of 6-8 hubs belonging to the category of cell wall related proteins standing out among all the rest. Remarkably, we also observe the presence of nodes that are hubs only in specific layers. The most evident example is, in the layer of NO exposure, formed by a couple of genes: Rv1738, of unknown function and Rv2925c, involved in information pathways.

In order to obtain a more rigorous picture in what concerns the role of these hubs and all the other proteins, we have represented, in figure 4.5 the overlap of each gene versus its participation index. The plot has been divided in three regions so as to

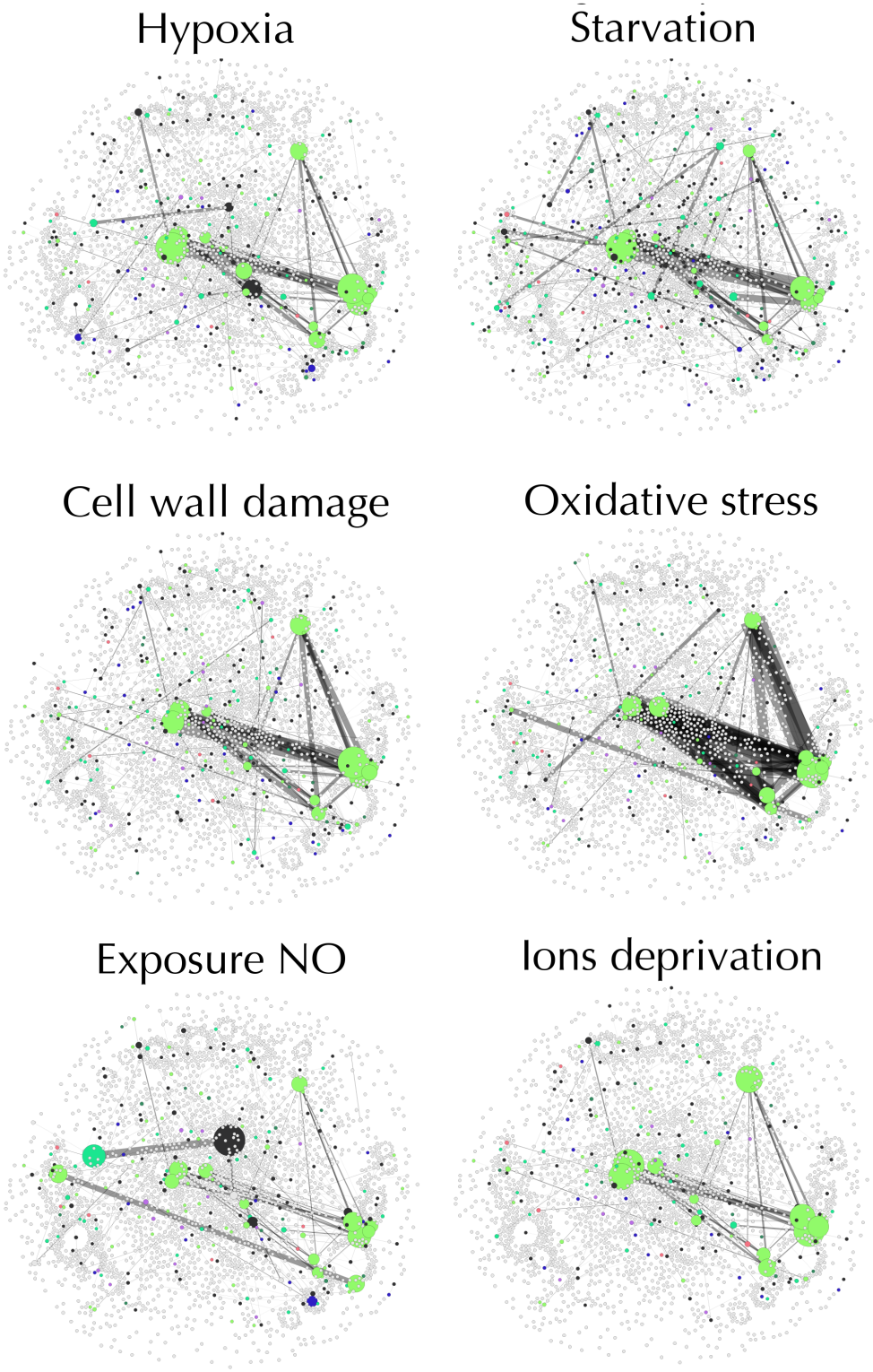


FIGURE 4.4: Stress response consensus multi-layer system. In each layer, only those genes and interactions systematically overexpressed in a majority of the experiments associated to a single type of stress will appear.

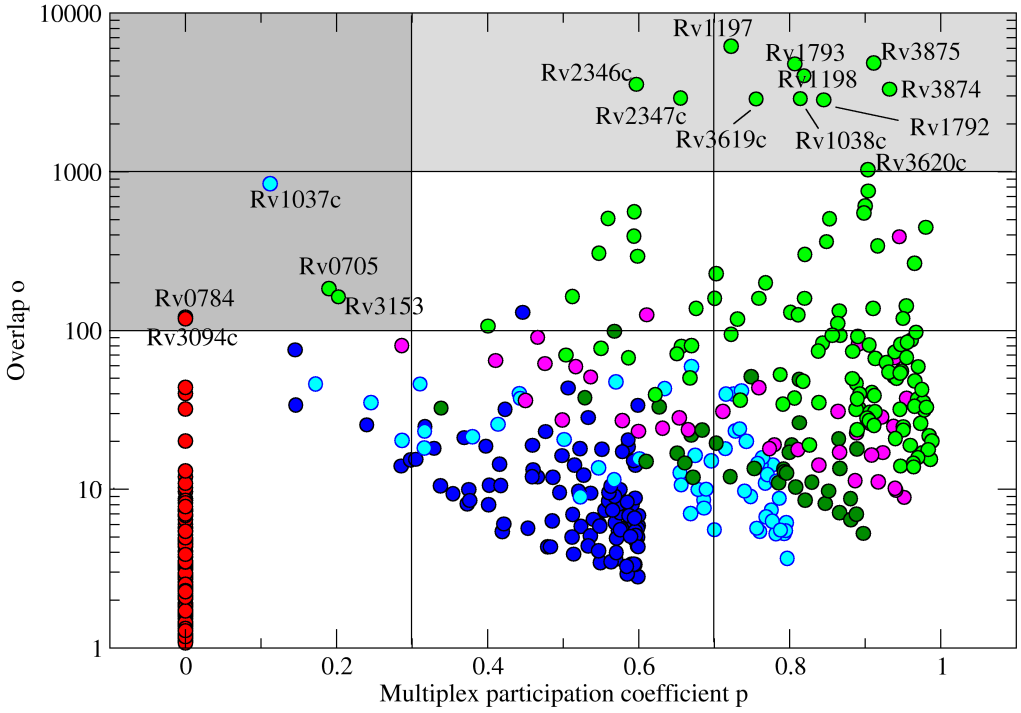


FIGURE 4.5: Overlap vs. participation coefficient for genes present in the stress-response multi-layer system. Two regions are highlighted: that of specialist hubs (dark grey) and that of generic hubs (light grey). All the genes classified in the second group (plus Rv1037c) form six pairs of adjacent genes analogous to ESAT-6 and cfp10.

identify atypical genes that may have important roles in this context. This classification distinguishes three types of nodes: regular nodes (white background in Figure 4.5), specialist hubs (dark-grey background) and generic hubs that respond equally to every stress studied (light grey). All the nodes appearing in the light grey or dark-grey regions were identified.

Furthermore, nodes are colored according to the number of layers they appear in, which defines the position of each node within the graph. Nodes that are absent in some layers cannot reach high values of p. In accordance with the definition of the participation coefficient, it is easy to show that nodes that appear in 1 to 5 layers can, at most, have values of equal to 0, 3/5, 4/5, 9/10 and 24/25 respectively, if the overlap of a node is homogeneously distributed among all the layers where it is present. On the other hand, "peaked" patterns emerge for each group, so it implies the emergence of "forbidden regions" for each one in the region of low p and low overlap. This is a

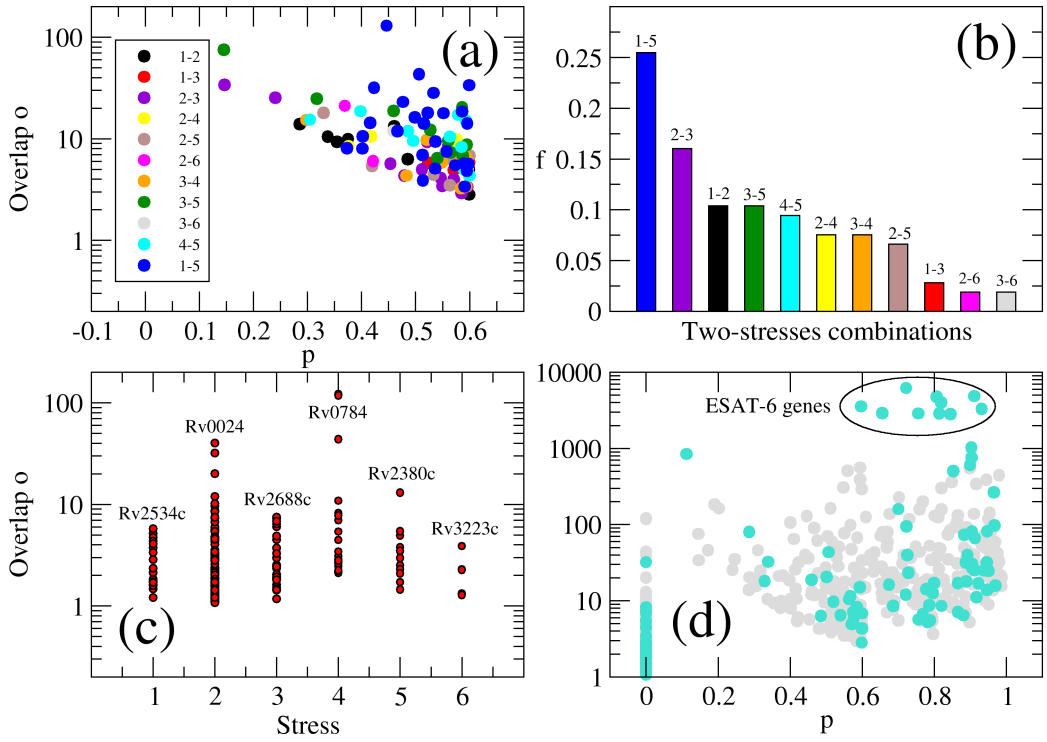


FIGURE 4.6: Overlap vs. participation coefficient analysis. A. For nodes appearing only in two layers, we represent the specific couple of layers in which they appear. Layers numbers: Hypoxia (1), nutrient starvation (2), cell wall damage (3), ions deprivation (4), NO exposure (5) and generic oxidative stress (6). B. Frequencies of association of each couple of layers. Hypoxia-NO exposure (layers 1 and 5) appears as the most frequent association. C. Overlap of genes appearing in only one layer. genes with maximum overlap in each layer are identified. 4. Cell wall related proteins highlighted in the graph. ESAT-6 protein family belongs to this functional cathegory.

straightforward consequence of the definition of these parameters. Let us consider a node preliminarily present in three layers (blue in the figure), as an example to illustrate this. A low p means that the overlap of the node is heterogeneously distributed among layers, and as a consequence, it is expected to have little strength in one of the layers, much lower than in the rest. If, besides, the overall overlap is reduced, the strength of the node in the weakest layer will not surpass the over-expression threshold, so it will disappear from the layer "jumping" from the group of nodes present in the three layers (light blue group) to the group present in only two (dark blue group). The graphical display of the network in figure 4.3 shows that more relevant nodes in the network are homogeneously distributed among all the layers. Surprisingly, all these genes, labelled as cell-wall related, belong to to the Early Secreted Antigen Target of 6kDa (ESAT-6) family, one of the most relevant antigens secreted by the bacterium (see Figure 4.5).

For those nodes that appear only in two layers, it would be interesting to know

whether there is any combination of two layers with more nodes in common than others, as this could indicate that some pair of stresses shares tighter common transcriptional response mechanisms than others. As Figure 4.6 (panels (a) and (b)) show, a high proportion of the nodes which are present only in two layers are in hypoxia and NO-exposure stress layers. This is not surprising, as the association between the bacterial responses to these stimuli is well known [238]. However, the different amount of genes that appear in each layer should be taken into account in a deeper analysis of these associations. Notwithstanding that, the six layers present a rather homogeneous number of nodes (267 in hypoxia, 402 in nutrient starvation, 297 in cell wall damage, 250 in ions deprivation, 262 in NO exposure and 192 in oxidative stress) and hypoxia (267) and NO layers (262) present modest number of nodes, which backs up the biological implications of the outcome displayed in figures 4.6, panels (a) and (b).

Likewise, proteins that appear only in one layer have been divided in accordance with the layer they belong to in figure 4.6 (c). Proteins with the greater overlap value of each layer were identified.

4.4 Discussion

In this chapter, we have analyzed the transformations suffered by the PPIN of *MTB* when dealing on different *MTB* strains facing different environmental conditions, both *in vitro* or *in vivo*.

From the analysis of our original multi-layer system, a first conclusion is that it is precisely the type of experiment performed -*in vitro* and *in vivo*– the factor more strongly affecting PPIN topologies. This has, however, the crucial exception represented by the existence of relevant differences between human and animal infection models, which are bigger than the differences between some of these *in vivo* models of *MTB* infection and certain *in vitro* experiments. This observation can be interpreted as a sign of the otherwise well known limitations of animal models in certain contexts of TB research, something that has constituted a frequent matter of discussion [239]. *In vitro* experimental setups that produced more similar PPINs to *in vivo* infection models include re-aireation and exposure to cholesterol.

In what regards re-aireation (GEO series GSE21590), experiments compare gene expression profiles of dormant bacteria (7 days under hypoxia) to those obtained after further ventilation during different times [240], which makes *MTB* cultures return to exponential growth. Folding rates are calculated dividing the dormant expression levels by subsequent exponential growth profiles obtained at the end of the experiment. Thus, it is not surprising that these reareation assays are able to mimick, at the level of the PPIN, *in vivo* infection models, as hypoxia-induced dormancy is known as one of best stablished models for *in vivo* infection [89]. Nevertheless, it is remarkable the fact that, among all experiments related to respiration arrest (green, in the fourth color bar of figures 4.2 and 4.3), reareation experiments appear systematicly closer to $M\phi$s infection assays. This seems to suggest that the process of entrance and ulterior exit from hypoxia-induced dormancy may not be totally (or instantaneously) simmetric, as

thoroughly studied precisely in [240].

Furthermore, the emergence of cholesterol as the *in vitro* stimulus closest to infection assays on human $M\phi$s is suggestive. Cholesterol metabolism pathways have been identified as a central piece of *MTB* adaptation to phagosomal environment. As we discussed in the introduction, host's cholesterol is used by *MTB* both as an energy source and as a precursor of more complex lipids that are essential for bacterial survival and virulence [117, 119]. These are essential ingredients of a global shift in *MTB* metabolism that substitutes carbo-hidrates to lipids as major carbon and energy sources [227]. The results of our analysis suggest that cholesterol exposure may constitute a relevant signaling mediator for the adaptation of *MTB* to intracellular conditions, able to generate a global transcriptional response that causes PPIN to closely resemble those corresponding to *in vivo* infection assays performed in human $M\phi$s to a superior extent than any other *in vitro* setup.

In what regards genes' roles analysis, our method has identified a group of 11 genes as generic hubs, appearing in all the layers and having both the greatest overlaps and high values of the participation coefficients. These eleven genes –along with Rv1037c, that appears in three layers–, conform six gene pairs of the family of ESAT-6 secreted antigens. This family of genes, as it has been discussed in the introduction to this thesis, encode small proteins (molecular weights of the order of 5-10 kDa) that are found in adjacent co-transcribed pairs across the mycobacterial genome. These pairs of proteins form hetero-dimers that are secreted by *MTB* to the intracellular environment. That is the case of the 12 genes (11 generic hubs plus Rv1037c) found in figure 4.6: Rv1037c-Rv1038c, Rv1792-Rv1793, Rv1197-Rv1198, Rv3619c-Rv3620c and Rv2346c-Rv2347c and Rv3874-Rv3875. Among these gene pairs, the last one –Rv3874-Rv3875, commonly denoted as ESAT-6 and cfp10 (culture filtrate protein of 10kDa)–, have received most of the attention during the last years [241, 242, 191, 243] for his strong immunogenicity and pathogenic activity as major *MTB* antigens, which has motivated its use for the development of novel vaccines [244]. ESAT-6 and cfp10 form hetero-dimers in vivo, which allows their correct secretion through a secretion system that is specific to this family of proteins [245]. Once secreted, cfp10 –whose principal role is thought to be related to ESAT-6 stabilization– splits, and ESAT-6 develops a strong cytolitic activity, opening pores in host's lipidic membranes [246]. In addition, BCG vaccine does not produce ESAT-6, [247, 248], which has indeed allowed the advent of new diagnostic tools based on ESAT-6 that, in principle, are able to discriminate between infected individuals and people vaccinated with BCG [249]. As an example of its central role in immunogenicity, one of the principal features of the vaccine candidate *MTB*VAC is that, through disruption of the global regulator PhoP, the proteins of the secretion system of ESAT-6 become insufficiently expressed, which turns into a *MTB* strain that, yet expressing ESAT-6 (which provides immunogenicity), is not able to secrete it into the medium (with reduces its pathogenicity) [129].

The fact that our method is able to capture the central importance of the gene pair ESAT-6/cfp10 constitutes an encouraging argument evidencing the need of systemic analysis of global cellular adaptation beyond standard transcriptional characterization. Nevertheless, our method also predicts a comparable role for five additional gene pairs

parents to ESAT-6/cfp10. These proteins have not received a comparable attention, and their biological role is assumed to be analogous to that of the pair ESAT-6/cfp10, being their essential purpose that of providing a source for antigenic diversity to the pathogen that could help it to evade host' immune response [435, 436]. Summarizing, our results suggest that the proteins mentioned could play a role comparable to their most relevant member ESAT-6 and cfp10; an hypothesis that deserves further exploration.

All these results, taken together, indicate that the methodology proposed in this chapter constitutes a promising approach to the observation of cellular adaptive responses to different conditions at a systemic scale. The observation of the transformations suffered by the PPIN of *MTB* under different environments, specially under different stresses, provides relevant biological insights of the underlying adaptation mechanisms experienced by the pathogen and helps to discriminate most relevant ingredients in such processes by proposing putative key genes and signaling routes that will need further experimental validation. Admittedly, our methodology relies on the correspondence of gene expression levels measured in micro arrays and protein concentration, and for this reason further de-correlations between mRNA and protein concentration due to post-transcriptional regulatory mechanisms can not be addressed from the data we manage.

Further work in this line is currently under progress, whereby we aim to elucidate, among other relevant aspects of our methodology, the following questions

- which part of the topological transformations suffered by PPIN as a consequence of environmental shifts is imputable to transcriptional changes and which part to the underlying topology of the PPIN?
- Is the convergent response of all the stresses listed a feature showed by other bacteria with similar (or different) life-styles?
- Are the topologies associated to other stimuli not associated to stress (e.g. cholesterol) similar to those of the stress consensus multi-layer system?
- Does the biological role of ESAT-6-like proteins present a comparable relevance in pathogenicity and immunogenicity to that of the couple ESAT-6/cfp10?

4.5 Appendix

4.5.1 Discarded data

We accessed GEO in March 2014 and collected all the series corresponding to microarray experiments conducted in *MTB*. We discarded the following 21 entire series for the reasons adduced in the main text (i.e. inadequate basal reference channels, mutants and/or antibiotics experiments): GSE7963, GSE17640, GSE29160, GSE31732, GSE32076, GSE32157, GSE32178, GSE42151, GSE54289, GSE40917, GSE45811, GSE24035,

GSE24045, GSE46212, GSE43466, GSE9776, GSE10897, GSE12364, GSE16626, GSE42293, GSE26166.

On the contrary, we recovered useful information from the following 22 series: GSE14840, GSE8732, GSE1642, GSE365, GSE16146, GSE8839 GSE8689, GSE15976, GSE6750, GSE8664, GSE50159, GSE10391 GSE5977, GSE21590, GSE8786, GSE9331, GSE8829, GSE13978 GSE6209, GSE35362, GSE21112, GSE21113.

Notwithstanding that, not all the samples within these series were useful for our analysis. In the following table, we enlist the samples discarded from each of the series used:

GEO-sample	Description
GSM5277	0.2 μg/mL MMC (8h) vs control
GSM5281	0.2 μg/mL MMC (12h) vs control
GSM5285	0.2 μg/mL MMC (4h) vs control
GSM5286	0.2 μg/mL MMC (6h) vs control
GSM5295	0.2 μg/mL MMC (8h) vs control
GSM5299	0.2 μg/mL MMC (4h) vs control
GSM155430	Rv-Control(1) vs. Rv-Δ981[mpr]-control(1)
GSM155431	Rv-Control(2) vs. Rv-Δ981[mpr]-control(2)
GSM155432	Rv-Control(3) vs. Rv-Δ981[mpr]-control(3)
GSM155433	Rv-Control(4) vs. Rv-Δ981[mpr]-control(4)
GSM155434	Rv-Control(5) vs. Rv-Δ981[mpr]-control(5)
GSM155435	Rv-Control(6) vs. Rv-Δ981[mpr]-control(6)
GSM155436	Rv-Δ981[mpr]-control(1) vs Rv-Δ981[mpr]-SDS(1)
GSM155437	Rv-Δ981[mpr]-control(2) vs Rv-Δ981[mpr]-SDS(2)
GSM155438	Rv-Δ981[mpr]-control(3) vs Rv-Δ981[mpr]-SDS(3)
GSM155439	Rv-Δ981[mpr]-control(4) vs Rv-Δ981[mpr]-SDS(4)
GSM155440	Rv-Δ981[mpr]-control(5) vs Rv-Δ981[mpr]-SDS(5)
GSM155441	Rv-Δ981[mpr]-control(6) vs Rv-Δ981[mpr]-SDS(6)
GSM155442	Rv-SDS(1) vs Rv-Δ981[mpr]-SDS(1)
GSM155443	Rv-SDS(2) vs Rv-Δ981[mpr]-SDS(2)
GSM155444	Rv-SDS(3) vs Rv-Δ981[mpr]-SDS(3)
GSM155445	Rv-SDS(4) vs Rv-Δ981[mpr]-SDS(4)
GSM155446	Rv-SDS(5) vs Rv-Δ981[mpr]-SDS(5)
GSM155447	Rv-SDS(6) vs Rv-Δ981[mpr]-SDS(6)
GSM214900	H37Rv wild type vs. sigE mutant
GSM214901	H37Rv wild type vs. sigE mutant
GSM214902	H37Rv wild type vs. sigE mutant
GSM214909	Exposed to 0.05% SDS sigE mutant
GSM214910	Exposed to 0.05% SDS sigE mutant
GSM214911	Exposed to 0.05% SDS sigE mutant
GSM214912	Exposed to 0.05% SDS sigE mutant
GSM214913	Exposed to 0.05% SDS sigE mutant

GEO-sample	Description
GSM214914	Exposed to 0.05% SDS sigE mutant
GSM215378	H37Rv wild type vs. sigH mutant
GSM215379	H37Rv wild type vs. sigH mutant
GSM215380	H37Rv wild type vs. sigH mutant
GSM215387	Exposed to Diamide 5 mM sigH mutant
GSM215388	Exposed to Diamide 5 mM sigH mutant
GSM215389	Exposed to Diamide 5 mM sigH mutant
GSM215390	Exposed to Diamide 5 mM sigH mutant
GSM215391	Exposed to Diamide 5 mM sigH mutant
GSM215392	Exposed to Diamide 5 mM sigH mutant
GSM216629	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216630	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216631	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216632	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216633	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216634	ST22 mutant vs. H37Rv (iron sufficient : 50 μM $FeCl_3$)
GSM216635	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM216636	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM216637	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM216638	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM216639	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM216640	ST52 mutant vs. ST22 mutant (iron sufficient : 50 μM $FeCl_3$)
GSM218911	Concanamycin A inhibition of phagosome acidification
GSM218912	phagocytosis inhibited with cytochalasin D
GSM219337	Hypoxia experiment with *MTB* dosR mutant
GSM219338	Hypoxia experiment with *MTB* dosR mutant
GSM219339	Hypoxia experiment with *MTB* dosR mutant
GSM219340	Hypoxia experiment with *MTB* dosR mutant
GSM219341	Hypoxia experiment with *MTB* dosR mutant
GSM219342	Hypoxia experiment with *MTB* dosR mutant
GSM886249	H37Rv 0 day ctrl vs Rv3132-34 KO hypoxia 4 day
GSM886251	H37Rv ctrl vs Rv3132-3134 KO 0,05 mM DETA-NO 40min
GSM886260	*MTB* strain 1254 Ctrl vs 0,5 mM KCN + 2hr hypoxia
GSM886265	H37Rv ctrl vs Rv3132-34 KO (complemented) 0,05 mM DETA-NO 40min

GEO-sample	Description
GSM886275	H37Rv 0 day ctrl vs Rv3132-34 KO hypoxia 4 day
GSM886277	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 20min + 0,05 mM DETA-NO 40min
GSM886279	H37 Rv3132-34 KO 0 days vs Rv3132-34 KO (complemented) hypoxia 4 day
GSM886285	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 1hr
GSM886287	H37Rv ctrl vs Rv3132-34 KO (complemented) 0,05 mM DETA-NO 40min
GSM886290	H37Rv 0 day ctrl vs Rv3132-34 KO hypoxia 4 day
GSM886292	H37Rv ctrl vs Rv3132-3134 KO 0,05 mM DETA-NO 40min
GSM886302	H37Rv ctrl vs Rv3132-3134 KO 0,05 mM DETA-NO 40min
GSM886318	H37Rv ctrl vs Rv3132-3134 KO 0,05 mM DETA-NO 40min
GSM886324	H37Rv ctrl vs Rv3132-34 KO (complemented) 0,05 mM DETA-NO 40min
GSM886329	H37 Rv3132-34 KO 0 days vs Rv3132-34 KO (complemented) hypoxia 4 day
GSM886336	H37Rv ctrl vs Rv3132-34 KO (complemented) 0,05 mM DETA-NO 40min
GSM886342	H37Rv 0 day ctrl vs Rv3132-34 KO hypoxia 4 day
GSM886368	H37 Rv3132-34 KO 0 days vs Rv3132-34 KO (complemented) hypoxia 4 day
GSM886370	H37 Rv3132-34 KO 0 days vs Rv3132-34 KO (complemented) hypoxia 4 day
GSM886304	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 1hr
GSM886321	*MTB* strain 1254 Ctrl vs 0,5 mM KCN + 2hr hypoxia
GSM886346	*MTB* strain 1254 Ctrl vs 0,5 mM KCN + 2hr hypoxia
GSM886361	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 20min + 0,05 mM DETA-NO 40min
GSM886363	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 1hr
GSM886378	*MTB* strain 1254 Ctrl vs 0,5 mM KCN 20min + 0,05 mM DETA-NO 40min
GSM237655	H37dosR mutant hypoxia 4hr
GSM237656	H37dosR mutant hypoxia 4hr
GSM237657	H37dosR mutant hypoxia 4hr
GSM237658	H37dosR mutant hypoxia 4hr
GSM237659	H37dosR mutant hypoxia 8hr
GSM237660	H37dosR mutant hypoxia 12hr
GSM237661	H37dosR mutant hypoxia 12hr
GSM237662	H37dosR mutant hypoxia 1day
GSM237663	H37dosR mutant hypoxia 1day
GSM237664	H37dosR mutant hypoxia 1day
GSM237665	H37dosR mutant hypoxia 1day

GEO-sample	Description
GSM237666	H37dosR mutant hypoxia 1day
GSM237667	H37dosR mutant hypoxia 4day
GSM237668	H37dosR mutant hypoxia 7day
GSM237672	H37dosR mutant hypoxia 8hr
GSM237673	H37dosR mutant hypoxia 12hr
GSM237674	H37dosR mutant hypoxia 1day
GSM351193	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs
GSM351195	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs
GSM351202	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs
GSM351208	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs plus cholesterol
GSM351277	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs plus cholesterol
GSM351278	CDC1551 (146) vs CDC1551 kstR mutant (162) 24 hrs plus cholesterol
GSM400324	H37Rv wild type vs H37Rv sigB null mutant
GSM400326	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM400327	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400329	H37Rv wild type vs H37Rv sigB null mutant
GSM400330	H37Rv wild type vs H37Rv sigB null mutant
GSM400332	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM400337	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400338	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400342	H37Rv wild type vs H37Rv sigB null mutant
GSM400343	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM400344	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400345	H37Rv wild type vs H37Rv sigB null mutant
GSM400346	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400347	H37Rv wild type vs H37Rv sigB null mutant
GSM400349	H37Rv sigB null mutant control vs 5mM Diamide for 60 min
GSM400352	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM400353	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM400355	H37Rv sigB null mutant control vs 0.05% SDS for 60 min
GSM27855	0.1 μg/mL MMC ctrl (0.33h)
GSM27856	0.2 μg/mL MMC ctrl (0.33h)
GSM27857	0.1 μg/mL MMC ctrl (0.75h)
GSM27858	0.2 μg/mL MMC ctrl (0.75h)
GSM27859	0.2 μg/mL MMC ctrl (1.5h)
GSM27860	0.2 μg/mL MMC ctrl (2h)
GSM27861	0.2 μg/mL MMC ctrl (2h)
GSM27862	0.2 μg/mL MMC ctrl (2h)

GEO-sample	Description
GSM27863	0.2 μg/mL MMC ctrl (2h)
GSM27864	0.2 μg/mL MMC ctrl (4h)
GSM27865	0.2 μg/mL MMC ctrl (4h)
GSM27866	0.2 μg/mL MMC ctrl (4h)
GSM27867	0.2 μg/mL MMC ctrl (4h)
GSM27868	0.2 μg/mL MMC ctrl (4h)
GSM27869	0.2 μg/mL MMC ctrl (6h)
GSM27883	0.12mg mL PZA pH5.6 (2h)
GSM27884	1.2mg mL PZA pH5.6 (2h)
GSM27885	0.12mg mL PZA pH5.6 (2.5h)
GSM27886	0.12mg mL PZA pH5.6 (4h)
GSM27887	1.2mg mL PZA pH5.6 (4h)
GSM27888	0.12mg mL PZA pH5.6 (4h)
GSM27889	1.2mg mL PZA pH5.6 (4h)
GSM27890	0.12mg mL PZA pH5.6 (4h)
GSM27891	0.12mg mL PZA pH5.6 (4h)
GSM27892	1.2mg mL PZA pH5.6 (4h)
GSM27893	1.2mg mL PZA pH5.6 (5h)
GSM27894	0.12mg mL PZA pH5.6 (5h)
GSM27895	0.12mg mL PZA pH5.6 (5.5h)
GSM27896	1.2mg mL PZA pH5.6 (6h)
GSM27897	1.2mg mL PZA pH5.6 (6h)
GSM27898	0.12mg mL PZA pH5.6 (6h)
GSM27899	0.12mg mL PZA pH5.6 (6h)
GSM27900	1.2mg mL PZA pH5.6 (7h)
GSM27901	1.2mg mL PZA pH5.6 (7h)
GSM27902	1.2mg mL PZA pH5.6 (10.5h)
GSM27903	0.12mg mL PZA pH5.6 (10.5h)
GSM27904	0.12mg mL PZA pH5.6 (11h)
GSM27905	1.2mg mL PZA pH5.6 (16h)
GSM27906	0.12mg mL NAM pH5.6 (2h)
GSM27907	0.12mg mL NAM pH5.6 (2h)
GSM27908	0.12mg mL NAM pH5.6 (4h)
GSM27909	0.12mg mL NAM pH5.6 (4h)
GSM27910	0.12mg mL NAM pH5.6 (4h)
GSM27911	1.2mg mL NAM pH5.6 (5h)
GSM27912	0.12mg mL NAM pH5.6 (5.5h)
GSM27913	0.12mg mL NAM pH5.6 (7h)
GSM27914	1.2mg mL NAM pH5.6 (10.5h)
GSM27980	0.12mg mL BZA pH5.6 (2h)
GSM27981	0.12mg mL BZA pH5.6 (4h)

GEO-sample	Description
GSM27982	0.12mg mL BZA pH5.6 (4h)
GSM27983	0.12mg mL BZA pH5.6 (5h)
GSM27984	0.12mg mL BZA pH5.6 (7h)
GSM27985	0.12mg mL BZA pH5.6 (11h)
GSM27986	40 μg/mL 5-CL-PZA pH5.6 (2h)
GSM27987	80 μg/mL 5-CL-PZA pH5.6 (2h)
GSM27988	40 μg/mL 5-CL-PZA pH5.6 (2h)
GSM27989	40 μg/mL 5-CL-PZA pH5.6 (4h)
GSM27990	40 μg/mL 5-CL-PZA pH5.6 (4h)
GSM27991	40 μg/mL 5-CL-PZA pH5.6 (4h)
GSM27992	pH4.8 pH6.8 (2h)
GSM27993	pH4.8 pH6.8 (2h)
GSM27994	pH4.8 pH6.8 (2h)
GSM27995	pH4.8 pH6.8 (4h)
GSM27996	pH4.8 pH6.8 (4h)
GSM27997	pH4.8 pH6.8 (4h)
GSM27998	pH4.8 pH6.8 (5h)
GSM27999	pH4.8 pH6.8 (5.5h)
GSM28000	pH4.8 pH6.8 (6h)
GSM28001	pH4.8 pH6.8 (7h)
GSM28002	pH4.8 pH6.8 (11h)
GSM28003	pH5.2 pH6.8 (2h)
GSM28004	pH5.2 pH6.8 (5h)
GSM28005	pH5.2 pH6.8 (5.5h)
GSM28006	pH5.2 pH6.8 (10.5h)
GSM28007	pH5.2 pH6.8 (11h)
GSM28008	pH5.2 pH6.8 (16h)
GSM28009	pH5.6 pH6.8 (2h)
GSM28010	pH5.6 pH6.8 (2h)
GSM28011	pH5.6 pH6.8 (4h)
GSM28012	pH5.6 pH6.8 (4h)
GSM28013	pH5.6 pH6.8 (4h)
GSM28014	pH5.6 pH6.8 (7h)
GSM28015	pH5.6 pH6.8 (5h)
GSM28016	pH5.6 pH6.8 (11h)
GSM28017	pH5.6 pH6.8 (16h)
GSM28018	0.1 μg/mL MMC ctrl (1.5h)
GSM28021	0.2 μg/mL MMC ctrl (8h) H37Rv
GSM28022	0.2 μg/mL MMC ctrl (8h) H37Rv
GSM28027	0.2 μg/mL MMC ctrl (12h) H37Rv
GSM28031	10 μg/mL EMB DMSO (12h)

GEO-sample	Description
GSM28032	DIPED 5 μg/mL vs DMSO 12h
GSM28033	ethambutol (10 μg/mL) vs DMSO 12h
GSM28034	5 μg/mL DIPED vs DMSO 12h
GSM28035	DIPED (10 μg/mL) vs DMSO 12h
GSM28036	DIPED (10 μg/mL) vs DMSO 12h
GSM28037	EMB (10 μg/mL) vs DMSO 12h
GSM28038	EMB (10 μg/mL) vs DMSO 12h
GSM28039	DIPED (50 μg/mL) vs DMSO 12h
GSM28040	DIPED (50 μg/mL) vs DMSO 12h
GSM28041	DIPED (50 μg/mL) vs DMSO 12h
GSM28042	DIPED (100 μg/mL) vs DMSO 12h
GSM28043	DIPED (100 μg/mL) vs DMSO 12h
GSM28044	DIPED (100 μg/mL) vs DMSO 12h
GSM28045	EMB (10 μg/mL) vs DMSO 12h
GSM28046	241 (0.1 μg/mL) vs DMSO 12h
GSM28047	241 (1 μg/mL) vs DMSO 12h
GSM28048	241 (10 μg/mL) vs DMSO 12h
GSM28049	241 (10 μg/mL) vs DMSO 12h
GSM28050	109 (0.1 μg/mL) vs DMSO 12h
GSM28051	109 (1 μg/mL) vs DMSO 12h
GSM28052	109 (1 μg/mL) vs DMSO 12h
GSM28053	109 (10 μg/mL) vs DMSO 12h
GSM28054	109 (10 μg/mL) vs DMSO 12h
GSM28055	Rif (5 μg/mL) vs DMSO 2.5h
GSM28056	Rif (5 μg/mL) vs DMSO 2.5h
GSM28057	Rif (0.5 μg/mL) vs DMSO 2.5h
GSM28058	Rif (0.5 μg/mL) vs DMSO 2.5h
GSM28059	10 μg/mL Amikacin EtOH 6h
GSM28060	5 μg/mL Streptomycin EtOH 6h
GSM28061	5 μg/mL Streptomycin EtOH 6h
GSM28062	0.2 μg/mL INH EtOH 6h
GSM28063	0.2mg mL Amp EtOH 6h
GSM28064	0.2mg mL Amp EtOH 6h
GSM28065	12 μg/mL Ethionamide EtOH 6h
GSM28066	12 μg/mL Ethionamide EtOH 6h
GSM28067	0.13mg mL TLM DMSO 6h
GSM28068	0.13mg mL TLM DMSO 6h
GSM28069	0.1 mg mL Triclosan EtOH 6h
GSM28070	0.1 mg mL Triclosan EtOH 6h
GSM28071	10 μg/mL Ofloxacin DMSO 6h
GSM28072	10 μg/mL Ofloxacin DMSO 6h

GEO-sample	Description
GSM28073	5 μg/mL Amikacin EtOH 6h
GSM28074	5 μg/mL Amikacin EtOH 6h
GSM28075	40 μg/mL Ethionamide EtOH 6h
GSM28076	40 μg/mL Ethionamide EtOH 6h
GSM28077	0.4 μg/mL INH EtOH 6h
GSM28078	0.4 μg/mL INH EtOH 6h
GSM28079	5 μg/mL Ofloxacin EtOH 6h
GSM28080	5 μg/mL Ofloxacin EtOH 6h
GSM28081	0.2 μg/mL Rif DMSO 6h
GSM28082	0.2 μg/mL Rif DMSO 6h
GSM28083	2 μg/mL Streptomycin EtOH 6h
GSM28084	2 μg/mL Streptomycin EtOH 6h
GSM28085	10 μg/mL Tetracycline EtOH 6h
GSM28086	10 μg/mL Tetracycline EtOH 6h
GSM28087	130 μg/mL TLM DMSO 6h
GSM28088	130 μg/mL TLM DMSO 6h
GSM28089	0.2mg mL Amp EtOH 6h
GSM28090	0.2mg mL Amp EtOH 6h
GSM28091	5 μg/mL Tetracycline EtOH 6h
GSM28092	0.2mg mL TLM EtOH 6h
GSM28093	0.2mg mL TLM EtOH 6h
GSM28094	0.15mg mL Triclosan EtOH 6h
GSM28095	0.15mg mL Triclosan EtOH 6h
GSM28096	10 μg/mL capreomycin EtOH 6h
GSM28097	10 μg/mL Capreomycin EtOH 6h
GSM28098	10 μg/mL Levofloxacin EtOH 6h
GSM28099	10 μg/mL Levofloxacin EtOH 6h
GSM28100	0.5 μg/mL Rifapentine EtOH 6h
GSM28101	0.5 μg/mL Rifapentine EtOH 6h
GSM28102	50 μg/mL Roxithromycin EtOH 6h
GSM28103	50 μg/mL Roxithromycin EtOH 6h
GSM28104	0.1mg mL Triclosan EtOH 6h
GSM28105	0.1mg mL Triclosan EtOH 6h
GSM28106	5 μg/mL Capreomycin EtOH 6h
GSM28107	5 μg/mL Capreomycin EtOH 6h
GSM28108	0.32 μg/mL Cerulenin EtOH 6h
GSM28109	0.32 μg/mL Cerulenin EtOH 6h
GSM28110	30 μg/mL Roxithromycin EtOH 6h
GSM28111	30 μg/mL Roxithromycin EtOH 6h
GSM28112	0.5 μg/mL Cerulenin EtOH 6h
GSM28113	0.5 μg/mL Cerulenin EtOH 6h

GEO-sample	Description
GSM28114	10 μg/mL Chlorpromazine EtOH
GSM28115	10 μg/mL Chlorpromazine EtOH 6h
GSM28116	60 μg/mL Chlorpromazine EtOH 6h
GSM28117	10 μg/mL clotrimazole EtOH 6h
GSM28118	10 μg/mL Clotrimazole EtOH 6h
GSM28119	10 μg/mL Econazole EtOH 6h
GSM28120	10 μg/mL Econazole EtOH 6h
GSM28121	0.1 μg/mL Rifapentine EtOH 6h
GSM28122	0.1 μg/mL Rifapentine; EtOH 6h
GSM28123	10 μg/mL Roxithromycin EtOH 6h
GSM28124	10 μg/mL Roxithromycin EtOH 6h
GSM28125	20 μg/mL Cephalexin Control 6h
GSM28126	20 μg/mL Cephalexin Control 6h
GSM28127	100 μg/mL Cephalexin EtOH 6h
GSM28128	100 μg/mL Cephalexin EtOH 6h
GSM28129	50 μg/mL Triclosan EtOH 6h
GSM28130	50 μg/mL Triclosan EtOH 6h
GSM28131	100 μg/mL Triclosan EtOH 6h
GSM28132	100 μg/mL Triclosan EtOH 6h
GSM28133	10 μg/mL chlorpromazine ctrl 6h
GSM28134	10 μg/mL chlorpromazine ctrl 6h
GSM28135	20 μg/mL Procept 6776 DMSO 6h
GSM28136	0.5 μg/mL 121940 DMSO 6h
GSM28137	0.5 μg/mL 121940 DMSO 6h
GSM28138	0.5 μg/mL 111891 DMSO 6h
GSM28139	0.5 μg/mL 111891 DMSO 6h
GSM28140	0.5 μg/mL 111895 DMSO 6h
GSM28141	30 μg/mL Procept 6776 DMSO 6h
GSM28142	30 μg/mL Procept 6776 DMSO 6h
GSM28143	5 μg/mL Procept 6778 DMSO 6h
GSM28144	5 μg/mL Procept 6778 DMSO 6h
GSM28145	20 μg/mL Procept 6778 DMSO 6h
GSM28146	0.5 μg/mL 111895 DMSO 6h
GSM28147	0.5 μg/mL 124196 DMSO 6h
GSM28148	0.5 μg/mL 124196 DMSO 6h
GSM28149	2 μg/mL 121940 DMSO 6h
GSM28150	200μM Dipyridyl ctrl 6h
GSM28151	12.5 μg/mL Antimycin DMSO 6h
GSM28152	25 μg/mL Antimycin DMSO 6h
GSM28153	20 μg/mL Procept 6776 DMSO 6h
GSM28154	20 μg/mL Procept 6778 DMSO 6h

GEO-sample	Description
GSM28155	20 μg/mL Procept 6778 DMSO 6h
GSM28156	40 μg/mL Procept 6778 DMSO 6h
GSM28157	40 μg/mL Procept 6778 DMSO 6h
GSM28158	2 μg/mL 111895 DMSO 6h
GSM28159	2 μg/mL 111895 DMSO 6h
GSM28160	9 μg/mL Ascididemin Natural product DMSO 6h
GSM28161	50 μg/mL Antimycin A DMSO 6h
GSM28162	50 μg/mL Antimycin A DMSO 6h
GSM28163	10 μg/mL PA-1 DMSO 6h
GSM28164	10 μg/mL PA-1 DMSO 6h
GSM28165	10 μg/mL PA-21 DMSO 6h
GSM28166	10 μg/mL PA-21 DMSO 6h
GSM28167	10 μg/mL Triclosan DMSO 6h
GSM28168	10 μg/mL Triclosan DMSO 6h
GSM28173	10 μg/mL PA-1 DMSO 6h
GSM28174	50 μg/mL PA-21 DMSO 6h
GSM28175	50 μg/mL PA-1 DMSO 6h
GSM28176	10 μg/mL Triclosan DMSO 6h
GSM28177	50 μg/mL PA-21 DMSO 6h
GSM28178	10 μg/mL Triclosan DMSO 6h
GSM28179	0.5 μg/mL Cerulenin DMSO 6h
GSM28180	0.5 μg/mL Cerulenin DMSO 6h
GSM28181	0.5 μg/mL Cerulenin DMSO 6h
GSM28182	150μM Deferoxamine ctrl 6h
GSM28183	150μM Deferoxamine ctrl 6h
GSM28184	150μM Deferoxamine ctrl 6h
GSM28185	100μM Dipyridyl ctrl 6h
GSM28186	100μM Dipyridyl Ctrl 6h
GSM28187	1 μg/mL 111891 DMSO 6h
GSM28188	1 μg/mL 111891 DMSO 6h
GSM28189	1 μg/mL 111895 DMSO 6h
GSM28190	1 μg/mL 111895 DMSO 6h
GSM28191	1 μg/mL 124196 DMSO 6h
GSM28192	1 μg/mL 124196 DMSO 6h
GSM28193	9 μg/mL Ascididemin Natural Product DMSO 6h
GSM28194	9 μg/mL Ascididemin natural product DMSO 6h
GSM28195	2 μg/mL 121940 DMSO 6h
GSM28196	2 μg/mL 111891 DMSO 6h
GSM28197	2 μg/mL 111895 DMSO 6h
GSM28198	50 μg/mL Antimycin DMSO 6h
GSM28199	0.5 μg/mL Cerulenin DMSO 6h

GEO-sample	Description
GSM28200	$250\mu M$ Deferoxamine DMSO 6h
GSM28201	$200\mu M$ Dipyridyl DMSO 6h
GSM28202	9 μg/mL Natural product DMSO 6h
GSM28203	4 μg/mL Natural compound DMSO 6h
GSM28204	1 μg/mL 109 DMSO 6h
GSM28205	10 μg/mL 109 DMSO 6h
GSM28206	1 μg/mL 109 DMSO 6h
GSM28207	10 μg/mL 109 DMSO 6h
GSM28208	1 μg/mL 241 DMSO 6h
GSM28209	1 μg/mL 241 DMSO 6h
GSM28210	10 μg/mL 241 DMSO 6h
GSM28211	1 μg/mL 59 DMSO 6h
GSM28212	1 μg/mL 59 DMSO 6h
GSM28213	10 μg/mL 59 DMSO 6h
GSM28214	10 μg/mL 59 DMSO 6h
GSM28215	10 μg/mL EMB DMSO 6h
GSM28216	10 μg/mL EMB DMSO 6h
GSM28217	1 μg/mL 121940 DMSO 12h
GSM28218	1 μg/mL 111891 DMSO 12h
GSM28219	1 μg/mL 111895 DMSO 12h
GSM28220	10 μg/mL clofazimine DMSO 6h
GSM28221	24 μg/mL clotrimazole DMSO 6h
GSM28222	20 μg/mL chlorpromazine DMSO 6h
GSM28223	2.5mM Dinitrophenol DMSO 6h
GSM28224	50 μg/mL Thioridazine DMSO 6h
GSM28225	50 μg/mL Triclosan DMSO 6h
GSM28226	20 μg/mL Procept 6776 DMSO 6h
GSM28227	20 μg/mL Procept 6778 DMSO 6h
GSM28228	100 μg/mL Cephalexin Ctrl 6h
GSM28229	50 μg/mL Triclosan EtOH
GSM28230	10 μg/mL 109 DMSO 6h
GSM28231	10 μg/mL 59 DMSO 6h
GSM28232	20 μg/mL ethambutol DMSO
GSM28233	10 μg/mL 241 DMSO 6h
GSM28234	24 μg/mL Econazole DMSO 6h
GSM28235	24 μg/mL Econazole DMSO 6h
GSM28236	20 μg/mL Procept 6778 DMSO 6h
GSM28239	13 μg/mL clofazimine DMSO 6h
GSM28240	13 μg/mL clofazimine DMSO 6h
GSM28241	24 μg/mL clotrimazole DMS 6h
GSM28242	24 μg/mL Clotrimazole DMSO 6h

GEO-sample	Description
GSM28243	20 μg/mL chlorpromazine DMSO 6h
GSM28244	0.4 μg/mL PA824 DMSO 6h
GSM28245	0.4 μg/mL PA824 DMSO 6h
GSM28246	2 μg/mL PA824 DMSO 6h
GSM28247	2 μg/mL PA824 DMSO 6h
GSM28248	0.2 μg/mL PA824 DMSO 6h
GSM28249	0.4 μg/mL PA824 DMSO 6h
GSM28250	0.2 μg/mL PA824 DMSO 6h
GSM28251	0.4 μg/mL PA824 DMSO 6h
GSM28252	10μM CCCP DMSO 6h
GSM28253	100 μg/mL Cephalexin DMSO 6h
GSM28254	100μM DCCD DMSO 6h
GSM28255	1mM DNP DMSO 6h
GSM28256	20 μg/mL KCN DMSO 6h
GSM28257	100 μg/mL Cephalexin DMSO 6h
GSM28258	10 μg/mL Thioridazine DMSO 6h
GSM28259	10 μg/mL Novobiocin DMSO 6h
GSM28260	5 μg/mL KCN DMSO 6h
GSM28261	0.5mM DNP DMSO 6h
GSM28262	20μM DCCD DMSO 6h
GSM28263	50μM CCCP DMSO 6h
GSM28264	Min med -Palmitate Min med - Glucose
GSM28265	Min med - Succinate Min med glucose
GSM28266	50μM CCCP DMSO 6h
GSM28267	20μM DCCD DMSO 6h
GSM28268	5 μg/mL KCN DMSO 6h
GSM28269	10 μg/mL Novobiocin DMSO 6h
GSM28270	10 μg/mL Thioridazine DMSO 6h
GSM28271	10mM Succinate (min med) 10mM Glucose (min med)
GSM28272	0.05mM Palmitate (min med) 10mM Glucose (min med)
GSM28273	1mM DNP DMSO 6h
GSM28274	25 μg/mL Thioridazine DMSO 6h
GSM28275	10 μg/mL Novobiocin DMSO 6h
GSM28276	10mM Succinate (min med) 10mM Glucose (min med)
GSM28277	0.05mM Palmitate (min med) 10mM Glucose (min med)
GSM28278	1 μg/mL ARP4 DMSO 6h
GSM28279	4 μg/mL ARP4 DMSO 6h
GSM28280	1 μg/mL ARP2 DMSO 6h
GSM28281	4 μg/mL ARP2 DMSO 6h
GSM28282	10mM Succinate (min med) 10mM Glucose (min med)
GSM28283	0.05mM Palmitate (min med) 10mM Glucose (min med)

GEO-sample	Description
GSM28284	10 μg/mL Methoxatin DMSO 6h
GSM28285	0.1mM GSNO 5 μg/mL KCN DMSO 6h
GSM28286	0.1mM GSNO DMSO 6h
GSM28287	7d Dubos NRP-1 Dubos log phase
GSM28288	7d Dubos NRP-1 Dubos log phase
GSM28289	20 μg/mL methoxatin DMSO 6h
GSM28290	10 μg/mL methoxatin DMSO 6h
GSM28291	0.1mM GSNO DMSO 6h
GSM28292	0.1mM GSNO 5 μg/mL KCN DMSO 6h
GSM28293	Dubos NRP-1 Dubos log-phase
GSM28294	1mM DTNB DMSO 6h
GSM28295	2mM DTT DMSO 6h
GSM28296	1μM Valinomycin DMSO 6h
GSM28297	10μM Valinomycin DMSO 6h
GSM28298	2mM b-mercaptoethanol DMSO 6h
GSM28299	2mM DTNB DMSO
GSM28300	2mM DTT DMSO 6h
GSM28301	50μM Nigericin DMSO 6h
GSM28302	1μM Valinomycin DMSO 6h
GSM28303	50 μg/mL Verapamil DMSO 6h
GSM28304	1μM Valinomycin DMSO 6h
GSM28305	20 μg/mL CPZ DMSO 6h
GSM28306	PBS Tw DMSO 7H9 6h
GSM28307	10 μg/mL menadione DMSO 6h
GSM28308	0.1mM GSNO 20 μg/mL CPZ DMSO 6h
GSM28309	0.1mM GSNO 10 μg/mL menadione DMSO 6h
GSM28310	50μM CCCP DMSO 6h
GSM28311	50μM Nigericin DMSO 6h
GSM28312	0.5μM Valinomycin DMSO 6h
GSM28313	0.5μM Valinomycin 70mM KCl DMSO 6h
GSM28314	0.1mM GSNO 20 μg/mL CPZ DMSO 6h
GSM28315	0.1mM GSNO 10 μg/mL menadione DMSO 6h
GSM28316	20 μg/mL CPZ DMSO 6h
GSM28317	10 μg/mL menadione (Dubos) DMSO 6h
GSM28318	4d Dubos NRP-1 Dubos aerobic log phase
GSM28319	4d 20mM nitrate in Dubos NRP-1 aerobic Dubos
GSM28320	25 μg/mL CPZ DMSO ctrl 6h
GSM28321	0.02mM GSNO DMSO 6h
GSM28322	25 μg/mL CPZ + 0.1mM GSNO DMSO ctrl
GSM28323	5 μg/mL Clofazimine + 0.1mM GSNO DMSO
GSM28324	0.4mM NaN_3 DMSO 4h

GEO-sample	Description
GSM28325	PBS Tween80 DMSO DMSO 6h
GSM28326	25 μg/mL CPZ+ 0.1mM GSNO DMSO
GSM28327	10 μg/mL menadione + 0.1mM GSNO DMSO
GSM28328	0.4mM NaN_3 DMSO
GSM28329	4d Dubos NRP-1 Dubos aerobic log phase
GSM28330	25 μg/mL CPZ DMSO 6h
GSM28331	0.1mM GSNO DMSO 4h
GSM28332	5 μg/mL Clofazimine + 0.1mM GSNO DMSO
GSM28333	0.4mM NaN_3 DMSO 4h
GSM28334	25 μg/mL CPZ DMSO 6h
GSM28335	6 μg/mL Menadione DMSO 6h
GSM28336	25 μg/mL CPZ + 0.1mM GSNO DMSO
GSM28337	6 μg/mL Menadione + 0.1mM GSNO DMSO
GSM28338	2mM NaN_3 DMSO 6h
GSM28339	2mM ZnSO4 DMSO 6h
GSM28340	25 μg/mL CPZ + 0.1mM GSNO DMSO
GSM28341	5 μg/mL Clofazimine + 0.1mM GSNO DMSO
GSM28342	0.1mM GSNO DMSO 3.5h
GSM28343	0.2mM NaN_3 DMSO
GSM28344	25 μg/mL CPZ + 0.1mM GSNO DMSO
GSM28345	10 μg/mL menadione + 0.1mM GSNO DMSO
GSM28346	0.4mM NaN_3 DMSO
GSM28347	25 μg/mL CPZ DMSO 6h
GSM28348	1mM DTT DMSO 6h
GSM28349	2mM NaN_3 DMSO 6h
GSM28350	50μM Nigericin DMSO 6h
GSM28351	10 μg/mL menadione DMSO 6h
GSM28352	0.1mM GSNO DMSO 6h
GSM28353	5 μg/mL Clofazimine + 0.1mM GSNO DMSO
GSM28354	25 μg/mL CPZ + 0.1mM GSNO DMSO
GSM28355	10 μg/mL Menadione + 0.1mM GSNO DMSO
GSM28356	PBS Tween 7H9

TABLE 4.1: GEO samples discarded

4.5.2 Layers dictionary

Once filtered the useful information at hand, we have algorithmically grouped samples constituting repetitions of the same experiments, so as to reconstruct one version (or layer) of the original PPIN for each experiment. In the following table, we enlist those layers (NA means that, for that experiment, no link passed the significance test):

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE14840	V14	24 hour phosphate starvation	CDC1551	ION DEPRIVATION	PHOSPHATE
GSE14840	V15	72 hour phosphate starvation	CDC1551	ION DEPRIVATION	PHOSPHATE
GSE8732	V97	inversion Comparison between iron-sufficient(50 uM FeCl3) and iron-deficient (2 uM FeCl3) grew conditions	H37Rv	ION DEPRIVATION	IRON
GSE1642	NA	100uM Dipyridyl control 6h	H37Rv	ION DEPRIVATION	IRON SCAVENGERS
GSE1642	NA	200uM Dipyridyl control 6h	H37Rv	ION DEPRIVATION	IRON SCAVENGERS
GSE1642	V34	150uM Deferoxamine control 6h	H37Rv	ION DEPRIVATION	IRON SCAVENGERS
GSE1642	V38	WT t=0h WT t=6h in TBST	H37Rv	RESPIRATION ARREST	NUTRIENT STARVATION
GSE1642	NA	Minimal medium (Succinate) 7H9-based medium	H37Rv	MINIMAL MEDIUM	MINIMAL MEDIUM
GSE1642	NA	20mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE1642	NA	20mJ cm2 UV control (8h) H37Rv	H37Rv	DNA DAMAGE	UV
GSE1642	NA	20mJ cm2 UV control (12h) H37Rv	H37Rv	DNA DAMAGE	UV
GSE1642	NA	25mJ cm2 UV control (2h)	H37Rv	DNA DAMAGE	UV
GSE1642	NA	25mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE1642	NA	40mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE1642	V35	40mJ cm2 UV control (8h) H37Rv	H37Rv	DNA DAMAGE	UV
GSE1642	NA	40mJ cm2 UV control (12h) H37Rv	H37Rv	DNA DAMAGE	UV
GSE1642	NA	60mJ cm2 UV control (2h)	H37Rv	DNA DAMAGE	UV
GSE1642	NA	60mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE1642	NA	4mM H2O2 control (1h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE1642	V36	4mM H2O2 control (2h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE1642	V37	4mM H2O2 control (4h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE1642	NA	4mM H2O2 control (8h) H37Rv	H37Rv	OXIDATIVE STRESS	H2O2
GSE1642	NA	4mM H2O2 control (12h) H37Rv	H37Rv	OXIDATIVE STRESS	H2O2
GSE365	NA	20mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	20mJ cm2 UV control (8h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	20mJ cm2 UV control (12h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	25mJ cm2 UV control (2h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	25mJ cm2 UV control (4h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	25mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	40mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE365	V86	40mJ cm2 UV control (8h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	40mJ cm2 UV control (12h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	60mJ cm2 UV control (2h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	60mJ cm2 UV control (4h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	60mJ cm2 UV control (6h)	H37Rv	DNA DAMAGE	UV
GSE365	NA	4mM H2O2 control (1h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE365	V87	4mM H2O2 control (2h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE365	V88	4mM H2O2 control (4h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE365	NA	4mM H2O2 control (8h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE365	NA	4mM H2O2 control (12h)	H37Rv	OXIDATIVE STRESS	H2O2
GSE16146	V19	*MTB* strain1254 vs *MTB* strain 1254 for DETA NO exp control	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V21	*MTB* strain 1254 control vs 0.005 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V22	*MTB* strain 1254 control vs 0.05 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V24	*MTB* strain 1254 control vs 0.5 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE16146	V27	*MTB* strain 1254 control vs 1.0 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V30	*MTB* strain 1254 control vs 5.0 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V25	*MTB* strain 1254 control vs 0.5 mM DETA NO 4hrs	1254	OXIDATIVE STRESS	DETA NO
GSE16146	V20	*MTB* strain 1254 vs *MTB* strain 1254 for H2O2 exp control	1254	OXIDATIVE STRESS	H2O2
GSE16146	V23	*MTB* strain 1254 control vs 0.05 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V26	*MTB* strain 1254 control vs 0.5 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V31	*MTB* strain 1254 control vs 5.0 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V28	*MTB* strain 1254 control vs 10.0 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V33	*MTB* strain 1254 control vs 50.0 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V29	*MTB* strain 1254 control vs 200 mM H202 40min	1254	OXIDATIVE STRESS	H2O2
GSE16146	V32	*MTB* strain 1254 control vs 5.0 mM H202 4hr	1254	OXIDATIVE STRESS	H2O2
GSE8839	V116	merged *MTB*1254 Day 0	1254	OXIDATIVE STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V117	merged *MTB*1254 Day 0 vs low Oxygen Day 4	1254	OXIDATIVE STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V130	merged *MTB* strain 1254 control vs 2hr Hypoxia	1254	OXIDATIVE STRESS	HYPOXIA
GSE8839	V115	merged H37Rv control vs 0,05 mM DETA NO 40min	H37Rv	OXIDATIVE STRESS	DETA NO
GSE8839	V114	merged CDC1551 control vs 0,05 mM DETA NO 40 min	CDC1551	OXIDATIVE STRESS	DETA NO
GSE8839	V118	merged *MTB* strain 1254 control vs 0,005 mM DETA NO 40min	1254	OXIDATIVE STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V119	merged *MTB* strain 1254 control vs 0,05 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V129	merged *MTB* strain 1254 control vs 1,0 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V131	merged *MTB* strain 1254 control vs 5,0 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V127	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 5min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V122	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 20min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V126	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V121	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 1hr	1254	OXIDATIVE STRESS	DETA NO

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE8839	V125	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 2hrs	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V128	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 8hrs	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V120	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 16hrs	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V123	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 24hrs	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V124	merged *MTB* strain 1254 control vs 0,5 mM DETA NO 24hrs + 0,5 mM DNO 40min	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V132	merged *MTB* strain 1254 control vs high aeration	1254	OXIDATIVE STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V133	merged *MTB* strain 1254 control vs high aeration + 0,001 mM DETA NO	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V134	merged *MTB* strain 1254 control vs high aeration + 0,005 mM DETA NO	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V135	merged *MTB* strain 1254 control vs high aeration + 0,01 mM DETA NO	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V136	merged *MTB* strain 1254 control vs high aeration + 0,05 mM DETA NO	1254	OXIDATIVE STRESS	DETA NO
GSE8839	V137	merged *MTB* strain 1254 control vs low aeration (0,2% oxigen 2hr	1254	DORMANCY MODEL OX STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V138	merged *MTB* strain 1254 control vs low aeration + 0,001 mM DETA NO	1254	DORMANCY MODEL OX STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V139	merged *MTB* strain 1254 control vs low aeration + 0,005 mM DETA NO	1254	DORMANCY MODEL OX STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V140	merged *MTB* strain 1254 control vs low aeration + 0,01 mM DETA NO	1254	DORMANCY MODEL OX STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8839	V141	merged *MTB* strain 1254 control vs low aeration + 0,05 mM DETA NO	1254	DORMANCY MODEL OX STRESS	OXIDATIVE STRESS AND RESPIRATION ARREST
GSE8689	V96	Exposed to Diamide 5 mM H37Rv wild type	H37Rv	OXIDATIVE STRESS	DIAMIDE
GSE15976	V16	H37Rv wild type Vs H37Rv wild type	H37Rv	CELL WALL DAMAGE	SDS
GSE15976	V17	H37Rv wild type control vs 0.05% SDS for 60 min	H37Rv	CELL WALL DAMAGE	SDS
GSE15976	V18	H37Rv wild type control vs 5mM Diamide for 60 min	H37Rv	OXIDATIVE STRESS	DIAMIDE
GSE6750	V94	Rv-Control vs Rv-SDS treated	H37Rv	CELL WALL DAMAGE	SDS

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE8664	V95	Exposed to 0.05% SDS H37Rv wild type	H37Rv	CELL WALL DAMAGE	SDS
GSE50159	V89	WT 140 mM NaCl	CDC1551	CELL WALL DAMAGE	NACL
GSE10391	V1	MS1 D01	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V2	MS1 D02	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V3	MS1 D03	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V4	MS1 D06	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V5	MS1 D12	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V6	MS2 D03	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V7	MS2 D09	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V8	MS2 D18	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V9	MS3 D03	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V10	MS3 D09	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE10391	V11	MS3 D18	H37Rv	DORMANCY MODEL HYPOXIA	RESPIRATION ARREST LOW PH AND GLYCEROL DEPRIVED
GSE5977	V90	Mtb late-log phase	H37Rv	RESPIRATION AREST	STACIONARY PHASE
GSE5977	V91	Mtb stationary phase	H37Rv	RESPIRATION AREST	STACIONARY PHASE
GSE21590	V74	H37Rv 1hour reareation	H37Rv	RESPIRATION AREST	REAREATION

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE21590	V76	H37Rv 4hour reareation	H37Rv	RESPIRATION ARREST	REAREATION
GSE21590	V77	H37Rv 6hour reareation	H37Rv	RESPIRATION ARREST	REAREATION
GSE21590	V73	H37Rv 12hour reareation	H37Rv	RESPIRATION ARREST	REAREATION
GSE21590	V75	H37Rv 24hour reareation	H37Rv	RESPIRATION ARREST	REAREATION
GSE8786	V98	Growth curve 0 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V102	Growth curve 6 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V103	Growth curve 8 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V99	Growth curve 14 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V100	Growth curve 24 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V101	Growth curve 60 days Control	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V109	oxygen-depleted Wayne Growth 4 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V110	oxygen-depleted Wayne Growth 6 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V112	oxygen-depleted Wayne Growth 8 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V104	oxygen-depleted Wayne Growth 10 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V105	oxygen-depleted Wayne Growth 12 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V106	oxygen-depleted Wayne Growth 14 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V107	oxygen-depleted Wayne Growth 20 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V108	oxygen-depleted Wayne Growth 30 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE8786	V111	oxygen-depleted Wayne Growth 80 days	H37Rv	RESPIRATION ARREST	WAYNE GROWTH
GSE9331	V145	H37Rv hypoxia 4hr	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE9331	V147	H37Rv hypoxia 8hr	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE9331	V142	H37Rv hypoxia 12hr	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE9331	V143	H37Rv hypoxia 1day	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE9331	V144	H37Rv hypoxia 4day	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE9331	V146	H37Rv hypoxia 7day	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE8829	V113	Hypoxic conditions experiment of *MTB* wild type	H37Rv	RESPIRATION ARREST	HYPOXIA
GSE13978	V13	*MTB* - cholesterol 3hrs	H37Rv	CHOLESTEROL	CHOLESTEROL
GSE13978	V12	*MTB* - cholesterol 24hrs	H37Rv	CHOLESTEROL	CHOLESTEROL
GSE6209	V93	H37Rv 4hrs after infection in $M\phi$ vs. H37Rv grown in 7H9 media biological 4h	H37Rv	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE6209	V92	H37Rv 24hrs after infection in $M\phi$ vs. H37Rv grown in 7H9 media biological 24h	H37Rv	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V78	14-day resting $M\phi$ infection 2hr	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE35362	V82	14-day resting $M\phi$ infection Day 2	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V83	14-day resting $M\phi$ infection Day 4	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V84	14-day resting $M\phi$ infection Day 6	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V85	14-day resting $M\phi$ infection Day 8	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V79	14-day resting $M\phi$ infection Day 10	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V80	14-day resting $M\phi$ infection Day 12	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE35362	V81	14-day resting $M\phi$ infection Day 14	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V55	merged H37Rv extra vs intracellular in resting $M\phi$s	H37Rv	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V39	merged CDC1551 extra vs intracellular in resting $M\phi$s	CDC1551	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V40	merged Clinical isolate 10514-01 Afri2 extra vs intracellular in resting $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V41	merged Clinical isolate 10517-01 Afri2 extra vs intracellular in resting $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V52	merged Clinical isolate 5468-02 Afri2 extra vs intracellular in resting $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V53	merged Clinical isolate 947-01 EAI extra vs intracellular in resting $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V44	merged Clinical isolate 1797-03 EAI extra vs intracellular in resting $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V51	merged Clinical isolate 4850-03 EAI extra vs intracellular in resting $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V42	merged Clinical isolate 12954-03 Beijing extra vs intracellular in resting $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V43	merged Clinical isolate 1500-03 Beijing extra vs intracellular in resting $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V45	merged Clinical isolate 1934-03 Beijing extra vs intracellular in resting $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V46	merged Clinical isolate 2169-99 Uganda extra vs intracellular in resting $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V47	merged Clinical isolate 2191-99 Uganda extra vs intracellular in resting $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V48	merged Clinical isolate 2333-99 Uganda extra vs intracellular in resting $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE

GEO-serie	Code	Description	Strain	Type of environment	Subtype
GSE21112	V49	merged Clinical isolate 2336-02 Haarlem extra vs intracellular in resting $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V50	merged Clinical isolate 4130-02 Haarlem extra vs intracellular in resting $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21112	V54	merged Clinical isolate 9532-03 Haarlem extra vs intracellular in resting $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	RESTING MACROPHAGE
GSE21113	V72	merged H37Rv extra vs intracellular in activated $M\phi$s	H37Rv	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V56	merged CDC1551 extra vs intracellular in activated $M\phi$s	CDC1551	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V69	merged Clinical isolate 5468-02 Afri2 extra vs intracellular in activated $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V57	merged Clinical isolate 10514-01 Afri2 extra vs intracellular in activated $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V58	merged Clinical isolate 10517-01 Afri2 extra vs intracellular in activated $M\phi$s	AFRI2	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V61	merged Clinical isolate 1797-03 EAI extra vs intracellular in activated $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V68	merged Clinical isolate 4850-03 EAI extra vs intracellular in activated $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V70	merged Clinical isolate 947-01 EAI extra vs intracellular in activated $M\phi$s	EAI	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V59	merged Clinical isolate 12954-03 Beijing extra vs intracellular in activated $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V60	merged Clinical isolate 1500-03 Beijing extra vs intracellular in activated $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V62	merged Clinical isolate 1934-03 Beijing extra vs intracellular in activated $M\phi$s	BEIJING	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V63	merged Clinical isolate 2169-99 Uganda extra vs intracellular in activated $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V64	merged Clinical isolate 2191-99 Uganda extra vs intracellular in activated $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V65	merged Clinical isolate 2333-99 Uganda extra vs intracellular in activated $M\phi$s	UGANDA	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V66	merged Clinical isolate 2336-02 Haarlem extra vs intracellular in activated $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V67	merged Clinical isolate 4130-02 Haarlem extra vs intracellular in activated $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE
GSE21113	V71	merged Clinical isolate 9532-03 Haarlem extra vs intracellular in activated $M\phi$s	HAARLEM	GRANULOMA MODEL/REAL INFECTION	ACTIVATED MACROPHAGE

TABLE 4.2: Dictionary of samples used.

For some series, we analized them, but it resulted that no sample within them contained any significant interaction. These were GSE8827, GSE10336, GSE14005, GSE15642, GSE32236 and GSE34363.

4.5.3 layers ordering in figures 4.2 and 4.3

In the following table we list the position of each of the listed layers in dendrograms represented in figures 4.2 and 4.3:

Position (from left)	Figure 4.2	Figure 4.3
1	V78	V13
2	V80	V92
3	V81	V93
4	V82	V84
5	V79	V79
6	V83	V85
7	V84	V83
8	V85	V82
9	V44	V81
10	V49	V80
11	V54	V78
12	V55	V54
13	V45	V48
14	V53	V43
15	V40	V50
16	V52	V49
17	V39	V45
18	V47	V42
19	V46	V47
20	V42	V55
21	V43	V46
22	V50	V53
23	V48	V39
24	V51	V40
25	V57	V41
26	V58	V52
27	V69	V51
28	V70	V44
29	V63	V62
30	V64	V59
31	V67	V60
32	V65	V65

Position (from left)	Figure 4.2	Figure 4.3
33	V62	V71
34	V68	V64
35	V61	V63
36	V72	V67
37	V60	V66
38	V66	V72
39	V71	V70
40	V59	V68
41	V56	V56
42	V41	V61
43	V74	V69
44	V76	V58
45	V75	V57
46	V73	V76
47	V77	V74
48	V93	V113
49	V92	V12
50	V12	V77
51	V87	V73
52	V36	V75
53	V88	V137
54	V37	V132
55	V86	V97
56	V35	V107
57	V124	V108
58	V25	V101
59	V120	V111
60	V123	V99
61	V128	V100
62	V141	V102
63	V30	V103
64	V131	V110
65	V24	V112
66	V27	V104
67	V129	V105
68	V122	V106
69	V126	V109
70	V121	V117
71	V125	V145
72	V127	V144
73	V114	V146

Position (from left)	Figure 4.2	Figure 4.3
74	V22	V143
75	V119	V142
76	V139	V147
77	V140	V91
78	V130	V90
79	V115	V5
80	V136	V6
81	V133	V7
82	V134	V4
83	V135	V1
84	V138	V2
85	V33	V3
86	V29	V8
87	V32	V10
88	V31	V9
89	V28	V11
90	V18	V96
91	V96	V28
92	V94	V31
93	V17	V95
94	V95	V17
95	V147	V94
96	V142	V18
97	V143	V33
98	V146	V29
99	V144	V32
100	V145	V141
101	V117	V140
102	V109	V136
103	V105	V115
104	V104	V139
105	V106	V130
106	V112	V135
107	V110	V133
108	V113	V138
109	V108	V134
110	V107	V119
111	V101	V22
112	V103	V114
113	V102	V125
114	V100	V121

Position (from left)	Figure 4.2	Figure 4.3
115	V99	V126
116	V90	V122
117	V5	V24
118	V7	V131
119	V6	V30
120	V10	V129
121	V8	V27
122	V11	V128
123	V9	V123
124	V3	V120
125	V2	V25
126	V1	V124
127	V4	V127
128	V91	V35
129	V89	V86
130	V13	V37
131	V97	V88
132	V132	V36
133	V137	V87
134	V19	V89
135	V21	V14
136	V118	V15
137	V34	V19
138	V38	V21
139	V98	V118
140	V116	V34
141	V16	V38
142	V20	V98
143	V23	V116
144	V26	V16
145	V14	V20
146	V15	V23
147	V111	V26

TABLE 4.3: Layers' positions in figures 4.2 and 4.3

4.5.4 Assignation to environment classification to stresses

In order to construct our consensus stress-response layers we have assigned some of the experiments of our original multi-layer network to each one of the six stress-related con-

Consensus layer	Environmental subtype (see table 4.2)
Hypoxia	Hypoxia
Starvation	Stationary phase
Starvation	Nutrient starvation
Hypoxia	Reareation
Hypoxia	Wayne growth (hypoxia model)
Cell wall damage	Surfactants
Cell wall damage	SPHEROPLAST INDUCED
Cell wall damage	NACL
Cell wall damage	SDS
Ions deprivation	Phosphate deprivation
Ions deprivation	Iron scavengers
Ions deprivation	Iron
NO exposure	DETA-NO
Oxidative stress	H_2O_2
Oxidative stress	Diamide
Starvation	Minimal medium

TABLE 4.4: Sub-types of experiments associated to layers in the stress response multi-layer system

sensus layers. In the following table, we list the environmental subtypes, as annotated in table 4.2 that joins each consensus layer:

Part III

Transcriptional regulatory networks: analysis and data reliability

No tengo a quién rezarle
pidiendo luz.
Ando tanteando
el espacio a ciegas.

Jorge Drexler.
Hermana Duda, en 12 segundos de Oscuridad, 2006.

Chapter 5

Topological effects of data incompleteness of gene regulatory networks

5.1 Background

In the second part of this thesis we undertook the study of different networks of biomolecular interactions of *MTB*. Specifically, on the one hand we characterized the transformations that take place on its PPIN as a consequence of transcriptional adaptation to different environmental conditions. On the other hand we compiled and analyzed a bibliography-based, updated version of the genome-wide TRN of the bacterium. In what regards this last system, despite the systematic character of the bibliographic revision made, current knowledge about *MTB* transcriptomics is only partial, and so, as it has already been mentioned, genes included in the network hardly cover the 40% of the genome. Additionally, the specific methodologies after which experimental evidence is considered enough to endorse the presence of a given interaction are highly heterogeneous, and certain link attributes -specifically signs- can not always be derived from them.

In this third part of the thesis, our intention is that of addressing the reach of this inherent data incompleteness of diverse nature that appears in TRNs; and its affection on the outcomes of different topological analysis methods typically performed in the field. In order to do so, we analyze the TRN of *MTB*, previously curated by ourselves [3], and compare it to two of the best known prokaryotic TRNs: those of the model bacteria *Escherichia coli* [314] and *Bacillus subtilis* [315].

Specifically, the general question we set to answer here is whether robust and biologically relevant conclusions about TRNs can be reached given the current incompleteness of the data, going a step further with respect to other works that had somehow addressed this question previously [317]. Besides, we also show that some topological metrics do depend on the level of detail incorporated in TR maps, in particular, the structure of the mesoscale. Our findings show that extreme care should be taken when strong claims are made based on partial data. This is the case of TRNs superfamilies, which we argue are indeed grouped into a single class.

5.2 Community detection and link attributes

The identification of modules in complex networks has attracted much attention of the scientific community in the last years. A modular view of a network offers a coarse-grained perspective in which nodes are gathered not due to knowledge-based decisions

–function, composition, etc.–, but rather on a topological basis –who is connected to whom. To this end Newman put forward the concept of modularity Q [36], which quantifies how far a certain partition is from a random counterpart. From this definition, algorithms and heuristics to optimize modularity (Q) have appeared ever faster and more efficient [318], and generalizations to directed, weighted and signed networks are also available in the literature [319, 320]. All these efforts have led to a considerable success regarding the quality of detected community structure in networks, and thus a more complete topological knowledge at this level has been attained. Behind this interest underlies the intuition that the relation between network structure and dynamics is strongly mediated by the mesoscale, and that community structure plays a central role in network formation and functioning. And yet, with few exceptions, link attributes are seldom taken into account.

In this section we intend to underline that interaction direction and sign critically shape the detected community structure of a network. This is even more dramatic in the case of TRNs, where a sharp distinction must be made between regulators (which mostly emit links) and the rest of the network, which mainly receives them. Also it is peculiar (though not exclusive) of these systems to allow for positive (activating) and negative (inhibitory) relationships. In practice, directions and signs are not always available in the datasets. Regarding directionality, we analyze a system –the TRN of *MTB* [3]– for which that is not an actual problem, as regulatory proteins are well identified, i.e. their function as link sources is known. Nevertheless, there are many cases of organisms whose regulatory pathways have not been explicitly identified, and in those cases the real topology is usually replaced by a co-expression network, which acts as an undirected proxy for the true underlying regulatory structure. Unavailability of interaction signs is, on the other hand, a more persistent problem: there exist many experimental approaches to infer a transcriptional regulation that do not inform about the sign of the interaction. Furthermore, there are interaction signs which depend on environmental conditions. Therefore, given the unavoidable incompleteness of the data, we explore whether link attributes determine the network modular structure, and to what extent.

To address the previous question, we perform a systematic comparison of the effects of preserving the original information (sign and direction) in modularity measures and community structure in TRNs. To this end, we will analyze the TRN of *MTB* [3], for which we will consider three different topologies: one that preserves all available information (directed-signed, DS); an intermediate one (preserving directions, but not signs –directed-unsigned, DU); and a last one where all fine-grained information is ignored (undirected-unsigned, UU). From the output of this analysis, we provide a way to quantify how much biological information is lost when directions and/or signs are dropped out. Note that the three versions of the network have the same number of nodes N and number of links L, the only differences being those regarding direction and/or the sign of the interactions. Interaction signs have been compiled from the experimental works enlisted in [3], although signs were not reported there (see [279]).

The modularity expression used hereafter corresponds to its most general definition, i.e. the one that accounts for the existence of directions, weights, signed relations and

self-loops, preserving the original information [320]:

$$Q = \frac{w^+}{w^+ + w^-} Q^+ - \frac{w^-}{w^+ + w^-} Q^- \tag{5.1}$$

This expression generalizes the concept of modularity, and simply computes the contribution to group formation of positive (w^+) and negative (w^-) interactions separately, Q^+ and Q^- respectively, which can be interpreted as the tendency to form communities (positive weights) and that of negative weights to dissolve them. For more detail, Q^+ is defined as

$$Q^+ = \frac{1}{2w^+} \sum_i \sum_j \left(w_{ij}^+ - \frac{w_i^+ w_j^+}{2w^+}) \right) \tag{5.2}$$

which accounts for the deviation of actual positive weights w_{ij}^+ against a null case random network; the negative counterpart Q^- is defined accordingly, just placing negative weights in the expression. As for our current object of study, links in the network can only take values +1 or -1, and are originally defined as directed.

An intrinsic limitation of modularity maximization, as posed in Eq. 5.1, is that it provides a single snapshot of the modular structure of the network. However, several topological descriptions of the network coexist at different scales, which is, in general, a fingerprint of complex systems, and particularly relevant in biological ones [321]. A method to overcome this fundamental drawback of typical modularity optimization was put forth in [319]. A parameter r is introduced as a constant self-loop to each node, thus changing the total strength in a network and avoiding the inherent resolution limit of Newman's modularity Q [322]. The shift only affects the property of each node individually and in the same way for all of them. Thus, the original adjacency matrix $\mathbf{A}$ is changed as a function of r: $\mathbf{A_r} = \mathbf{A} + \mathbf{Ir}$. The interesting property of the rescaled topology is that its characteristic scale in terms of modularity has changed. Then the topological structure revealed by optimizing the modularity for $\mathbf{A_r}$ is that of large groups for small values of r, and smaller groups for large values of r, all of which are strictly embedded in the original topology. As an example, the method can uncover each significant resolution level in the well-known synthetic hierarchical network model RB [323], see Figure one in [319]. To perform these costly calculations we have used a mixture of heuristics, including extremal optimization and Newman's fast algorithm, as implemented in [324].

Figure 1 (top) represents the number of modules N_c that a combination of Q-maximization heuristics [324] has detected for the three versions of the TRN of *MTB*. Each topology has been scrutinized at different scales, screening the parameter r for 200 possible values, in a range such that it yielded an interpretable amount of modules. This range changes for different topologies, thus r is normalized in the plot to allow for comparison. On visual inspection it is apparent that the three topologies present plateaus, where different r values yield similar partitions in terms of N_c. This indicates that certain topological scales are robust and persistent, which might be a clue to identify functionally relevant groups of nodes [319]. Notably, the UU topology presents

a single plateau at $N_c = 205$ and then fails to stabilize for larger r's. On the contrary, DU and DS, which retain more information, yield stable partitions at many levels. Although for different r values, these topologies exhibit almost the same behavior regarding plateaus and the number of communities N_c these plateaus present. At this point, one can say that the mesoscale analysis for DU and DS networks allows a richer interpretation in terms of the grouping of nodes, but there is no way to confirm if these are more or less biologically sound, than, for example, the UU topology.

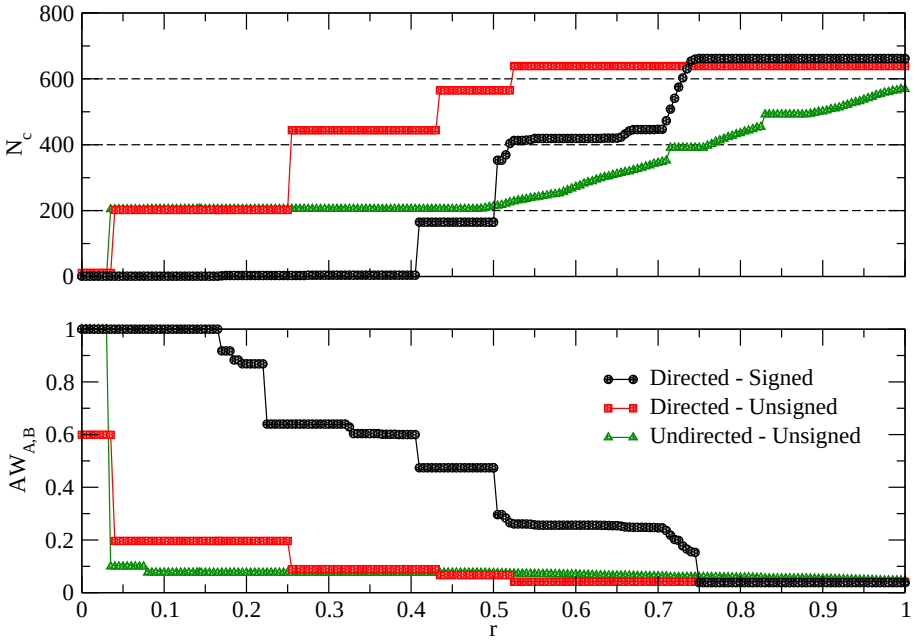


FIGURE 5.1: Top: number of detected modules N_c as a function of the normalized rescaling parameter r. The mesoscale has been screened only for a range of r yielding an interpretable amount of modules. DU and DS topologies show more than one persistent mesoscopic plateau, whereas the UU topology only has a single plateau made up of around $N_c \approx 200$ communities. Beyond $r = 0.5$, no other stable plateau can be found for this topology. Bottom: the detected community structures for the three versions of the TRN of *MTB* are compared to the functional partitions in Tuberculist [325]. DS is the only network that shows significant values of similarity, in terms of the Asymmetric Wallace Index, against the functional partition for a large range of r values.

To address this last question, we asked whether the partitions inferred by our method group genes with similar biological functions. The reason underlying this possibility is that genes within a topological community are connected among them by more regulations than average. This fact should imply that they tend to transcript together, as a response to common stimuli and eventually, to perform closely related functions. To do this, we compared the identified communities to the functional classification provided in the Tuberculist database [325]. There are many metrics and indices to compare two clusterings [326, 327, 328, 329, 330]. However, we need to rely on an

index that does not severely punish different resolution scales: our reference partition categorizes genes in only $N_c^F = 7$ groups, which yields a coarse-grained functional classification of the genome of *MTB*. We note that, for the comparison between topological partitions and functional classification, only genes belonging to truly structural functions have been considered. So, "conserved hypotheticals" and "unknown" genes have been excluded, as well as genes with regulatory roles ("regulatory proteins" and "information pathways" genes), which are expected to join transversally the TRN. Any partition with significantly more modules will show low resemblance to the functional one if the index is biased toward literally similar partitions. Thus, we present our results using the Asymmetric Wallace Index AW, which shows the inclusion of a partition into the other. The Asymmetric Wallace Index [331] is the probability that a pair of elements in one cluster of partition A is also in the same cluster of partition B. Let be a clustering A with c_A communities and a clustering B with c_B communities, and let us define the confusion matrix M whose rows correspond to the communities of the first clustering (A) and columns correspond to the communities of the second clustering (B). Let the elements of the confusion matrix, $M_{\alpha\beta}$, represent the number of common nodes between community α of the clustering A and community β of the clustering B; the partial sums being $M_{\alpha\cdot} = \sum_\beta M_{\alpha\beta}$ and $M_{\cdot\beta} = \sum_\alpha M_{\alpha\beta}$. Then, $AW_{A,B}$ (how much partition A is embedded in B) is defined as follows:

$$AW_{A,B} = \frac{\sum_{\alpha=1}^{c_A}\sum_{\beta=1}^{c_B} M_{\alpha\beta}(M_{\alpha\beta}-1)}{\sum_{\alpha=1}^{c_A} M_\alpha(M_\alpha - 1)}. \tag{5.3}$$

The Asymmetric Wallace index can also be defined the other way around ($AW_{B,A}$), but in this case this is not considered, because detected partitions are systematically more divisive than the functional one, i.e. we are interested in seeing how detected partitions are embedded in the functional one.

Figure 1 (bottom) shows the results for the proposed scheme. Initial results (early r) for the UU and DS networks are artificially high, because $N_c < N_c^F$. Besides this, the plot indicates that only the partitions obtained from the DS topology are significantly similar to the functional one. In fact, beyond the initial stages of the resolution levels, both DU and UU's community structures are far from being embedded in the functional categorization. Quite surprisingly, resolution levels with similar N_c do not entail similar $AW_{A,B}$ values. For instance, the three topologies show at some point a plateau with $N_c \approx 200$. But $AW_{UU,F} \approx 0.1$, $AW_{DU,F} \approx 0.2$ and finally $AW_{DS,F} \approx 0.5$.

These results suggest that the more complete knowledge about link attributes, the richer representation of the mesoscale, in which different levels of topological coarse-graining can be well identified, with possible bio-dynamical implications that need to be explored.

5.3 Motifs significance robustness vs. network growth

Exhaustive search of topologically common footprints and systematic differences between different real systems constitutes an important topic in network theory since its very beginning [332]. Along these lines, the classification of networks in families bring light into the evolutionary principles that ultimately yield to the complex topologies that real, evolving systems like TRNs show today [90]. In this sense, the work by Alon and coworkers [267] constitutes a milestone.

In their work, the statistical significance of 3-nodes motifs –triads– was analyzed. The number of appearances of each of the thirteen possible directed structures in real systems was compared to those observed in a null model. The null ensemble was constructed by randomly rewiring the links of the original networks, preserving the number of single links and mutual interactions (as it is done in [267]). The statistical significance of each motif h is then defined as the Z-score of its number of appearances when compared to the results found in the null ensemble:

$$Z_{\mathrm{score_h}} = \frac{n_h - \langle n_{rand,h} \rangle}{\sigma_{rand,h}} \tag{5.4}$$

Therefore, computing the Z_{score} for all possible triads in a network yields a 13-dimensional vector that, when normalized, represents the so-called triad significance profile (TSP). From the analysis of different systems' profiles, four superfamilies were identified with common TSPs: two families of non-biological networks –semantic adjacency words maps and social systems– and two families of biological, information processing networks.

Regarding the two biological networks superfamilies originally identified, TRNs of three unicellular organisms were found to conform the first one: yeast, *B.subtilis* and *E.coli*. In Figure 2, panel A, we plot the TSPs that belong to two of the four datasets analyzed by the authors in their original work: yeast [333] and *E.coli* (available at the authors' web site [313]). The second group contains developmental TRNs of eukaryotic cells belonging to pluricellular organisms, signal transduction maps and synaptic networks. In Figure 2, panel B, the TSP of the synaptic wiring map of the nematode *C.elegans* [334] is plotted, as an example of this second superfamily, evidencing the differences with respect to panel A.

The biological interpretation of the emergence of the two superfamilies of TRNs –or more generally, bio-information processing networks– proposed in [267] has to do with the typical response times developed by each group of systems. These times are similar to those of single interactions for the networks in the first group (rate-limited networks) but remarkably greater than characteristic interaction times for the systems within the second superfamily (unrate-limited networks).

The recent addition to this scheme of the TRN of *MTB* poses an intriguing question. As it is visible to the naked eye in Figure 2 (panel B) its TSP, although belonging to an unicellular organism, has a greater correlation with the representative of the unrate-limited superfamily. The fact that *MTB* has these developmental-like topological features at its TRN might be interpreted under a coherent biological picture [3].

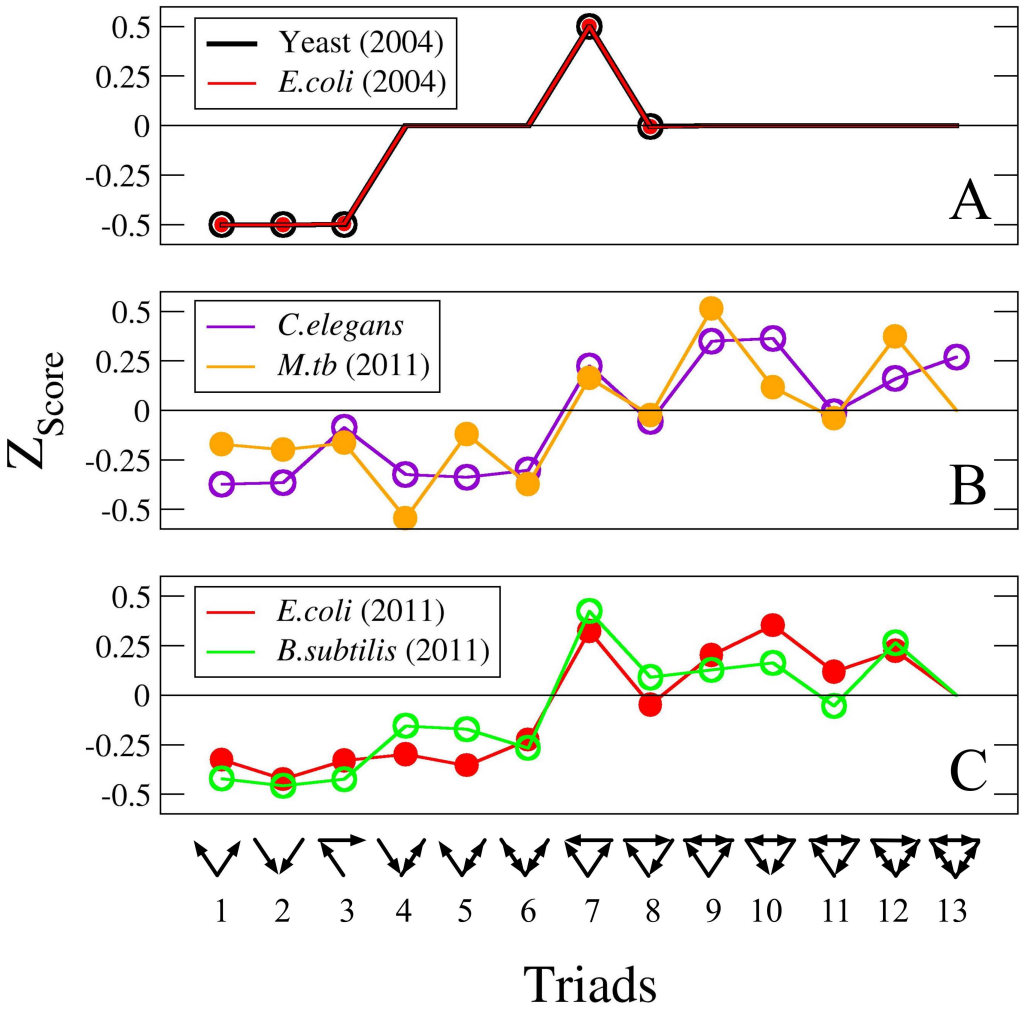


FIGURE 5.2: Panel A: *E.coli* (2004). TRN of *E.coli* as firstly published in [313, 269] ($N = 423$ operons, $L = 519$ regulations plus $SL = 59$ self-regulations). Yeast (2004) TRN of budding yeast [262] ($N = 688$ genes, $L = 1079$ regulations [313, 335]). Panel B: *C.elegans* neural network [334] ($N = 279$ neurons, $L = 2990$ synapses). *MTB* TRN [3] ($N = 1624$ genes, [336] $L = 3169$ regulations plus $SL = 43$ self-regulations). Panel C: *E.coli* (2011). Updated TRN of *E.coli* based on RegulonDB, release 7.2, dated on May, 2011 ($N = 1037$ operons, $L = 2574$ regulations plus $SL = 113$ self-regulations). *B.subtilis* (2011), updated TRN of *B.subtilis*, based on DBTBS database [315], (accessed in October, 2011) ($N = 814$ operons, $L = 1294$ regulations plus $SL = 80$ self-regulations).

The pathogen has an evolutive history tightly bound to its condition of a human intracellular obligate parasite, which could eventually have caused an adaptation of the bacterium to the rhythms and response dynamics of host cells. Indeed, certain stimuli, like hypoxia, yield anomalously slow shifts in *MTB* gene expression patterns, which can take as much as 80 days until stabilization [89].

The third panel in Figure 2 invalidates the previous hypothesis, and presents the TSPs of the updated TRNs of two bacteria which were initially characterized as rate-limited according to their TSPs. Visible at a glance, the update of the datasets has shifted their TSPs from one superfamily to another, in a way that suggests that the division of the information processing networks into two groups was an effect of data incompleteness.

The key of the change observed in the TSPs stems from the small number of two nodes feedback loops (FBLs) that are observed in unicellular organisms TRNs. Indeed, this possibility was already foreseen in [267] (see footnote 12 there). When FBLs are absolutely absent from the system under study, as the randomizing algorithm preserves the number of them, FBLs will also be absent in the null ensemble. This situation makes the Z-scores associated to triads 4, 5, 6, 9, 10, 11, 12 and 13 undefined, as in Figure 2, panel A. As time goes by, such cases have become obsolete: new links have been discovered and added to the growing datasets, and some of them generate FBLs, which are now present in the triads listed before. In the three updated systems studied, we have found as many as 12 FBLs in *E.coli* TRN, 9 in *B.subtilis* and 6 in *MTB*. The result, after the incorporation of these new FBLs, suppose that the division between two superfamilies of biological information processing networks according to their TSPs disappears, affecting the biological interpretation about the eventual relationship between time responses and motifs statistics.

Beyond the discussion on the robustness of motifs statistics that is faced here with a similar spirit of other previous works [317], much has been written about the eventually deep biological implications of anomalous network motifs' statistics as a ubiquitous, topological property of gene regulatory networks. On the one hand, environmental evolutionary adaptation has been claimed to lie underneath this ubiquitous topological treat in gene regulatory circuits [275]. According to this point of view, different environmental requirements could exert different evolutionary pressures to gene expression dynamics which may be correlated to network‘s topologies at the level of motifs, each of which is believed to offer different dynamical performances, as it has been observed in several precise cases [275, 269, 337, 273]. Complementarily, recent theoretical studies have addressed how functional, artificial networks required to drive different dynamical functions yield divergent motifs contents [338].

However, as it has been stressed in several works, evolutionary pressures are not the sole mechanism able to generate not-random statistics in networks motifs. Simple models incorporating spatial distribution of nodes [339] or typical mechanisms of network growth assimilable to those which drive gene-regulatory changes upon evolutionary time [340] have been found to generate network motifs without any evolutionary pressure. Under this kind of interpretation, network motifs could appear, not as a consequence of environmental adaptation but rather as a side-effect of some “intrin-

sic constraints" related to typical mechanisms of genetic material transformation like DNA fragments duplication, deletion, inversion etc [341]. Supporting this hypothesis, a simple but powerful argument is often put forward: topological-bias at the level of TRNs could hardly be a consequence of dynamics-based, natural selection, as in a vast amount of cases transcriptional regulatory mechanisms constitute only one layer of more complex regulatory pathways also coupling translational and post-translational interactions, which are the ultimate responsible of the complex gene expression dynamical patterns observed in the cell [342]. However, comparisons between motifs in gene regulatory networks of different bacteria which should have suffered the effects of entirely comparable "intrinsic constraints" yield slight "fine-tuning" differences in motifs statistics that can be reasonably related to environmental adaptation [3].

In this chapter, we do not intend to introduce any additional argument in the debate, which can hardly be considered closed. The reason may be that, as it has been pointed elsewhere, intrinsic constraints and evolutionary pressures are not, definitely, mutually exclusive mechanisms of network transformation [341], and to quantify the relative relevance of each mechanism could result in even a harder task. Our main purpose in this section is, however, to increase our understanding [317] about the robustness of certain topological treats of gene regulatory networks against data incompleteness, as well as to warn about how this analysis affects the network taxonomy scheme proposed in [267].

5.4 Systematic correlations between topology and experimental evidence

Experimental techniques used in transcriptional regulation inference are numerous and often subtle [343]. However, usual approaches can be grouped within two main categories. The first approach is based on the explicit detection of the physical protein-DNA interaction between regulators and promoters of target genes. This presents the advantage that only direct operations of regulators on targets can be observed. However, the existence of a protein-DNA interaction under certain in-vitro conditions does not guarantee that it is physiologically relevant in terms of target expression levels.

The alternative approach is essentially based on the generation of mutant strains in which the functionality and/or the expression levels of a certain binding factor are significantly altered with respect to those of the wild type. Then, expression levels of genes which are potentially regulated by the binding factor under study are registered and compared between wild type and mutant strains. In this way, if these different levels of regulator activity yield significantly different target expression measures, one might assume that the regulator is actually acting on the target.

The main advantage of the latter approach is that the sign and strength of the interaction can be determined. However, the analysis cannot distinguish direct regulatory interactions from indirect influences regulator-target mediated by secondary regulatory pathways. Nonetheless, as it can be seen in Table 1, this second kind of methods is responsible for the characterization of an important fraction of the links

in our systems. Therefore, a relevant question is whether or not the appearance of indirect, spurious links (as if they were real interactions) might suppose a systematic error responsible of topological bias at a global level.

	E.coli	**B.subtilis**	***MTB***
BA	914	499	726
TELC	1272	856	1290
WCLs	656	323	191
PCLs	1044	262	1344

TABLE 5.1: Number of links reported based upon binding assays (BA) or target expression levels comparison (TELC). Well characterized links (WCLs) are those characterized at least by one methodology of each group (for an exhaustive list of the experimental methods in each group see [279]). Poorly characterized links (PCLs) are reported under methodologies that can not be considered within neither of the two main groups (too generic methods, orthologies based deductions, absence of experimental support etc.). Whilst *B.subtilis* and *E.coli* are relatively well characterized systems, in order to get enough statistics for the *MTB* case, we consider identification of consensus sequences as binding proofs.

These hypothetical spurious interactions should appear connecting nodes for which a secondary regulation pathway exists, and its sign should be the same of that secondary route (see Figure 3). So, in our networks, we can identify those "suspicious" links (SLs) connecting nodes for which some secondary via has been already registered, and verify for sign coherence. We will restrict our analysis to those secondary pathways formed by a two-links cascade. The question is how we can know whether this subset of SLs presents a higher rate of spurious links than on average. Indeed, among the topologically SLs, only those that are characterized by at least one technique of each methodological category –henceforth referred to as well-characterized links (WCLs)–, can safely be considered as non-suspicious direct regulations.

Therefore, the idea is to compare the proportion of WCLs within and outside the subset of topologically SLs using Fisher's exact test (see Table 2). As it can be seen, SLs systematically present a slightly lower proportion of WCLs than non-SLs, which could be associated to random fluctuations with respect to the average values with probabilities lower than 2% in each of the systems, being remarkably lower in the case of the TRN of *E.coli*.

	E.coli		**B.subtilis**		***MTB***	
	WCLs	No WCLs	WCLs	No WCLs	WCLs	No WCLs
SLs	104	628	43	188	11	252
No SLs	552	1285	280	774	180	2141
H_o p value $< 10^{-17}$			H_o p value < 0.007		H_o p value < 0.019	

TABLE 5.2: Null hypothesis H_o assumes that the proportion of WCLs is the same for SLs or no-SLs. Only signed links are considered.

This indicates that SLs constitute a topologically defined subset of interactions

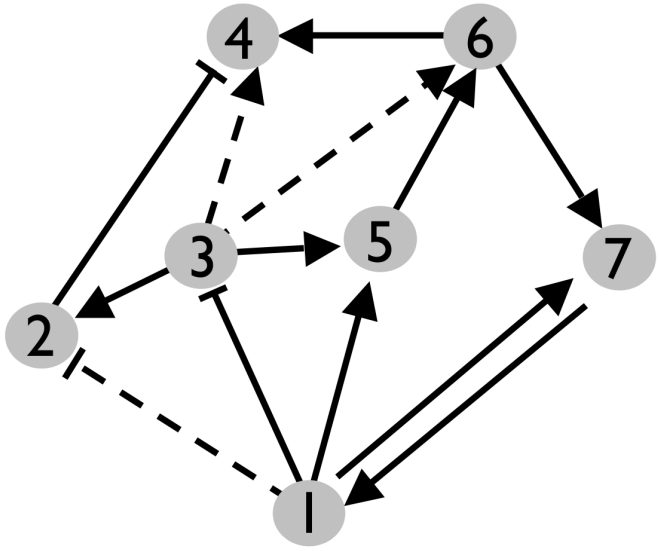


FIGURE 5.3: Arrows represent activations and right angles inhibitions. Dotted lines correspond to the links that are topologically suspicious, i.e., those for which a secondary regulatory pathway mediated by a third gene is registered and whose sign is coherent. For example, the link connecting nodes 1 and 2 is considered suspicious because of the existence of the pathway 1 to 3 and 3 to 2. The same happens for the link between nodes 3 and 4. In both cases, the condition that the product of the two links making the secondary pathway should coincide with that of the link being considered as suspicious is verified. In this sense, we have also included a case in which the latter condition does not hold: the edge linking node 1 to node 5 is not suspicious because although a two-nodes cascade connecting the same nodes exists, (i.e. 1 to 3 and 3 to 5) it is not sign coherent.

which is systematic and significantly less reliably characterized than on average in all the systems under study. This observation is in agreement with the hypothesis that insufficient experimental methods of transcriptional regulation inference can suppose the systematical observation of topologically-biased spurious links. The problem addressed here seems to critically affect the characterization of the activity of sigma factors. In fact, when we reconstruct the networks under study by considering only transcription factors as regulators and exclude sigma factors, the whole picture significantly changes. Indeed, the percent of SLs which are better characterized is even greater than the background, both for *B.subtilis* (45.0% vs 42.2%) and for *E.coli* (46.0% vs 43.1%). For the case of *MTB*, the analysis can be hardly conclusive due to the loss of statistics after sigma factors removal (no WCL is located within the set of suspicious interactions, now, less than 100 in the whole signed network). These findings, put together, suggest that characterization of sigma factor regulons is more sensitive to the aforementioned issues.

Another issue of interest is whether this experimental bias is topologically relevant.

More precisely, we question if this systematic error could quantitatively affect motif statistics in our systems. The key is that these spurious interactions could recurrently transform some motifs into others, and more precisely, focusing on most prominent motifs in number of appearances, this would suppose the systematic, spurious transformation of three-nodes cascades (triad 3) into coherent feedforward loops (FFL, triad 7). To test the robustness of the TSPs to the presence of spurious links, we delete in each network a fraction of partially characterized SLs up to the point in which the proportion of WCLs among them is comparable to the average background level. This suppose the removal of 324 SLs in the TRN of *E.coli*, 59 in the TRN of *B.subtilis* and 114 for the *MTB* case. The links to be deleted are randomly chosen within the set of partially characterized SLs. Finally, we recalculate the Z_{scores} of all the motifs and compare TSPs with their original values. The results of this process are shown in Table 3, where it can be seen that the statistical significance of cascades and FFLs are systematically affected. The interesting fact is that, after the correction, cascades are yet significantly underrepresented while FFLs as a whole (i.e. independently of the signs) continue to appear much more frequently than expected by random. Obviously, the Z_{score} associated to the other triads also varies. But the striking point is that, after normalization, in all the systems the effect of the correction on the TSPs are very limited, as we can see in Figure 4. So, the conclusion is that this kind of systematic error, although modifying the absolute values of motifs' Z_{scores}, does not affect their ratios that are recovered after normalizing the TSPs.

	E.coli	**B.subtilis**	**MTB**
Cascade (original)	-4.7 ± 0.2	-6.9 ± 0.4	-2.2 ± 0.1
Cascade (corrected)	-2.9 ± 0.4	-6.9 ± 0.8	-1.1 ± 0.3
FFL (original)	4.7 ± 0.2	6.9 ± 0.4	2.2 ± 0.1
FFL (corrected)	2.5 ± 0.7	7.0 ± 0.8	1.1 ± 0.3

TABLE 5.3: Changes in the Z_{scores} of cascades and FFLs due to systematic mischaracterization of Sls.

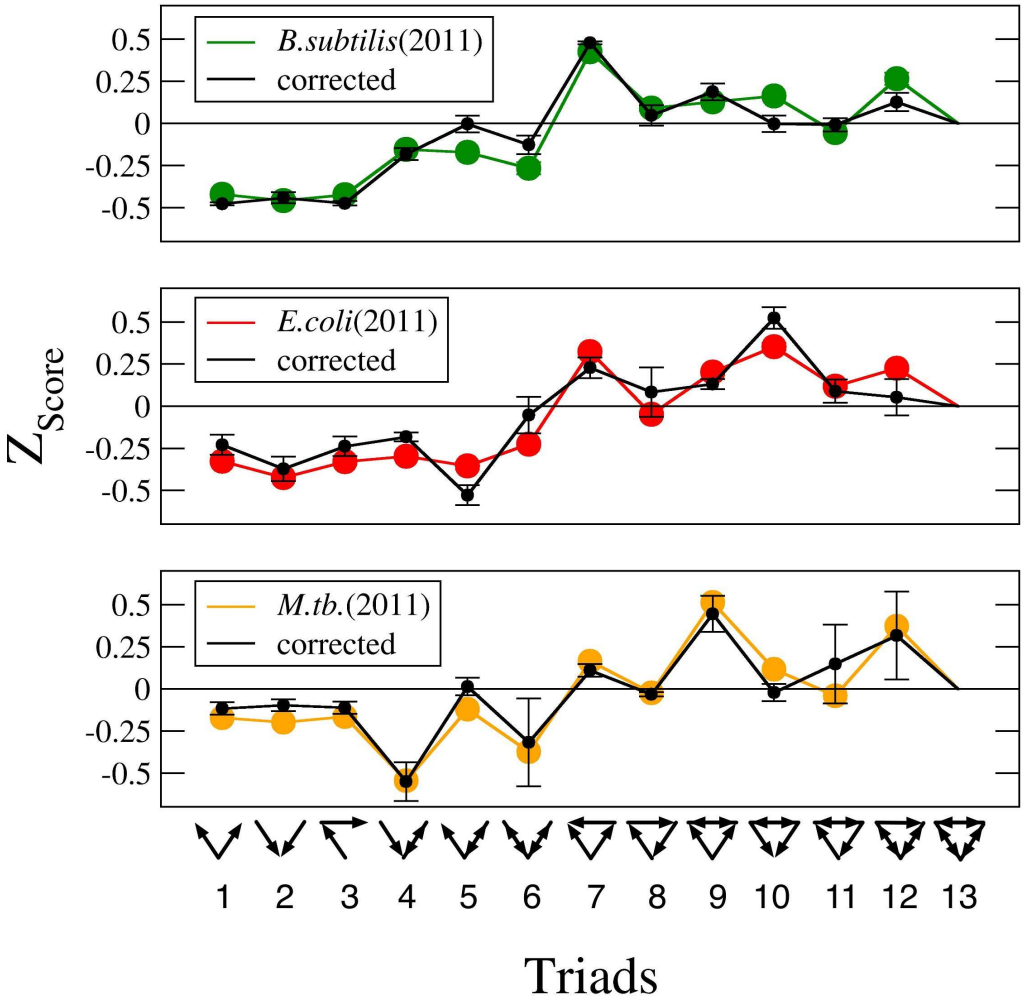


FIGURE 5.4: Original systems present a lower proportion of WCLs than the background in the set of links that connect nodes for which a secondary coherent regulatory pathway has been registered. In corrected networks, the proportion of WCLs is paired to background levels by randomly deleting poorly characterized interactions. This, however, do not affect the TSPs significantly

5.5 Conclusions

As we have shown here, sources of unreliability can be of diverse nature: from the often unjustified lack of details in link attributes to the lack of key interactions, whose inclusion radically modify motifs' TSPs. As a matter of fact, our first finding convincingly shows that data incompleteness could exert a relevant influence on the topological characterization of the mesoscale in prokaryotic TRNs. More precisely, we have shown how a complete knowledge of link attributes (directions and signs) can yield richer mesoscale structures in TRNs. Secondly, we have also shown that a mere updating of the interactions that make up a TRN in which key regulatory interactions are incorporated, radically modifies previous results based on the analysis of motifs appearances. In fact, some of the previous conclusions do not hold anymore. We have observed that prokaryotic TRNs show motifs significance profiles very similar to those belonging to multicellular, developmental TRNs, signal transduction and neural systems. Finally, experimental mischaracterization of the links has also been studied, and yet, we have found that its influence on motifs statistics is reduced. These results suggest that the evolutionary interplay between topology and dynamics is more similar between regulatory systems of multicellular and unicellular organisms than expected.

TRNs have been increasingly studied during the last several years. Nowadays, however, their characterization can only be considered provisional, as they consist of incomplete annotations of often heterogeneous and unreliable experimental evidences, computational inferences and theoretical predictions. While working with still incomplete networks could be of valuable help to uncover unknown biochemical pathways, there are situations in which reliable conclusions cannot be obtained. Moreover, we don't even know when the latter is the case. Accuracy and robustness of the results thus require us to be able to assess what results are dependent on the noisy and uncertain nature of some annotated links. This is crucial if deep biological implications are to be claimed.

5.6 Methods

5.6.1 Topological analysis

We find the community structure of the networks studied using the modularity concept introduced by Newman [36]. To perform these costly calculations we have used a mixture of heuristics, including extremal optimization and Newman's fast algorithm, as implemented in [324]. On the other hand, the statistical significance of motifs has been calculated as it is customarily done [262, 267]. Finally, for an exhaustive list of the experimental methods that have been categorized in different groups, see [279].

5.6.2 Experimental methods for link characterization in prokaryotic TRNs

Experimental methods used in transcriptional regulation inference are numerous. For the three updated networks analyzed in this chapter [314, 315, 3], we have listed as many as 25 methods from the dataset of *E.coli*, 73 in *B.subtilis* repository and 22 method annotations for *MTB* (*MTB*) case. A complete list of them is included in a separate file (see section 2).

Despite these relatively high numbers, methodology annotations of most of the links in the three networks belong to two main groups: target expression level comparisons (TELCs) and protein-DNA binding assays (BAs).

Concerning the first group, typical experiments have two phases. First, a mutant of the organism has to be generated, in which the regulator under study presents expression levels or functionality significantly altered with respect to the wild type. The first common option is to completely disrupt the transcription factor gene so as to impair its function in the mutant strain. In the case that the regulator is essential for cell survival, additional copies of the transcription factor gene can be attached to a strong promoter and inserted into the organism within plasmids in order to produce a mutant in which the regulator has an enhanced expression level. The second phase consists of studying the expression of other genes, in order to identify those whose expression is modified as a consequence of the changes associated to the transcription factor. In order to do that, mRNA concentrations of a precise gene can be measured using RT-qPCR, or mRNA concentrations of large amounts of genes can be simultaneously checked using a microarray. In order to register expression changes in single genes, reporter genes are usually attached to the promoter of the target under study. By measuring the signal associated to the reporter –for example a fluorescence signal–, one obtains a relative measure of how expression levels of that gene change due to the regulator function. Alternative expression measurement approaches rely on mixture separation blotting techniques to isolate the precise mRNAs of eventual targets (northern blot), or even separate the target proteins from an heterogeneous cell extract (western blot). All the method annotations that can be assimilated to this scheme have been marked in red in the document "MethodsSummary.xls" (available at [279] and we call them TELCs.

The second approach –BAs– consists essentially of explicitly testing whether a protein-DNA interaction between the transcription factor and target gene's promoter exists. Electrophoretic mobility shift assay (EMSA) represents the paradigm of this approach to test whether a regulator binds to the promoter of a precise gene. ChiP-on-chip technique gives the same information in a parallel fashion, on a high amount of possible targets. In turn, foot-printing assays, besides proving a transcription factor is binding to the DNA, informs about the precise position of the binding site. Once the binding site of the regulator has been found –or once one has only partial evidence of it–, directed mutagenesis techniques at the cis-regulatory element can be used to prove the specificity of the interaction. Usually, a regulator recognizes specific sequences, with small variations, which also allows us to computationally infer transcriptional regulations according to the similarity of gene promoters sequences to those already

identified as binding sites for the regulator under study. This kind of evidences can be considered as binding proofs, as they are based on previous sequences experimentally proved to be actually recognized by the regulator; but they are only weak evidences. In our work, we only have considered them as binding proofs for the case of *MTB*. The reason to proceed in this way is that this bacterium is clearly the worst characterized of the three under study here, and explicit binding evidence for it are yet rare [3]. So, in order to recover a minimum statistics, we are forced to include computational inferences of binding sites as evidences in this case. If we also consider computational inferences as binding proofs in the other two systems, results do not significantly change (results not shown). An alternative way to prove the physical implication of a certain regulator on the transcription of a certain gene is to reproduce an in vitro transcription assay (or a run off transcription assay) in which one can verify whether the presence of a regulator (transcription factor or sigma factor) is required to start transcription of a certain gene(s). All the annotations associated to this kind of methods have been marked in blue at "MethodsSummary.xls" [279], and they have been associated to BAs in the analysis.

For all datasets under study, there are method annotations that can not be associated to any of the described groups for many reasons. Some of them refer to methods that can be used in combination to expression measures but also to binding assays. That is the case of transcriptional initiation mapping techniques. These methods use mRNA samples to infer the initiation sites of the genes from which they were transcribed. The explicit implication of a transcription factor in the transcription of the mRNA samples could be determined or not, using additional techniques; whilst the technique can also be used (or not) to measure the mRNA concentrations and eventually compare it to that of other set-up. Other method annotations that have been excluded from our analysis are those which are not direct references to experimental support (like previous bibliographic generic references or orthology based inferences). For last, some method annotations are too general or barely informative, or refer to measurements related to precise physiological processes in which the genes involved in the link take place. We have also excluded these annotations from our analysis.

In conclusion, we define WCLs as those which have been annotated, at least, with a method belonging to each of the two main groups: TELCs and BAs. Thus, it must be stressed that we focus, beyond the "strength" of the experimental evidences, on the information about the aspects of the regulatory interaction each evidence gives us: TELCs tell us what the influence of the presence/absence of a certain regulator on target levels is like; while BAs tell us whether the interaction is direct or whether it has not yet been found to be so.

5.7 Online data reporistory

In order to provide an open access to the methodoligal details related to this chapter and chapter 3, we have rendered on-line a data repository [279] containing the following items:

- "MethodsSummary.xls" Contains the list with all methodological annotations provided by the databases [314, 315, 3] and analyzed here

- Updated version of the TRN of *MTB* first compiled in [3]. In this update we provide interaction signs when possible. Exhaustive information ordered by experimental sources is provided in the spreadsheet "bibliographicalRevision.xls". A list of the genes of the network is found at "IDlistMtb.txt", and the network itself at "MtbTRnetwork.txt". At "readme.txt" information about the changes incorporated to this second version of the dataset can be found.

Chapter 6

Data reliability in complex directed networks

6.1 Introduction

The lack of data quality and complete information about interactions is an ubiquitous problem in most research areas where the framework of network modeling is applied, that, as we have seen in the last section, often causes relevant problems and even incorrect conclusions. For example, classical social survey methods must deal with problems like sampling biases [349], or data loss [350, 351], which can compromise network-level analyses. The problem is even more acute when moving from social to biological systems like TRNs, in which the promise of high-throughput biochemical techniques of revealing the system backbone (i.e., transcriptomes) has to deal with the inaccuracy that these methods often show (or have, historically, shown). Microarray essays −the main tool to quantitatively measure the activity of large amounts of genes in a highly parallel fashion during the last decade− constitute a paradigmatic example of a powerful, but sometimes inaccurate or not totally reproducible technique [352, 353, 354].

Focusing on the subfield of gene regulatory networks, one additional limitation to the network approach is the diversity -even conceptual- of the high number of different techniques used to infer regulatory interactions [3, 314, 6]. Lastly, the most important issue is probably the fact that the environmental conditions under which regulatory interactions take place are, in general, different for each interaction, and for a high proportion of cases only roughly known. This leads to the paradox that in many cases, reported regulations [3, 314, 358] identified through very diverse experimental techniques, and under specific experimental conditions, are rarely similar when links identified through different experiments are compared.

It is then of utmost importance to develop new ways to assess data reliability in complex directed networks, specially because most of the efforts up to now have been directed towards solving the problem in undirected, unweighted graphs [363, 360, 361, 362], and there are only few works devoted to address the problem in directed systems [365, 364]. In this chapter, we capitalize on a previous method proposed to study the very same problem but for undirected systems [276]. Specifically, we generalize the method proposed by Guimera & Sales-Pardo [276] based on Stochastic Block Models (SBMs) to the case in which links are directed, like in a regulatory network. By doing so, we are able to successfully identify missing and spurious interactions in several real-world networks.

In this chapter we compare the performance of our method to that of previous approaches [365, 364], one of which [365] we also adapt, in this case, so as to render it able to analyze self-loop containing networks. Our results point out that, when using our approach we obtain, with some exceptions, better results at predicting missing and spurious interactions, paying the prize of greater computational requirements. This exhaustive comparison allows us to give a general outlook of the problem of data reliability determination on directed networks, identifying the strengths and weaknesses of each method. Finally, we test whether the methods can be used to predict new links in a genome-wide TRN like the one we have compiled and studied in chapters 3 and 5, providing a robust methodology that could help and guide the experimental search for unnoticed regulations. Our results indicate that the approach proposed here is also the best performing method when facing this kind of situation.

6.2 Models for identifying missing and spurious links in directed networks

6.2.1 SBMs for reliability determination in directed networks

Following [276], let us suppose that we are working on a certain graph whose adjacency matrix is A^o, which is just an imperfect realization of a certainly ideal, "true" network A to which we have no access. Being X a certain measurable property of the network, we will call $p(X = x|A^o)$ the probability that, once observed the graph A^o, X is equal to x in the ideal system A. Then we have:

$$p(X = x|A^O) = \int_M p(X = x|m)p(m|A^O)dm \tag{6.1}$$

where $p(m|A^O)$ is the probability that m is the model in a class M that gave the observation A^O, and $p(X = x|m)$ stands for the probability of model m to generate networks in which $X = x$. It is worth noticing that, depending on the way the family of models M is defined, eq. 6.1 can adopt the form of a sum instead of an integral. Since the term $p(m|A^O)$ is certainly difficult to estimate, we must reformulate the problem

by using the Bayes theorem, to get:

$$p(X = x|A^O) = \frac{\int_M p(X = x|m)p(A^O|m)p(m)dm}{\int_{M'} p(A^O|m')p(m')dm'} \tag{6.2}$$

where $p(A^0|m)$ is the probability that m gave A^O among all possible adjacency matrices and $p(m)$ is the a priori probability of model m.

At this point, we need to select a class of models to integrate the former expression. The main hypothesis that lies beneath this method consists of assuming that the required family is that of stochastic-blocks-models (SBM). In the case of undirected networks, any of these SBM can be characterized by a partition P of the set of nodes into blocks, and a probability matrix $\mathbf{Q}$ such that the element $Q_{\alpha,\beta}$ defines the probability that any of the nodes belonging to the block α be connected to any of the nodes within block β. So, the probability of two nodes being connected depends only on the blocks these nodes belong to within the partition P. Note that under these assumptions, $\mathbf{Q}$ is symmetric.

In order to deal with directed networks several possibilities are conceptually feasible. Here, we propose the following variation of the model. Instead of considering one single partition P of the nodes' space, we will consider two partitions, a senders partition P_s and a receivers partition P_r. Every node i must then belong, independently, to a block in each partition: $i \in \sigma_i$ with $\sigma_i \in P_s$ and $i \in \rho_i$ with $\rho_i \in P_r$, as it is sketched in figure 6.1. The partitions just take into account the fact that in directed networks, out-going and incoming links are treated separately. Thus out-going links of node i will be determined by block σ_i to which it belongs in the partition P_s. On its turn, and in an independent way, the in-degree will be given by the block ρ_i in the other partition P_r in which the node i is located. Within this scheme, the probability of node i sending a link to node j is Q_{σ_i,ρ_j}. Remarkably, the probability of observing the opposite link is different, and equal to Q_{σ_j,ρ_i}.

This scheme, yet having the virtue of its computational tractability, conceptually captures the behavior of systems like TRNs in which the statistics associated to in-degrees are very different to those regarding out-degrees [3, 314], being both relatively uncorrelated. This can be easily understood if one considers that the biochemical properties that define the susceptibility of a protein to be regulated by others are different to those that make the protein a regulator. While the information that will ultimately define the identity and the strength of the transcriptional regulations affecting a protein reside in its promoter region, its eventual ability to bind to the promoters of other target proteins depends on the presence and identity of a regulator domain within its protein sequence. Consequently, these two eventual roles of the protein are determined by DNA sequences that are independent and that, at least in principle, can evolve separately, both in prokaryotic [90] and eukaryotic cells [366].

Links reliabilities

Each of the SBM is fully defined by determining the two partitions above and the probability matrix, hence $m = (P_s, P_r, \mathbf{Q})$. Additionally, we define the reliability of a

certain link $i \to j$ as the probability:

$$R_{i\to j} = P(A_{i,j} = 1|A^o). \tag{6.3}$$

On the other hand, let us consider a couple of nodes (i, j) so that $i \in \sigma_i$ in the senders partition P_s and $j \in \rho_j$ in the receivers partition P_r. The probability of observing a link from node i to node j in a network generated by our model is:

$$P(A_{i,j} = 1|P_s, P_r, \mathbf{Q}) = Q_{\sigma_i,\rho_j}. \tag{6.4}$$

Consequently, the probability of observing the graph A^o as a realization of the same model is given by the binomial product:

$$P(A^O|P_s, P_r, \mathbf{Q}) = \prod_{\sigma\in P_s,\rho\in P_r} Q_{\sigma\rho}^{l^O_{\sigma\rho}}(1 - Q_{\sigma\rho})^{r_{\sigma\rho}-l^O_{\sigma\rho}}, \tag{6.5}$$

where $l^O_{\sigma,\rho}$ is the number of links observed between nodes placed in σ in P_s, and nodes placed in ρ in P_r. Regarding $r_{\sigma,\rho}$ it is the maximum possible value for $l^O_{\sigma,\rho}$, that is, the product of the sizes of blocks $\sigma \in P_s$ and $\rho \in P_r$. Substituting the three last expressions into Eq. 6.2, we get, after integration over all possible probability matrices for each case, that the reliabilities of links are:

$$R_{i\to j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)\frac{l^O_{\sigma_i,\rho_j} + 1}{r_{\sigma_i,\rho_j} + 2}e^{-H(P_s,P_r)}, \tag{6.6}$$

with P_S and P_R standing, respectively, for the spaces of all possible partitions of nodes as link senders (S) and link receivers (R). Node i belongs to block σ_i in P_s; while node j is located in ρ_j at P_r. Finally, $P(P_s, P_r)$ is here the a priori probability of observing a subset of models defined by P_s and P_r, under the assumption that once partitions are fixed, all possible models that one can get by changing the probability matrices are equally probable. In addition, the partition function Z in the last equation takes the form:

$$Z = \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)e^{-H(P_s,P_r)} \tag{6.7}$$

and the hamiltonian function is:

$$H(P_s, P_r) = \sum_{\substack{\sigma \in P_s \\ \rho \in P_r}} \left[ln\left(r_{\sigma\rho} + 1\right) + ln\binom{r_{\sigma\rho}}{l^O_{\sigma\rho}}\right] \tag{6.8}$$

Up to this point, the scheme of the method is totally analogous to the baseline method for undirected systems presented in [276]. However, the generalization of the method to directed networks requires further refinements. More precisely, as it is detailed in

the Appendix, we must adopt here the following hypothesis. Let $\vec{\chi}_{P_x}$ be the vector whose components are the (ordered) number of nodes present in each of the blocks within partition P_x. We have that

$$\begin{aligned} P(P_s, P_r) = constant \;\; \forall (P_s, P_r) \;\; with \;\; \vec{\chi}_{P_s} = \vec{\chi}_{P_r}, \\ P(P_s, P_r) = 0 \;\; \forall (P_s, P_r) \;\; with \;\; \vec{\chi}_{P_s} \neq \vec{\chi}_{P_r}. \end{aligned} \tag{6.9}$$

Then, the a priori probabilities cancel out in Eqs. (6.6) and (6.7), and thus, the mathematical forms of these expressions are identical to those given in [276], except for the fact that here, sums and products are taken over the combination of two partition spaces: P_s and P_r, with the additional constraint that the only couple of partitions (P_s, P_r) that computes are those for which $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$ (See Appendix).

Nevertheless, the reliabilities sums have always the form of a canonical ensemble average, which allows us to use again a Metropolis algorithm to sample among all the possible pairs of partitions compatible with the condition $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$, those yielding to smaller hamiltonians and thus contributing the most to the sum (see Appendix). When the sampling finishes, we recover the reliabilities of all possible directed links in the network despite of their directionality –obviously, in general $R_{i,j} \neq R_{j,i}$–. Moreover, by ranking the links one can test which are the more reliable ones, no matter whether a given link was observed in our graph A^o or not.

6.2.2 Alternative methods: Zhang's approach

In order to evaluate the performance of our proposed approach for links reliabilities determination, we have to compare it with the more recent alternatives approach available in the literature. According to the first of these alternative methods, due to Zhang et al. [365], the reliability of a link is thought to be proportional to the number of bi-fans (graph formed by two senders and two receivers each one of which receives a link from each sender[262]) in which each link participates. Similarly, the reliability of a non existing link can be evaluated as the number of bi-fans that would be generated by adding the absent interaction. The so evaluated scores are integers and obviously have no absolute probabilistic interpretation; nevertheless, pairs of nodes can be ranked by their scores and, in this sense, it is a useful tool for missing and spurious links identification also.

In order to understand in depth the relationship between SBM and Zhang's method, let us recall some technical details of our approach. As it is thoroughly explained in the Appendix, in the Metropolis algorithm used by the SBM-based method, the partitions that give lower hamiltonians and thus mostly contribute to the reliability sums are those that find a better compromise between two conflicting constraints. The first of these constraints is that blocks in the partitions have to be as large as possible. The second constraint forces the amounts of links $l^o_{\sigma,\rho}$ existing between any pair of blocks $\sigma \in P_s$ and $\rho \in P_r$ to be either close to the maximum (i.e. equal to $r_{\sigma\rho}$, the maximum possible value given the size of the blocks) or to the minimum. Once said that, given a certain partition, the bigger the quotient $r = (l^o_{\sigma_i,\rho_j} + 1)/(r_{\sigma_i\rho_j} + 2)$, the bigger the reliability of the link $i \to j$ will be. In the partition depicted in figure 6.1, for example,

for the absent link $3 \to 1$, we have $r = (8+1)/(9+2) = 9/11$, while, for the link $3 \to 6$, $r = (1+1)/(12+2) = 2/14$. The example is relevant because it evidences that a pair of nodes with a high link-reliability between them also tends to form a high number of bi-fans, as it is the case of pairs of nodes between blocks $1 \in P_s$ and $1 \in P_r$. The reason is that, to have a high reliability, the number of links between the blocks involved have to be saturated, or nearly saturated, (i.e. $l^o_{\sigma_i,\rho_j}$ near to $r_{\sigma_i\rho_j}$) and that, additionally, as it has been said before, blocks tend to be as large as possible.

To show that this relationship between both methods exist, we have calculated the correlation coefficients between Zhang scores and SBM-based reliabilities. The results, given in table 6.1 for the six networks analyzed show a high correlation between the outcome of both methods: links with high Zhang-scores tend to have high SBM-based reliabilities and vice-versa. In order to perform an additional test, we can calculate the probability of any pair of nodes (i, j) to co-occur in a common block either at the senders partition:

$$P_{\sigma_i=\sigma_j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)\delta(\sigma_i - \sigma_j)e^{-H(P_s,P_r)} \tag{6.10}$$

or at the receivers partition:

$$P_{\rho_i=\rho_j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)\delta(\rho_i - \rho_j)e^{-H(P_s,P_r)} \tag{6.11}$$

where δ stands for the Kronecker delta function. For the same pair of nodes (i, j) we calculate the number of bi-fans that are generated with nodes i and j playing the receivers roles $N^{rec}_{bf}(i, j)$, and we can compare it with the expected number of bi-fans that they would generate at random in a network of N nodes given their in-degrees $k_{in}(i)$ and $k_{in}(j)$. This expected value is $N^{rec}_{bf}|_{exp}(i, j) = k_{in}(i) \cdot k_{in}(j)/N$, and the deviation of the observed number of bi-fans coming from nodes (i, j) and the expected value is $\Delta^{rec}_{bf}(i, j) = N^{rec}_{bf}(i, j) - N^{rec}_{bf}|_{exp}(i, j)$. To test whether bi-fans tend to be formed by couples of receivers that share a common block in the senders partition, we calculate the average of $\Delta^{rec}_{bf}(i, j)$ for those receiver couples having a co-occurence probability $P_{\rho_i=\rho_j} \leq 0.5$, and we compare it to the same quantity averaged on couples with $P_{\rho_i=\rho_j} > 0.5$. The results of this test, given in table 6.2, show that in most networks a larger probability of co-existence at receivers' partitions comes together with a higher number of bi-fans formed by the couple of receivers.

After these observations, the reasons behind Zhang's method performance could be reinterpreted as a simple consequence of the existence of an underlying block structure. Under this assumption, the mapping between both methods allows us to overcome one of most clear limitations in [365], i.e. its inability to evaluate self-loops reliabilities (a self-loop never joins a bi-fan). Once we have seen that the appearance of bi-fans around highly reliable links can be rooted in the underlying block structure, we notice that the structures sketched in figure 6.2, are, from the perspective of SBMs, absolutely iden-

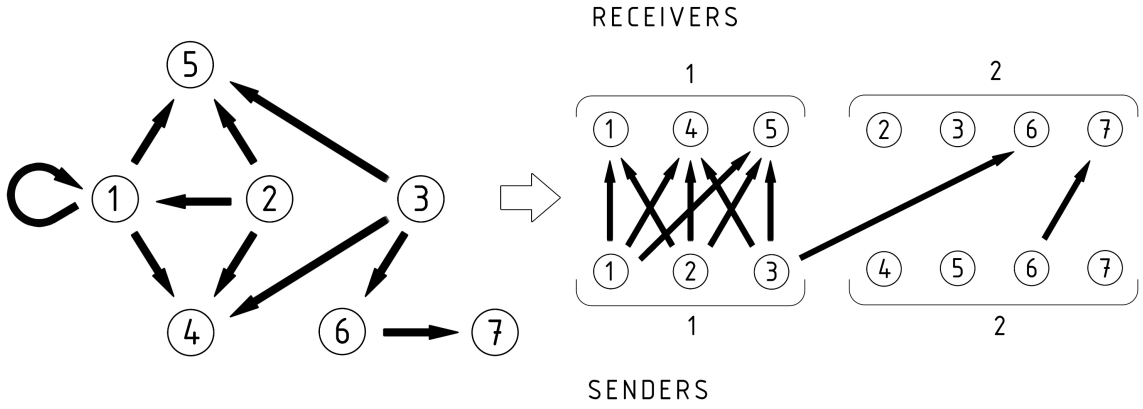


FIGURE 6.1: Stochastic block models for directed networks. In this example, partitions count with two blocks of three and four nodes, respectively. Notice that the number of occupied blocks and the number of nodes within them have to be the same both in senders and receivers partitions. The partition represented has a low hamiltonian, because blocks are large enough and the number of links between each pair of blocks $l^0_{\sigma,\rho}$ is either close to 0 or to the maximum possible value in each case $r_{\sigma,\rho}$. We have: $(l^o_{1,1}, r_{1,1}) = (8, 9)$, $(l^o_{1,2}, r_{1,2}) = (1, 12)$, $(l^o_{2,1}, r_{2,1}) = (0, 12)$, $(l^o_{2,2}, r_{2,2}) = (1, 16)$. According to this picture, links from nodes of block $1 \in P_s$ to nodes in block $1 \in P_r$ are much more reliable than any others: thus, the link $3 \to 1$ is a prototypical missing link, while the links $3 \to 6$ and $6 \to 7$ are prototypical false positives.

Network	Pearson coefficient
Radio calls	0.912
Glossary	0.917
Narragansett bay	0.904
Everglades	0.850
A.thaliana	0.920
MTB	0.943

TABLE 6.1: Correlations between SBM-based reliabilities and Zhang scores. To obtain these correlation coefficients we calculate the average SBM-based reliabilities of all pairs of nodes sharing a common Zhang score. The Pearson coefficients represented show the existence of correlations between the scores and the averaged reliabilities. For the calculations, we have considered the generalized Zhang scores, which also take into account the degenerated bi-fans in figure 6.2.

Network	$\langle\Delta^{rec}_{bf}(i,j)\rangle$ for $P_{\rho_i=\rho_j} \leq 0.5$	$\langle\Delta^{rec}_{bf}(i,j)\rangle$ for $P_{\rho_i=\rho_j} > 0.5$	p-value
Radio calls	5.42 ± 0.37	10.3 ± 2.37	$1.20 \cdot 10^{-4}$
Glossary	$(1.80 \pm 0.85) \cdot 10^{-1}$	$(3.12 \pm 1.61) \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$
Narragansett bay	$(4.47 \pm 1.10) \cdot 10^{-1}$	1.63 ± 0.26	$< \cdot 10^{-5}$
Everglades	6.85 ± 0.58	31.1 ± 2.94	$< \cdot 10^{-5}$
A.thaliana	$(-5.11 \pm 3.75) \cdot 10^{-1}$	2.28 ± 1.60	$8.48 \cdot 10^{-2}$
MTB	$(3.31 \pm 1.1) \cdot 10^{-2}$	$(1.40 \pm 0.64) \cdot 10^{-1}$	$7.63 \cdot 10^{-3}$

TABLE 6.2: Bi–fans tend to aggregate around pairs of receivers sharing common blocks. $\Delta^{rec}_{bf}(i,j)$ is the difference between the number of bi–fans in which nodes (i,j) participate as receivers $N^{rec}_{bf}(i,j)$ and the expected value of the same quantity $N^{rec}_{bf}|_{exp}(i,j) = k_{in}(i) \cdot k_{in}(j)/N$ in the null case. Positive values of $\Delta^{rec}_{bf}(i,j)$ mean that nodes (i,j) have more common in-neighbors than expected at random. In this table we show that receiver pairs with greater co-existence probabilities $P_{\rho_i=\rho_j}$ have greater $\Delta^{rec}_{bf}(i,j)$ on average. The p-values stand for the probability of the mean value of $\Delta^{rec}_{bf}(i,j)$ for the first population (pairs (i,j) with $P_{\rho_i=\rho_j} \leq$) being equal or greater than the mean of the second population. We can repeat the exercise to test the correlation between $P_{\sigma_i=\sigma_j}$ and the deviation of the number of bi–fans generated by receivers couples $\Delta^{send}_{bf}(i,j) = N^{send}_{bf}(i,j) - N^{send}_{bf}|_{exp}(i,j)$; the results are very similar; with all p-values under 20% and 4 out of 6 under 5%. Degenerated bifans have also been taken into account.

tical. Recalling the example partition in figure 6.1, the "pure" bi-fan formed by nodes $(1,2,4,5)$, appears as a consequence of links saturation between blocks $1 \in P_s$ and $1 \in P_r$, just in the same way that the degenerated structure formed by nodes $(1,2,4)$ does. Thus, what we propose here is a variation of Zhang's method in which degenerated bi-fans are treated in the same way that "pure" are, and therefore counted when it comes to evaluate the scores, even when they violate one of the main requirements of Zhang et al.'s original approach [365]. The last four rows of figure 6.3 show results obtained for networks that contain self-loops. As it can be seen, the generalization of the method has little impact on the food webs because of their low number of self-loops, while in the regulatory networks analyzed, in which self-loops are more frequent, the generalized method noticeably outperforms the original Zhang et al.'s proposal, thus supporting our hypothesis.

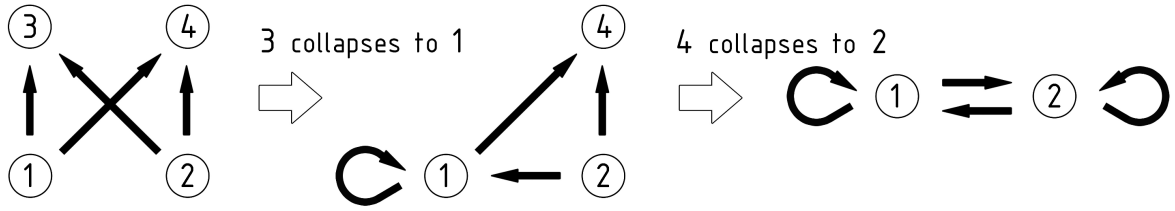


FIGURE 6.2: Bi-fan degeneration. From a "pure" bi-fan (left), if the identity of one –or both– of the couples sender-receiver coincides, we obtain these motifs, that we call degenerated bi-fans.

Network	Nodes	Time (s)
Radio calls	44	155
Glossary	67	113
Narragansett bay	32	50
Everglades	69	218
A.thaliana	15	2
MTB	65	302

TABLE 6.3: Computational time required for generating one single reliabilities rank using SBM approach. In each case, 1000 partitions are sampled.

6.2.3 Alternative methods: KronEM algorithm

Instead of Zhang's approach, –essentially based on local topological information–, the so-called Kronecker expectation-maximization (KronEM) algorithm [364], is based on a family of stochastic network models. These models have two main ingredients: a Kronecker matrix built after the expansion of a low-dimensional matrix θ verifying $\theta_{i,j} \in (0,1) \forall (i,j)$ via Kronecker multiplication by itself, and a bijection of the node set $\Sigma : i \to \sigma(i)$. In order to describe a network of N nodes, if we are working with a matrix θ of dimensions $n \times n$ (typically $n = 2$), we will need to iterate the Kronecker product k times, being k the lowest integer higher than $log_n(N)$. Once done so, the matrix elements $\theta^k_{i,j}$ –always verifying $\theta^k_{i,j} \in (0,1) \forall (i,j)$– can be interpreted as links reliabilities; more precisely, the matrix entry $\theta^k_{\sigma(i),\sigma(j)}$ is the probability assigned to the link $i \to j$.

It is worth noticing that, as exposed before, the so-constructed kronecker probability matrix has, –unless N is an integer power of n–, more rows than nodes exist in the network. This situation is used in [364] to make inferences on what they call "the hidden part of the graph", i.e. the part of the graph artificially described by the surplus of rows of the matrix θ^k. However, the adaptation of the algorithm to the problem that we face here is trivial, and allows us to focus only on the "real part" of the graph (as it is already done in [364], section 4.3).

The algorithm proposed in [364] to determine link reliabilities is based on finding, among all possible Kronecker models able to describe a certain network of adjacency matrix $A_{i,j}$, one that maximizes the overall likelihood:

$$P(A|\Sigma,\theta) = \Pi_{A_{i,j}=1}(\theta^k_{\sigma(i),\sigma(j)})\Pi_{A_{i,j}=0}(1-\theta^k_{\sigma(i),\sigma(j)}) \tag{6.12}$$

and, as it can be shown from [364] such a maximization is feasible on computational time of the order of the number of nodes. The computational requirements, as it can be seen in table 6.3, are heavier for the SBM approach, as it was already determined in [364] for the undirected version of the algorithm presented here.

The reason for this situation does not come from the behavior of the elemental step of our Metropolis algorithm itself; which indeed scales with the number of occupied blocks (i.e., at most, with the number of nodes as well). Most of the time requirements of our algorithm come from the amount of iterations needed for uncorrelating two

subsequent partitions in the sampling procedure. These decorrelated intervals depend in not trivial ways not only on the number of nodes, but also on the number of links and, in general, on the topology of the graph. This behavior explains the inexact correlations between computational times and system sizes in table 6.3.

KronEM algorithm does not present, in principle, these problems. The reason is that although both approaches are model-based bayesian methods, while the SBM approach bases its predictions on recovering a whole ensemble of stochastic models, kronEM algorithm aims at simply pick one optimal model to optimize the likelihood. This situation, yet having the virtue of reducing the computational requirements of a single run, makes the prediction of the algorithm more volatile, and sensitive to initial conditions as it was already admited in [364].

6.3 Results

6.3.1 Method accuracy

In order to check the performance of our approach, we perform a series of tests on top of different networks as in [276]. To this end, we use six well-known directed networks (see Appendix): a social network of radio calls among a closed set of operators[26], a network of hyperlinks in an on-line glossary [367, 368], the trophic webs of Narragansett bay[369] and the Everglades[370], the cell-fate determination gene network of flower development of *Arabidopsis thaliana*[371] and the regulatory network among transcription and sigma factors of *MTB* [3].

Assuming that these networks are error-free, we randomly remove a certain proportion of links. Then, we run our algorithm and rank the links reliabilities as coming out of the algorithm. We define the accuracy of the method when it comes to identify missing interactions as the probability that removed links are assigned a higher reliability ranking -i.e., they are false negatives- as compared to those that are true negatives. On the contrary, to test whether the method is able to identify spurious interactions accurately, we randomly add a proportion of links between nodes which are already senders and receivers in the original network. As before, link reliabilities are computed and the ordered ranking is used to check the accuracy of the method, which in this case is given by the (mean) probability that spurious interactions -now they are like false positives- are ranked lower than true links.

In order to evaluate the performance of our method, we compare its accuracy with two of the latest (and to the best of our knowledge the only two dealing with directed networks) alternative approaches to the problem, due to Zhang et al. [365] and Kim and Leskovec [364], respectively.

Results of the accuracy tests are shown in Fig. 6.3. In the left panels, we have represented the accuracy of the methods regarding the identification of missing links, and in the right panels, we show the accuracies related to spurious links. Black series correspond to the SBM-based algorithm presented here, red data series correspond to the method in [365] and green series to KronEM algorithm developed in [364].

As we can see, in the one hand, the SBM-based method systematically outperforms that of Zhang et al., except for the case of the social network of radio calls, for which the performance of the latter is slightly better, mainly regarding the prediction of spurious links. In fact, a deeper analysis of the two methods, as we discuss next, show that they give highly correlated outcomes. In the other hand, the SBM-based approach outperforms KronEM algorithm in eight panels, and underperforms it in three. At the last case -spurious links in the glossary network- both methods perform very similarly.

6.3.2 Guiding experiments

Once we have tested the general performance of the SBM-based method when compared to KronEM, Zhang's, and generalized Zhang's approaches, we discuss their application in an important and specific domain, that of TRNs. In this field of research, computational data reliability tools could help mitigate either the relatively poor quality and reduced size of some networks available [372, 373] or to integrate vast amounts of information coming from high-throughput experimental techniques.

On the other hand, there are several organisms, –even relevant pathogens– for which the whole TRN is not at hand, despite the fact that having the network would help in the search of new drug targets or vaccines. This is the case of the TRN of *MTB*. The bacillus of tuberculosis, responsible of one of the most threatening diseases worldwide, is probably one of the bacteria whose transcriptome has been best studied during the last years [316, 89, 3]. In 2008 the TRN of the pathogen consisted of 782 genes and 937 interactions [89], but the last updated version, published in 2011, contains as many as 1624 genes and 3212 interactions [3]. Moreover, the updated version, also added 357 new links between some of the 782 genes that were reported in 2008.

All the aforementioned facts, together with the running costs of experiments are calling for methods that could optimize the search of new interactions. To test whether and to what extent our algorithm could contribute to cure new datasets and guide the experimental search of new transcriptional relations and regulators, we perform a simple exercise with the *M. Tb* datasets of 2008 and 2011. Specifically, we check whether the appearance of the 357 links in the 2011 compilation that connects pairs of genes already integrated in the 2008 network could have been inferred from the analysis of the 2008 network itself.

To simulate the way in which our method could help to identify these new interactions, let us suppose that we are interested on a certain gene of the 2008 network and we look for undiscovered regulations it might receive from any of the regulators already present in the network in 2008 –obviously excluding those that had been already found to regulate its activity at the moment–. If no biological clue is available about what regulators are the more likely candidates to act on our gene, we are forced to experimentally try, one after another, all the possibilities. If the result of some of these experiments is positive, and so the interactions exist, we will identify them at a linear rate, as it is represented in grey in figure 6.4, panel a. In the same figure, the black curve represents the rate at which all these novel interactions are detected when the possible targets are checked according to their reliabilities calculated using

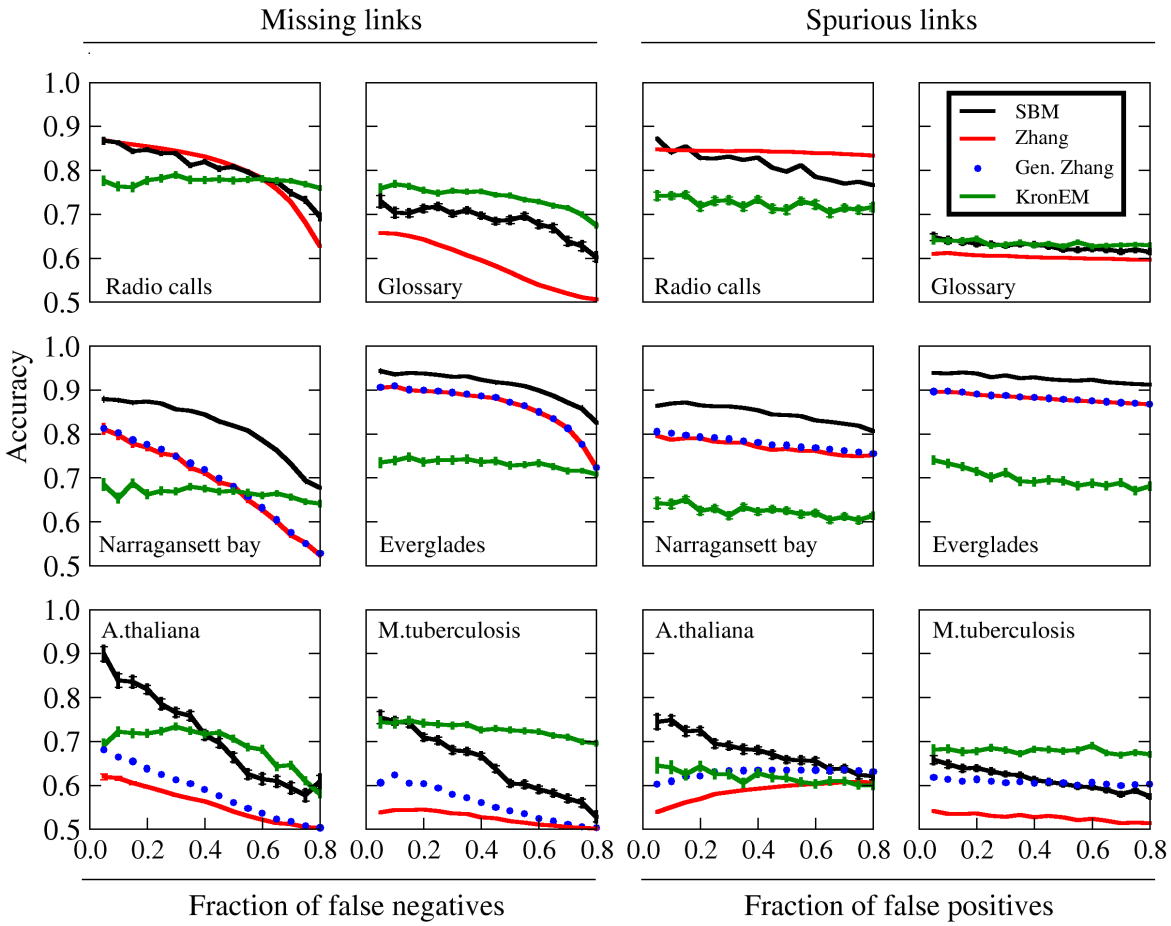


FIGURE 6.3: Method accuracy in the detection of missing (left) and spurious interactions (right) in six directed networks, according to the four methods explored in the text. The first two networks (radio calls and glossary) lack self-loops by construction, and hence, it makes no sense to generalize Zhang's approach there. For the same reason, self-links are not allowed as spurious interactions for these two networks. Not shown error bars are smaller than symbol size or line thickness.

the SBM approach. As it can be seen, the SBM-based method greatly enhances the rate at which new links are discovered, with respect to the random case but also, to a lower extent, with respect to KronEM and Zhang's algorithms .

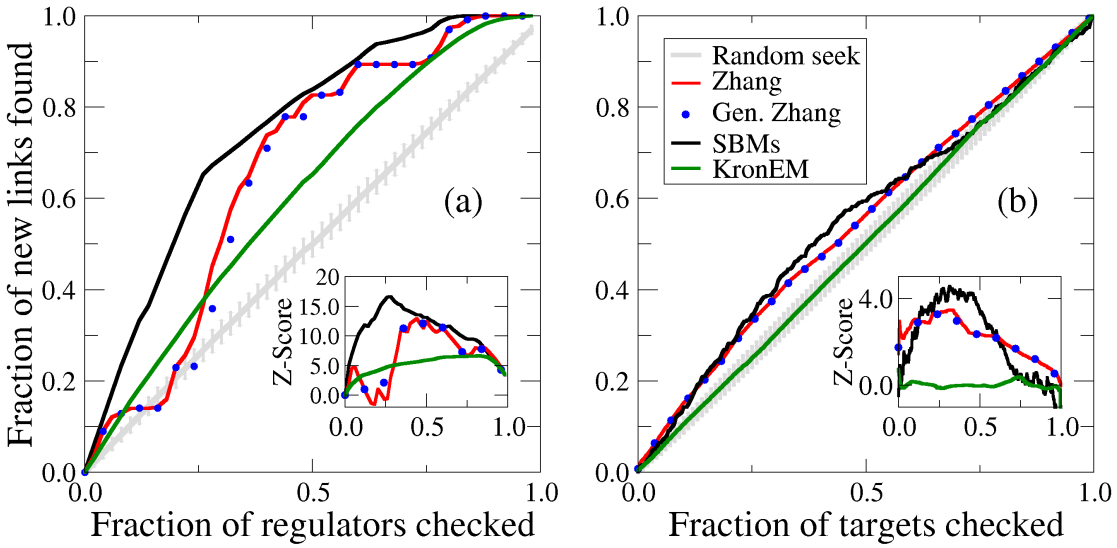


FIGURE 6.4: TRN of *MTB*: update analysis. Black: SBM-based seek. Red: Zhang's scores. Blue: generalized Zhang's scores. Green: KronEM algorithm. Grey: random seek. (a): Proportion of regulators checked versus proportion of new links found, when focusing on targets receiving the new links.(b): Regulators based search: Proportion of targets checked versus proportion of new links found, when focusing on regulators sending the new links. In the insets, the Z-Score of the methods' performance is computed, when compared to the random procedures, whose error bars ($\sigma = 1$) are represented in grey. As it can be seen, all three methods outperform the random procedure, mostly at first stages, and more remarkably in the case of target based search (panel a).

If the situation is the opposite, and we are interested in unveiling new links coming from any of the regulators of the network in 2008, the rate at which targets of the new links could be experimentally found is represented in figure 6.4, panel b, when choosing the candidates according to their SBM-based reliabilities, according to the alternative methods and when the order is random. In this case, the performance of SBM and Zhang's methods slightly outperforms KronEm, which is not better than the random procedure. These differences, though moderate, in practice could represent saving time and resources. In fact, starting from the regulators (Fig. 6.4, panel b) and aiming at finding 50% of the new targets, one has to seek the 39% of the targets with the highest SBM-based reliabilities in each case. This implies that the SBM-based method uses 78% of the time and resources needed if the identification is made randomly. Going back to the results shown in Fig. 6.4, panel a, that case produces even better results: to find the 50% of the regulations received by a target gene, one must only seek a 20%

of the total of regulators. Therefore, the SBM-based method remarkably outperforms the random search by using as less as 40% of the resources spent in the null case, but also the search orderings based on the alternative methods.

6.4 Conclusions

We have proposed an extension of the method in [276] to determine link reliabilities in directed networks. This opens the path to the potential application of our technique for the detection of missing and spurious interactions in systems as important as food-webs, TRNs or certain social networks, all of which are directed networks. A related and interesting problem that however remains to be explored is whether reliability rankings are correlated with significance measures [374, 375, 376] of the links identified. For instance, a genuine question is whether finding a highly ranked but lowly significant link is worth the computational cost involved in the calculation. We let this kind of questions for future works.

The accuracy and robustness of the method has been tested exhaustively on networks of different sizes and topological properties. Results of intensive numerical simulations have shown that missing and spurious interactions can be detected successfully, with higher precision than previous approaches in most cases. Additionally, we have numerically shown that the method can be used to guide the experimental search for missing links, as the reliability ranking resulting from the application of the algorithm to an incomplete network provides a very good guideline for experimental tests that eventually lead to the discovery of new interactions in a highly efficient way. This potentiality has important implications for our current efforts to map out transcriptional regulations, particularly, in cases such as that of *MTB*, where experimental lab protocols are very slow and expensive. At a conceptual level, this exercise makes explicit the ability of our method to predict real, arbitrarily correlated errors in complex directed networks, rather than randomly generated missing or spurious interactions.

On the other hand, after an exhaustive comparison of our method with the method proposed in [365], we have been able to provide a rationale for the latter approach: when the topology of a real system can be successfully described by a block model, bi-fans systematically appear around highly reliable links. Additionally, the mapping between both methods makes it immediate the generalization of Zhang's approach to deal with self-loops. This generalization enhances the performance of the original algorithm when self-loops are statistically relevant, as it happens in some gene regulatory networks (see figure 6.3 lower row).

Our SBMs-based model has however an important limitation. It is prohibitively costly in terms of computational time for large systems, and for sure much more expensive than Zhang's approach, or even than single runs of kronEM algorithm. Therefore, the method proposed here is mainly aimed at relatively small systems. For larger networks, Zhang's generalized method, whose outcomes are highly correlated to our approach in small systems, can be used as a low-cost resource. KronEM algorithm, in turn, represents an intermediate solution, both in terms of computational expenses

and method accuracy and consistence. This kind of situation in which the prize to pay for reaching more accurate tools also appears when facing other problems such as community detection in networks [318].

Alternatively, if more accuracy is required, we believe that the SBM-based method presented in this chapter could also be applied to subgraphs, overcoming in this way the size limitations. For instance, one can try to partition the whole system first by using one of the many algorithms available for community detection and then apply the reliability technique only to the detected communities. This kind of solutions will be explored in future work.

6.5 Materials and methods

6.5.1 Phase space

In [276], the mathematical form of the hamiltonian, in the undirected model is, as said before, equivalent to 6.8, except for the fact that there is only a partition family to sum over. Let us write it as:

$$H_u(P) = \sum_{\alpha<\beta} \left[ln\left(r_{\alpha\beta}+1\right) + ln\binom{r_{\alpha\beta}}{l^O_{\alpha\beta}} \right] \tag{6.13}$$

The restriction $\alpha < \beta$ (both blocks belonging to the partition P) appears only in order not to sum each term of the sum twice. Let's inspect the two different terms:

$$H_{1u}(P) = \sum_{\alpha<\beta} ln\left(r_{\alpha\beta}+1\right) \tag{6.14}$$

$$H_{2u}(P) = \sum_{\alpha<\beta} ln\binom{r_{\alpha\beta}}{l^O_{\alpha\beta}} \tag{6.15}$$

The first term depends, essentially, on how "concentrated" the partition is. Briefly, it is minimal when the nodes tend to concentrate in a few number of blocks. In the case of having all the nodes on the same block, Eq. 6.14 gives $ln(1+N(N-1)/2)$, where N is the number of nodes, which is approximately equal to $2ln(N)$ when N is large enough. Instead, if we have the opposite situation in which each node is assigned to a different block, then $H_1 = N(N-1)ln(2) >> 2ln(N)$. So, the term H_{1u} minimizes when the partitions are compact, and maximizes in the opposite case. As for the second term, the picture is the opposite. The presence of the combinatory number implies that, to minimize H_{2u}, the partitions of nodes should be a kind of "straight fit" for the links connecting blocks: given any two random blocks α and β, there should be a number of links between the blocks near to the maximum -the product of the block sizes, i.e. $r_{\alpha,\beta}$- or to the minimum (i.e. no link between the blocks). So, if we aim at getting the minimum of this term alone, one must go to the segregated partition in which each node belongs to a different block, for which the term directly vanishes.

Therefore, minimizing the hamiltonian implies finding a compromise between aggregation and segregation of nodes into blocks, as the two terms have clear opposite effects, and no one of the extreme situations are globally convenient. How this picture change when we move to the bipartite scheme? The addition of new degrees of freedom to the system generates an undesirable situation in which, if we perform a Metropolis algorithm letting freely evolve the two partitions, we will reach a situation in which in the P_s space, all nodes gather together into a single block, while in the P_r space we will get an split into as many blocks as nodes are. The reason is that, for the system, such configuration is globally stable, because the two hamiltonian terms, under this configuration, reach values that are far away of the possible maximum. However, in this case, the final configuration is absolutely uninformative.

The above problem comes from the fact that the system is not constrained enough and it is allowed to adopt partitions in each one of the subspaces with very different degrees of aggregation. So, we should impose a further constraint so that the system can only adopt couples of partitions with the same aggregation state (i.e. $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$), the stable. This will allow to get partitions that give rise to minimum hamiltonians being at the same time fully informative and having a compromise at intermediate levels of aggregation between links assignments and block sizes. In this case, the algorithm will be qualitatively analogous to that of the undirected case.

6.5.2 Metropolis algorithm

In order to perform our Metropolis algorithm, we start by assigning, at random, each node to one block, for example in the space P_s. Then we copy the partition generated to P_r. To ensure independence between the partitions but always verifying the constraint $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$, we proceed to randomize the partition P_r by iteratively changing the block of couples of nodes (also chosen randomly) a high enough number of times. In this way, the blocks numbered equally in both partitions contains the same number of nodes. Thus, at each Metropolis step, we choose a couple of nodes belonging to the same block in both the partitions P_s and P_r and we try to change both at the same time to the same destination block (each one on its own partition). To ensure that any couples of nodes has the same probabilities of being chosen, we proceed as follows: we start by choosing randomly one node n_1 in one partition. Then we move it to the twin block containing the very same node n_1 in the complementary partition. Inside this twin block, we randomly choose the second node to move, n_2. After the nodes n_1, n_2 are selected and tentatively moved, we recalculate H and accept the move if $H(t+1) < H(t)$. As usual, if the hamiltonian raises up, we accept the move with probability $P = e^{(H(t)-H(t+1))}$ in the standard case. Such an algorithmic scheme guarantees an ergodic exploration of the phase space, and ensures without problems detailed balance. In this way, after a certain transient, the hamiltonian reaches its equilibrium value and at that point, we start the sampling procedure, taking care that two consecutive samples are uncorrelated enough.

6.5.3 Technical aspects

While the method does not raise any problem when analyzing systems of small size (let us say $N < 200$ nodes and $E < 1000$ links approx.), as those studied in section 2, for larger networks, there sometimes appear some conceptual problems that can make the sampling procedure more difficult. First, it has been observed that the amplitude of oscillations of stationary hamiltonians, in general, increases with the size of the network analyzed. This range can be near 1000 hamiltonian units for systems of less than 2000 nodes, such as those of E.coli [314] or *MTB* [3] TRNs. Since the distribution of the hamiltonians is qualitatively normal around the average value (results not shown), the higher the amplitude of the oscillation is, the lower the proportion of samples that will contribute significantly to the sum is (let us say, those with H, at most, 10 units greater than the minimum). This problem, when it comes to analyze big networks, will force us to get a too high number of samples to get a minimum amount of relevant ones. The latter can be prohibitive in terms of computational time (recall, in addition, that the computational time of a single Metropolis step also increases with the size of the system).

Here we propose an alternative procedure that can be implemented when the networks under study are too large and computational resources do not allow a full exploration of the phase space. The alternative is as simple as discarding all the samples with $H < \langle H_{stat} \rangle - \gamma \cdot \sigma_{H_{stat}}$, where γ is a coefficient that can be chosen depending on the computational time we require and the number of samples we are looking for. This resource, although in principle could limit the performance of the method, does not affect it significantly, as showed in Fig. 6.5, panel a.

The black bars in figure 6.5, panel a, show the consistency of the standard method of sampling without any threshold. We define this consistency as the proportion of reliabilities pairs $R_{i,j}$ $R_{k,l}$ whose relative ordering is preserved in successive reliability ranks obtained with the same method. Moreover, in red bars, the comparison is made between a rank obtained with the standard procedure and another rank for which only the samples that lie over a threshold $\langle H_{stat} \rangle - \gamma \cdot \sigma_{H_{stat}}$ have been preserved and considered (here, $\gamma = 2$). Finally, the bars in blue show the internal consistency of the threshold method, that is, the mean proportion of reliability pairs whose order is conserved when we compare pairs resulting from two independent rankings generated using the threshold criterium. As it can be seen, the three measures, for the six systems shown, are consistently high and quantitatively similar between them, thus providing evidence that the threshold method could help in situations where the required computational time is prohibitively large if we aim at getting enough samplings. This kind of procedure has been used for the analysis of the TRN update represented in figure 6.4, considering 10.000 partitions with hamiltonian over $\langle H_{stat} \rangle - \gamma \cdot \sigma_{H_{stat}}$ with $\gamma = 2$.

There is an additional problem that generally appears when the networks have high mean connectivities, or, strictly speaking, when the mean connectivity is of the order of half the number of total possible links in the network, that is, in a directed network, $N^2/2$. In these cases, the information stored in the adjacency matrix is high, and so, being high the number of constraints, the dependency of the hamiltonian on

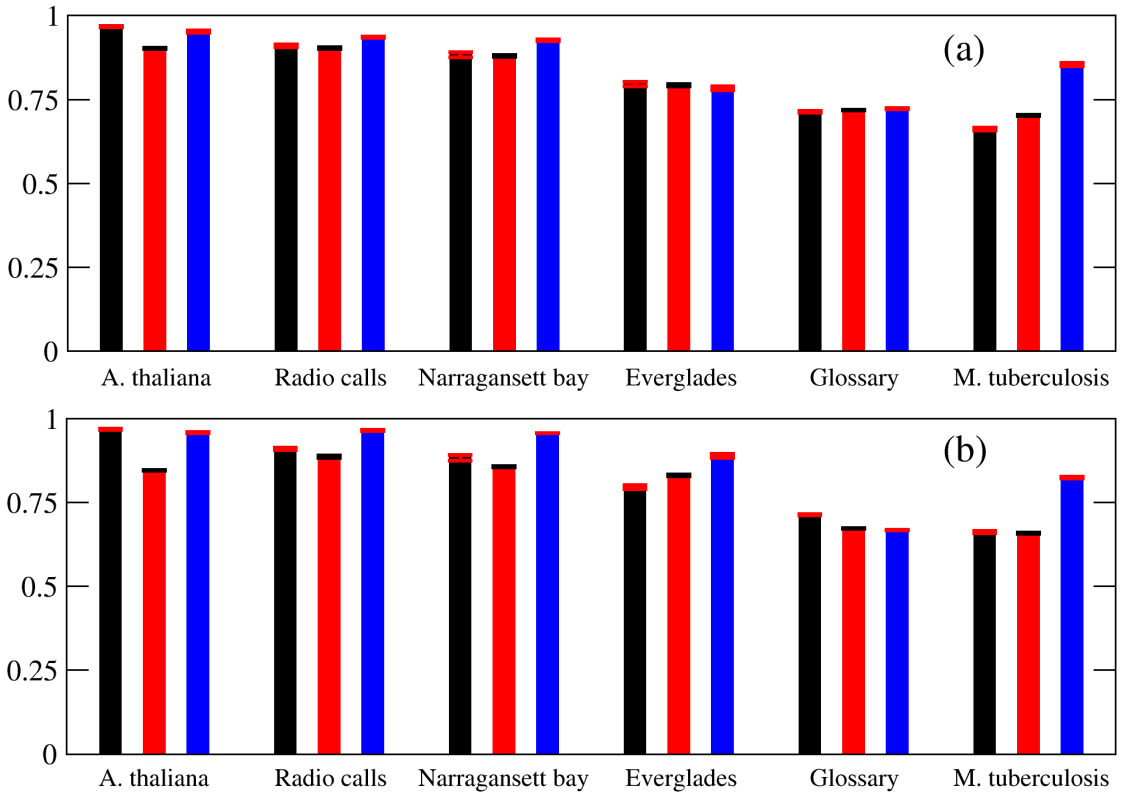


FIGURE 6.5: Coherence of ranks defined as the proportion of reliabilities that preserve ordering in successive realizations obtained with diverse sampling strategies: Panel (a): black bars: standard sampling procedure. Blue bars: Threshold sampling. We have set $\gamma = 2$. See the text for further details. Red bars: relative coherence of standard sampling ranks vs. threshold sampling ranks ($\gamma = 2$). Panel (b): black bars: standard sampling procedure. Red bars: relative coherence of standard sampling ranks vs. Hot-threshold sampling ranks ($T = 2$, $\gamma = 2$). Blue bars: Hot-threshold sampling.($T = 2$, $\gamma = 2$). See the text and Appendix for details.

the partitions defines a rough energy landscape that sometimes can become difficult to deal with. This situation can lead the system to fall into a local minimum after the thermalization process, and get trapped there. So, once arrived to the stationary state, if the basin of that local minimum is small, we will observe that the system is not able to uncorrelate sufficiently, and thus, even if its energy is small enough to consider it an acceptable minimum, our sampling will be very poor. One solution to this issue would be that of parallelizing the algorithm starting each parallel process from a random initial configuration. In this way, the process will ideally reach independent minima and thus the sampling would be N times reacher, being N the number of parallel processes.

In the above solution is is not possible, the strategy would be to introduce a pseudo temperature $T > 1$ in the Metropolis algorithm, just to ensure the system is able to

abandon local minima and explore the whole configurational space looking for other ones. The adoption of this strategy has the problem that, the higher is the temperature, the higher is also the oscillation of amplitudes of the stationary hamiltonian, and therefore the application of a threshold might also be needed.

In Fig. 6.5, panel b, we show the consistency of our method when the above strategy is implemented (using $T = 2$) in combination with a threshold criterium to select the samples, accepting only those with $H > \langle H_{stat} \rangle - 2\sigma_{H_{stat}}$. Though these operations, again, could compromise the quality of our sampling, we found that the consistency of the ranks generated with the method (Fig. 6.5, panel b, red bars) compared to those generated by the standard procedure is higher than 85%. On its turn, when we check the internal consistency of the ranks generated with the method is even better and could be greater than that reached with a standard sampling.

6.5.4 Network models

- **Killworth-Bernard radio calls network**. In their work [26], the authors asked to 44 radio operators (nodes) to rank from 0 to 9 the frequency they had used to call the rest of operators during last month. We have reconstructed our network by assigning a link when the rank associated to it was greater than 1, which produces 400 connections. Dataset available at [368].

- **Graph theory glossary network**. This network is constructed based upon an on-line glossary of definitions of technical terms about graph theory [367]. In the network, each node represents one concept; and a link points from one concept to another if the latter is linked in the definition of the former. The network has 67 nodes (5 of the 72 terms in the glossary are not connected to any other) and 122 links (114 unidirectional links and 4 reciprocal interactions). No self-link is allowed since a defined concept cannot be used in its own definition. Dataset available in [368].

- **Narragansett bay trophic web**. The system [369] originally contained 220 interactions between 35 nodes. We have removed the links involving the nodes associated to input, output and respiration fluxes, in order to take into consideration only the trophic relationships between organisms. The effective size of our system is, thus, 32 nodes and 158 links. Dataset available in [368].

- **Everglades trophic web**. The network described in [370] contains 69 nodes and 916 interactions. It describes the trophic interactions of the Everglades ecosystem in the wet season. Dataset available in [368].

- ***Arabidopsis thaliana* flower development cell fate determination network**. The network contains 15 nodes and 37 interactions among the genes that control cell-fate during the process of flower development of the model plant *Arabidopsis thaliana*[371].

- ***MTB* transcriptional regulatory backbone**. From the whole genome wide network compiled in [3], we have extracted the subnetwork that connects the transcription and sigma factors. The dataset analyzed is the giant component of that network and contains 65 genes and 130 interactions.

Part IV

Epidemic models of disease spreading on complex populations

–Brava comparación –dijo Sancho–, aunque no tan nueva, que yo
no la haya oído muchas y diversas veces, como aquella del juego del ajedrez,
que mientras dura el juego cada pieza tiene su particular oficio,
y en acabándose el juego todas se mezclan, juntan y barajan,
y dan con ellas en una bolsa, que es como dar con la vida en la sepultura.
–Cada día, Sancho –dijo don Quijote–, te vas haciendo menos simple y más discreto.

Miguel de Cervantes.
El Ingenioso Hidalgo don Quijote de la Mancha, 1605.

Chapter 7
Spreading of persistent diseases

7.1 Introduction

As it has been discussed in the introduction of this thesis, one of the main characteristics of TB spreading is a strong conceptual "bimodality" in what regards the latency periods that span between bacterial infection and development of the disease. In this sense, depending of the kind and the intensity of immune response that the host immune system performs after initial infection with MTB bacillus, the individual can suffer different fates. In most cases, after infection, the immune system succeeds at containing bacterial growth, and bacteria remain under a growth-arrest regime within lung granuloma. In such a case, the individual neither suffer any clinical symptom nor becomes infectious for long periods of time, even for his whole life. Instead, if immune system's response is not strong enough, bacteria rapidly proliferate in the lungs and the host suffers clinical symptoms and can transmit the pathogen by air [379, 382] rapidly after infection, in a process that is commonly referred to as primo-infection. In addition, latently infected individuals can, generally after an immune-depression episode, reach the active disease phase, even many decades later they were infected first.

These different phenomenologies use to be modeled in the context of mathematical epidemiology of TB by the introduction of two different latent states through which individuals pass after infection, which corresponds to the periods in which within-host infection process is evolving but the individuals neither are infectious yet nor develop any clinical symptom. These states use to be labeled as "latent-fast" and "latent-slow" [422, 22]. In some cases, specially for more schematic modeling schemes [45], and taking into account the huge differences between the typical latency periods associated to fast and slow progressions from latency to disease, the latent-fast state is omitted, and individuals suffering primo-infection are considered to directly bypass latency.

Estimating the probability of developing primo-infection after a contact, or lifetime risk of progressing from latency to disease are not easy tasks; although there exists studies that reach to often accepted estimations. So, it is commonly considered that only 5-10% of the infections directly produce active TB [379, 382], while the ranges in concerning the estimation of typical "half life" of latent state rounds about 500 y [60], which means that most of infected individuals die from causes foreign to TB. Remarkably, relevant divergences exist for some of these quantities when measured in different local settings [392, 393], and there have been observed relevant dependences for some of these parameters with individuals' age too [22]. Nevertheless, the consequences of such variations are far beyond the scope of the present chapter, and they will be further discussed in chapter thirteen.

Here, our first objective is to perform an exhaustive characterization of the dy-

namical properties of the spreading of a persistent disease whose disease's infection cycle presents the mentioned cardinal properties of TB transmission, both on homogeneous and heterogeneous populations. In order to do so, we introduce a simplified model in which latent-fast state is omitted, that is not directly assimilable neither to a pure susceptible-infected-removed scheme (SIR), nor to a SEIR (susceptible-exposed-infected-removed), but it can be conceived as an interpolation between these two cases mediated by a primo-infection probability parameter. Additionally, latency times associated to slow progression to disease are so long (from 5 centuries for the case of TB [60, 22]) that encourage us to consider open populations explicitly accounting for births and deaths caused for any cause foreign to TB in the populations under study. Very remarkably, the variability of total population's volume over time makes proportions of susceptible, latent and sick individuals with respect to the varying population size appear in the model as more suitable variables from a mathematical point of view, rather than numbers of individuals of each type -an abundant approach commonly found in the literature [48, 60, 378]-.

In what follows, we focus on the analytical characterization of the epidemic threshold of the model, that adequately reduces to those of the SIR and SEIR cases when the adequate parameters are turned off. For homogeneous systems, we also perform numerical simulations on randomly rewired networks simulating the aerial spreading of the disease within a demographically homogeneous area and show that they agree with the stationary proportions of susceptible, latent and sick individuals analytically predicted by the model. In what regards the case of structured populations, as we will see, this kind of dynamics introduce new problems essentially different from other treated before. In spite of these new problems, that constrict us to search new both analytical and numerical methods, the results achieved in this chapter are consistent and points in the same direction that previous works related to other kinds of spreading dynamics did; leading to the conclusion that virtually unbounded fluctuations in contacts networks have a important enhancing effect on the epidemic spreading also in this case of persistent infections.

The increasingly alarming situation about TB epidemiology evidences the need to increase the effort in TB research in a global way. In the context of the study of its epidemiology, new models must be developed in order to gain predictive skills; incorporating the recent theoretical advances referring to disease spreading on complex heterogeneous substrates as well as meta-population approaches and new computational tools for numerical analysis and simulation. In this sense, we aim to do a first contribution by addressing the main influences that the topological structure of social contacts in the network can exerts on its dynamics; and more specifically on one of the most important parameters in epidemiological description: the epidemic threshold.

7.2 The model

In our model, we consider that individuals in the population are compartmentalized into three groups: healthy, o uninfected -$U(t)$-, infected but not infectious -*or latently*

infected $L(t)$- and sick individuals with TB $T(t)$ which are infected and are infectious as well. The transition between these subpopulations proceeds in such a way that a healthy individual acquires the bacteria through a contact with an infectious subject with probability λ. In its turn, this newly infected individual may join the class T directly with probability p. However, the most common case is the establishment of a dynamic equilibrium between the bacillus and the host's immune system, which allows the survival of the former inside the latter. When this happens, newly infected individuals become latently infected, because despite harboring the bacteria within them, neither becomes sick nor is able to infect others, as his sputum lacks bacterial presence.

On the other hand, after a certain period of time (which may be, as we have already noted, many years) and usually following an episode of immunosuppression, the balance between the bacterium and its host can be broken. In this case, bacteria grow and the individual falls ill beginning to develop the clinical symptoms of the disease. In addition, if the infection attacks the lungs (pulmonary TB, which we consider as the only possibility), the bacillus is present in the sputum, making the guest infectious.

Once the contagion dynamics is defined, we also consider that the population size varies concurrent to the disease spreading. To take this variation into account, we add to the model demographical fluxes −births and deaths not related with the disease−. Hence, bN new individuals per unit time are added to the population −all of them healthy− and μN are removed (i.e., deceased individuals are homogeneously distributed among the three classes). According to this scheme, b represents the birth rate of the population and μ its natural death rate. Finally, latent individuals go to the active phase at a relapse rate r that represents, essentially, the inverse of the mean latency period. This implies a flux of rL individuals per unit time abandoning the latency class and entering into the active phase. Moreover, sick individuals die at a rate μ_{tb}.

As it has been pointed out before, the model here described can be interpreted like an interpolation between a simple SIR model and a chain-like SEIR model, in which any infected individuals join the SIR or the SEIR branch according to the primo-infection probability p. A formal characteristic of the model is that class R is not explicitly considered, but associated to the class of individuals who have dead by the disease.

7.2.1 Model formulation for homogeneous populations

Once we have sketched the principal characteristics of the spreading dynamics of a persistent disease like TB, and once we have outlined how to model it, one principal feature to consider is the structure of the network of contacts upon which the disease is going to spread over hosts populations.

The classical, first and simplest hypothesis regarding this particular consists of considering an homogeneous population within which each individual has the same number of contacts with all his neighbors (let us call this homogeneous connectivity β). This way, all the individuals in the system can be considered to be dynamically equivalent as a first approximation. Such an hypothesis, thoroughly studied across XXth cen-

tury since the seminal works by Kermack and McKendrick [61, 62], is commonly called mean field (MF) hypothesis, as according to it, for example, the risk of infection of any healthy individual is the same for all of them, and depends on the overall state of the population, which defines an approximated mean field related to disease burden levels.

Figure 7.1 shows the flux diagram of the different transitions that define our epidemic model, according to a MF formulation:

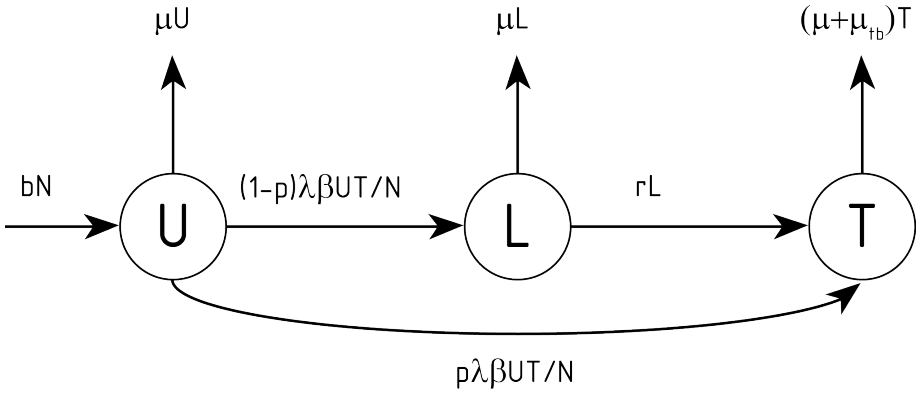


FIGURE 7.1: Flux diagram of our model for spreading of persistent infections on homogeneous populations. Labels represent the possible transitions between the compartments in which the whole population is divided according to the individuals' state. The model assumes a well mixed population of varying size. Parameters are introduced in the main text.

The corresponding mathematical description is given by the following system of ODEs:

$$\begin{aligned} \frac{dU(t)}{dt} &= bN(t) - \lambda\beta U(t)\frac{T(t)}{N(t)} - \mu U(t), \\ \frac{dL(t)}{dt} &= (1-p)\lambda\beta U(t)\frac{T(t)}{N(t)} - (\mu + r)L(t), \\ \frac{dT(t)}{dt} &= p\lambda\beta U(t)\frac{T(t)}{N(t)} + rL(t) - (\mu + \mu_{tb})T(t), \end{aligned} \tag{7.1}$$

in which:

$N(t) = U(t) + L(t) + T(t)$ represents the total population at time t,

β is the number of contacts per time unit,

λ is the probability that the bacteria is transmitted to a new host after a contact with an infectious subject,

b is the birth rate per capita and per unit time,

μ is the natural death rate per capita and unit time,

μ_{tb} is the rate of disease-related deaths per capita and unit time,

r is the transition frequency of latent infection (i.e., the probability that a latently infected individual becomes infectious),

with the closure relationship:

$$\frac{dN(t)}{dt} = (b-\mu)N(t) - \mu_{tb}T(t). \tag{7.2}$$

However, the description of the dynamical system in terms of the number of individuals is nevertheless not an optimal choice. To evaluate the impact of the spreading process on a population of varying size, it is more reasonable -and mathematically kinder- to study the temporal evolution of the densities of healthy, latent and sick individuals, rather than the number of individuals of each type. In order to do that, we define the respective densities as:

$$\begin{aligned} u(t) &= \frac{U(t)}{N(t)}, \\ l(t) &= \frac{L(t)}{N(t)}, \\ t(t) &= \frac{T(t)}{N(t)}. \end{aligned} \tag{7.3}$$

In this way, we recover a non-dimensional closure relationship $u+l+t=1$. Taking into account that

$$\begin{aligned} \frac{du(t)}{dt} &= \frac{1}{N(t)}\left(\frac{dU(t)}{dt} - u\frac{dN(t)}{dt}\right), \\ \frac{dl(t)}{dt} &= \frac{1}{N(t)}\left(\frac{dL(t)}{dt} - l\frac{dN(t)}{dt}\right), \\ \frac{dt(t)}{dt} &= \frac{1}{N(t)}\left(\frac{dT(t)}{dt} - t\frac{dN(t)}{dt}\right), \end{aligned} \tag{7.4}$$

the temporal evolution of the system in terms of densities is finally given by:

$$\begin{aligned} \frac{du}{dt} &= b - (\lambda\beta - \mu_{tb})ut - bu, \\ \frac{dl}{dt} &= (1-p)\lambda\beta ut + \mu_{tb}lt - (b+r)l, \\ \frac{dt}{dt} &= p\lambda\beta ut + \mu_{tb}t^2 - (b+\mu_{tb})t + rl. \end{aligned} \tag{7.5}$$

This kind of approach is specially suitable for open populations [48, 18], though it is not the only possible choice [60, 381]. Note that in the previous equations the natural death rate does not appear anymore, as natural deaths affect equally all the subpopulations.

This model is probably the most simple that can be built, yet resembling TB spreading dynamics at a conceptual level. An straightforward step to improve it would consist on considering such eventual recovery fluxes in our model explicitly introducing the "hidden" class R; as well as the possibility of further relapses (the so called endogenous reactivation). These phenomenologies are important, mainly for diseases like TB, which only feasible treatment in many areas consists on the supplies of large antibiotics series; the often scarce efficiency of which depends strongly on the constancy of patients and the tracking capacity of public health systems. Further refinements, like the inclusion of varieties of less infectious extra-pulmonary disease, could also have important consequences on disease dynamics.

7.2.2 Reformulation for heterogeneous populations

The model exposed in the precedent lines describes the dynamics of the epidemics in the homogeneous MF case. However, as argued in the introduction, the number of contacts of a given individual in a population can vary; which is reflected in an heterogeneous distribution of the number of contacts in the system. To account for this fact, we substitute the single parameter β, corresponding to the number of contacts of all individuals, with a connectivity parameter k, that is different for distinct individuals, following a connectivity (or degree) distribution $P(k)$. The system of equations 7.2 has to be modified accordingly. Assuming that all individuals with the same number of contacts, i.e., belonging to the same connectivity class k, are dynamically equivalent, we pass from a pure MF approach to the so-called heterogeneous mean field approximation (HMF), and the new system of differential equations is formulated for each degree class. Therefore, for a structured population, we have that:

$$N_k(t) = P(k)N(t), \tag{7.6}$$

with:

$$U_k(t) + L_k(t) + T_k(t) = N_k(t). \tag{7.7}$$

Moreover, it is convenient to reformulate the model in terms of densities, also defined within each connectivity class:

$$\begin{aligned} u_k(t) &= \frac{U_k(t)}{N_k(t)}, \\ l_k(t) &= \frac{L_k(t)}{N_k(t)}, \\ t_k(t) &= \frac{T_k(t)}{N_k(t)}, \end{aligned} \tag{7.8}$$

so that the following closure relation for any value of k is verified:

$$u_k(t) + l_k(t) + t_k(t) = 1 \qquad \forall (k,t). \tag{7.9}$$

On the other hand, the probability Θ that any given link points to an infectious individual is given by:

$$\Theta(t) = \frac{\sum_k kT_k(t)}{\sum_k kN_k(t)} = \frac{\sum_k kP(k)t_k(t)N(t)}{\sum_k kP(k)N(t)} = \frac{\sum_k kP(k)t_k(t)}{\langle k \rangle}, \tag{7.10}$$

which leads to the following set of equations that describes the dynamics within each connectivity class:

$$\begin{aligned} \frac{dU_k(t)}{dt} &= bP(k)N - \lambda k\Theta(t)U_k(t) - \mu U_k(t), \\ \frac{dL_k(t)}{dt} &= (1-p)\lambda k\Theta(t)U_k(t) - (\mu + r)L_k(t), \\ \frac{dT_k(t)}{dt} &= p\lambda k\Theta(t)U_k(t) + rL_k(t) - (\mu + \mu_{tb})T_k(t). \end{aligned} \tag{7.11}$$

Finally, the number of individuals with connectivity k evolves according to:

$$\frac{dN_k(t)}{dt} = (b-\mu)N_k(t) - \mu_{tb}T_k(t) = (b - \mu - \mu_{tb}t_k)N_k(t). \tag{7.12}$$

At this point, and building on the previous equation, it is important to point out a feature of the model: the influence of the infection dynamics on the connectivity distribution $P(k)$. First, if we add the above equation for all k, we obtain that the total population evolves as:

$$\frac{dN(t)}{dt} = (b-\mu)N(t) - \mu_{tb}\sum_k T_k(t) = \left(b - \mu - \mu_{tb}\sum_k P(k)t_k\right)N(t). \tag{7.13}$$

However, if we substitute $N_k(t) = P(k)N(t)$ directly into Eq. (7.12) and assume $P(k)$ to be constant, we would arrive to:

$$P(k)\frac{dN(t)}{dt} = P(k)\left(b - \mu - \mu_{tb}t_k\right)N(t). \tag{7.14}$$

The last expression is only compatible with Eq. (7.13) under the unrealistic assumption that all connectivity classes have the same proportion of sick individuals. We must therefore assume that the distribution of connectivity is also a function of time: $P(k,t)$, and therefore:

$$\frac{dN_k(t)}{dt} = \frac{d[P(k,t)N(t)]}{dt} = N(t)\frac{dP(k,t)}{dt} + P(k,t)\frac{dN(t)}{dt}, \tag{7.15}$$

so, if we substitute Eq. (7.15) into Eq. (7.12) we get:

$$N(t)\frac{dP(k,t)}{dt} + P(k,t)\frac{dN(t)}{dt} = P(k)\left(b - \mu - \mu_{tb}t_k\right)N(t), \tag{7.16}$$

expression from which, if we replace $dN(t)/dt$ from Eq. (7.13), we get the temporal evolution of $P(k,t)$ as:

$$\frac{dP(k,t)}{dt} = -P(k,t)\mu_{tb}\left[t_k(t) - \langle t_k\rangle(t)\right], \tag{7.17}$$

where

$$\langle t_k\rangle(t) = \sum_k P(k,t)t_k(t). \tag{7.18}$$

The last step is to reformulate all the equations so as to express all of them in terms of densities. Using the definitions of the densities given above, we rewrite the system of equations for each dynamical state and degree class as:

$$\begin{aligned}
\frac{du_k(t)}{dt} &= b - u_k(t)(b + \lambda k\Theta(t) - \mu_{tb}t_k(t)),\\
\frac{dl_k(t)}{dt} &= (1-p)\lambda k\Theta(t)u_k(t) - (b+r)l_k(t) + \mu_{tb}l_k(t)t_k(t),\\
\frac{dt_k(t)}{dt} &= p\lambda k\Theta(t)u_k(t) + rl_k(t) - (b+\mu_{tb})t_k(t) + \mu_{tb}t_k(t)^2.
\end{aligned} \tag{7.19}$$

7.3 Dynamics of persistent infections in homogeneous populations

7.3.1 Epidemic threshold

Epidemiological models are aimed at reproducing actual epidemic outbreaks as accurately as possible. Their final goal is to anticipate the course of an outbreak or even make predictions in real time, which will provide health authorities with new means to fight disease contagion. However, compartmental models in epidemiology share, despite of particularities of each model, a common dynamical outcome [43, 44, 45]. Typically, in these kind of models the parameter (phase) space is divided in two regions. In one of them, an initially healthy population remains macroscopically unaffected after the addition of a small fraction of infectious individuals, while, in the other, the disease is able to spread to affect a macroscopic fraction of the population. That is, there are two regions in the parameter space: a disease-free region and an active region. The critical epidemic point or epidemic threshold divides the two regions of the phase space. The determination of the epidemic threshold is one of the key goals of epidemiology, for this would allow designing efficient vaccination campaigns and other countermeasures [381]. This will also be our main objective. To this end, let us first characterize the fixed points of the dynamics described in 7.6, which have to verify:

$$\begin{aligned}
\frac{du^*}{dt} &= 0 = \quad b - (\lambda\beta - \mu_{tb})u^*t^* - bu^*, \\
\frac{dl^*}{dt} &= 0 = \quad (1-p)\lambda\beta u^*t^* + \mu_{tb}l^*t^* - (b+r)l^*, \\
\frac{dt^*}{dt} &= 0 = \quad p\lambda\beta u^*t^* + \mu_{tb}t^{*2} - (b+\mu_{tb})t^* + rl^*.
\end{aligned} \tag{7.20}$$

In Eq. 7.21. only two the equations are independent due to the closure relationship $u + l + t = 1$. The trivial solution for the system 7.21 is the fixed point corresponding to a disease free population: $(u^*, l^*, t^*) = (1, 0, 0)$. In order to look for not trivial fixed points, we can work out u^* from the first equation in 7.21 to obtain:

$$u^* = \frac{b}{(\lambda - \mu_{tb})t^* + b}, \tag{7.21}$$

which, after substitution in the third equation in 7.21 gives

$$\frac{dt^*}{dt} = 0 = t^*\left[\mu_{tb}(\lambda\beta - \mu_{tb})t^{*2} + [\mu_{tb}b - (\mu_{tb} + b + r)(\lambda\beta - \mu_{tb})]t^* + [\lambda\beta(pb + r) - (r + b)(\mu_{tb} + b)]\right]. \tag{7.22}$$

This equation is quadratic except for the common factor t^*, which in turns guarantees stationarity of the disease free fixed point. So, we could write down explicitly the analytical expressions of the additional two solutions t_1^* and t_2^* of the quadratic trinomial. However, if we call:

$$\mu_{tb}(\lambda\beta - \mu_{tb})t^{*2} + [\mu_{tb}b - (\mu_{tb} + b + r)(\lambda\beta - \mu_{tb})]t^* + [\lambda\beta(pb + r) - (r + b)(\mu_{tb} + b)] = \zeta t^{*2} + \eta t^* + \Theta, \tag{7.23}$$

we can more easily determine the sign of the three coefficients ζ, η and θ depending on the possible values of $\lambda\beta$:

- Main term $\zeta = 0 \longleftrightarrow \lambda\beta = (\lambda\beta)_1 = \mu_{tb}$.
- First order term $\eta = 0 \longleftrightarrow \lambda\beta = (\lambda\beta)_2 = \mu_{tb}\frac{\mu_{tb}+2b+r}{\mu_{tb}+b+r}$.
- Independent term $\Theta = 0 \longleftrightarrow \lambda\beta = (\lambda\beta)_3 = \frac{(r+b)(\mu_{tb}+b)}{pb+r}$.

which, noting that for any parameter combination, it is verified that

$$(\lambda\beta)_1 < (\lambda\beta)_2 < (\lambda\beta)_3, \tag{7.24}$$

allows us to construct the following sign table for the phase portrait description of the model's dynamics:

Let us focus firstly on the point $t^* = 1$, that apparently can appear as a dynamical attractor for the physically meaningful range $0 \leq t \leq 1$ in all the regions, with a basin of attraction that corresponds, for regions 2 to 4, to values of t that are greater than the largest solution of Eq. 7.22, that is, when $\zeta > 0$ (i.e. $\lambda\beta > \mu_{tb}$):

$$t^*_{max} = \frac{-\eta + \sqrt{\eta^2 - 4\zeta\theta}}{2\zeta}, \tag{7.25}$$

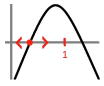
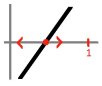
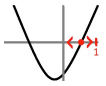
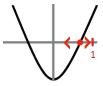
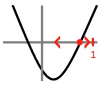
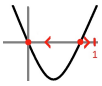
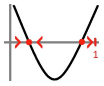
	Region 1: $\lambda\beta < \lambda\beta_1$	$1 \to 2$ $\lambda\beta = \lambda\beta_1$	Region 2: $\lambda\beta_1 < \lambda\beta < \lambda\beta_2$	$2 \to 3$ $\lambda\beta = \lambda\beta_2$	Region 3: $\lambda\beta_2 < \lambda\beta < \lambda\beta_3$	$3 \to 4$ $\lambda\beta = \lambda\beta_3$	Region 4: $\lambda\beta_3 < \lambda\beta$
ζ	-	0	+	+	+	+	+
η	+	+	+	0	-	-	
θ	-	-	-	-	-	0	+
ζ	-	0	+	+	+	+	+
$t_1^* + t_2^* = -\eta/\zeta$	+	N.d.	-	0	+	+	+
$t_1^* t_2^* = \theta/\zeta$	+	N.d.	-	-	-	0	+
Phase portrait							

TABLE 7.1: Stability characterization for t^* values at stationarity.

or with minus sign before the root, when $\zeta < 0$. In region 1, the basin of this hypothetical disease free attractor corresponds to the values of $t > t^*_{min}$, where we denote by t^*_{min} the smallest solution of 7.22. Note that, provided that $t < 1$, entering the basin of attraction leads to $\dot{t} > 0$ until $t = 1$. However, we should also consider the temporal evolution of N. More precisely, it is easy to see that, when proportion of infectious individuals exceeds the threshold:

$$t_{limit} = \frac{b - \mu}{\mu_{tb}}, \tag{7.26}$$

the spreading process is able to cause a demographical decay in the population, i.e., $\dot{N} < 0$. This behavior continues to be so until the annihilation of the whole population. Therefore, as it can be seen from Table 7.1, when the proportion of sick individuals is greater than t^*_{min} −in region 1− or t^*_{max} −in regions 2 to 4−, the proportion of sick individuals grows up indefinitely. This growth of the density of infectious individuals eventually causes that the whole population dies out. Therefore, the point $t = 1$ is everything but an stable point of the dynamics, as it leads to population's extinction. So, at least in regions 1 to 3, the only −also stable− possible fixed point corresponds to the disease free state, i.e., to $t^* = l^* = 0$.

The situation is different in region four, where there exist an additional stable stationary value for $t^* > 0$. Hence, in this region, the previous fixed point $t^* = l^* = 0$ is unstable and the transition between regions three and four defines the epidemic threshold, which is given by the condition:

$$(\lambda\beta) = (\lambda\beta)_c = \frac{(r + b)(\mu_{tb} + b)}{pb + r}. \tag{7.27}$$

As a matter of fact, there is a simpler way to obtain the threshold Eq. 7.27. The condition for the singularity of the Jacobian matrix in the vicinity of the disease-free

fixed point reduces to:

$$\begin{vmatrix} -b & 0 & -(\lambda\beta - \mu_{tb}) \\ 0 & -(b+r) & (1-p)\lambda\beta \\ 0 & r & -(b+\mu_{tb}) + p\lambda\beta \end{vmatrix} = 0, \tag{7.28}$$

that leads to the same threshold in Eq. 7.27., or expressed as it is usually found in the literature:

$$(\lambda)_c = \frac{1}{\beta}\frac{(r+b)(\mu_{tb}+b)}{pb+r}. \tag{7.29}$$

Finally, it is worth noticing that the result Eq. 7.29 reduces to the well-known threshold for the SIR model [45] when we take $b = \mu = 0$ and $p = 1$. In turn, when taking $p = 1$.

7.3.2 Numerical simulations.

In this section, we compare the analytical, stationary proportions predicted by the model −those that constitute the solution of the system Eq. 7.21 and that can be easily derived by explicitly working out t^* at 7.22 and substituting it into 7.21− with the results of numerical Montecarlo (MC) simulations.

We consider an initial population of N_o individuals distributed in the different classes. At each time step, each sick individual contacts β randomly chosen individuals. When one of these contacts is a healthy node −i.e., an individual belonging to U class− the contagion is produced with a probability equal to the spreading rate λ. In the case that contagion takes place, the newly infected node goes directly to the T class with probability p. In the complementary case (with probability $1-p$), the contagion causes the individual to enter into latency. The results that follow have been obtained using an initial population of $N_o = 1000$ individuals and we have taken $\beta = 6$. In addition to the contagion dynamics, at each time step individuals of classes U and L leaves the system with probability μ, while sick individuals die with a higher probability equal to $\mu + \mu_{tb}$. Births are also simulated by introducing bN individuals −all of them into class U−. Finally, the transition from latency to the infectious phase takes place with probability r.

Regarding the parameter values, for birth and natural death rates $-b$ and $\mu-$ we have taken as a reference the typical values of a developed country like Spain: $b = 0.01$ and $\mu = 0.009$ events per capita and year. The rest parameters −those directly related to the disease spreading− are not easily measurable on real populations. In spite of this, plausible approximations can be made, and usually, typical validity ranges are accepted in the literature [60]: $r = 0.002$, $p = 0.07$ and, finally $\mu_{tb} = 0.8$ deaths due to TB per capita and year. Therefore, the spreading rate will be our free, control parameter (in part because it is the most harder to obtain). In particular, we explore the region between $\lambda = 0.5$ and $\lambda = 1$.

Given the previous selection of parameters, it can be easily shown that we are between regions 3 and 4. So, taken into account the analytical characterization of the dynamics of the model made in the previous section, the only caution one should have in mind is that the initial proportion of sick individuals should not be bigger than t^*_{max}.

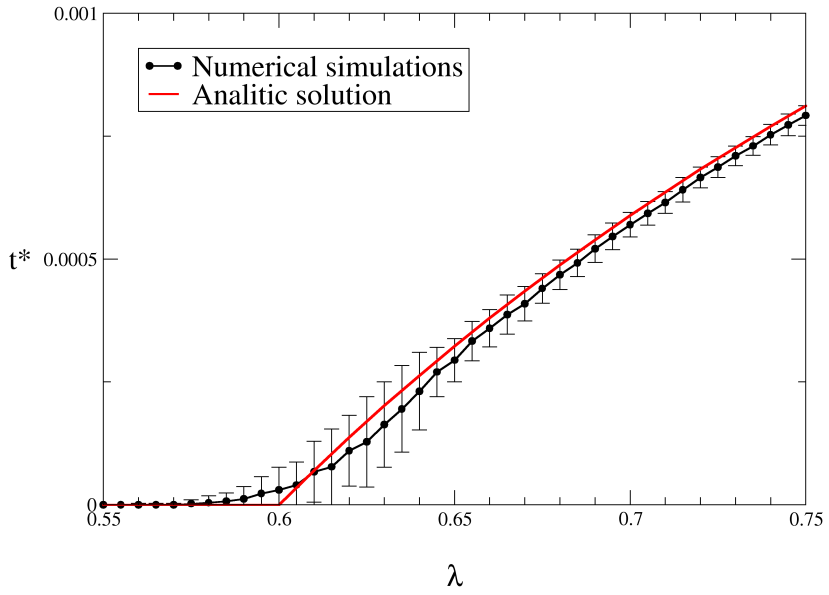


FIGURE 7.2: t^* values at stationarity vs the spreading rate. The figure shows the predicted values of t^* from the model and those obtained from MC simulations. Error bars correspond to two times the standard deviations for the 100 realizations carried out for each numerical point. Results were obtained for an initial population of $N_o = 1000$ individuals distributed as $(u_o, l_o, t_o) = (0.5, 0.4, 0.1)$. The rest of parameters are those discussed in the main text.

This will guarantee that the state $t = 1$ will not become an attractor of the system dynamics. In this sense, it can be easily calculated that, for our set of parameters, the largest solution of 7.22 is always greater than unity and thus the eventual state $t^* = 1$ will never be an attractor for the dynamics. Moreover, substituting the chosen model parameters in Eq. 7.29 predicts $\lambda_c = 0.6$.

Figure 7.2 compares the results derived from the analytical solution of the model to those obtained from MC simulations. As it can be seen, although the analytical curve is systematically above the numerical values, it lies within the limits of the error bars, and therefore both are in agreement. Regarding the epidemic threshold, MC results predict a somewhat smaller value for λ_c. Although the differences are not large and the agreement can also be considered good, it is likely that these deviations come from finite size effects and that working with larger system sizes will reduce the gap.

7.4 Spreading of persistent diseases on heterogeneous populations

7.4.1 Evolution of the degree distribution

After having successfully characterized epidemic thresholds on homogeneous populations, at this point it is appropriate to point out one aspect that will hinder any equivalent analysis on heterogeneous systems. Here, our main goal will be to calculate both numerically and analytically the critical value λ_c beyond which the population presents an endemic proportion of sicks individuals. However, we expect λ_c to be dependent on the ratio $\frac{\langle k \rangle}{\langle k^2 \rangle}$, (as in previous works [8, 57]) which is in turn a function of the connectivity distribution $P(k)$. The degree distribution, as previously shown, changes in time as the dynamics of infection progress. As we should see, we can handle this time dependence analytically, but we should be forced to design a simulation method to account for the rate of births and deaths and the effects of these two processes on the degree distribution.

The aforementioned features might lead to a situation in which the infection dynamics would modify the underlying structure of the network through which the disease is being spread. Therefore, λ_c could also vary as it is intrinsically related to the first two moments of a seemingly time-dependent degree distribution. The reason why we consider the distribution of contacts per unit time as heterogeneous, even for the current airborne-transmitted disease is based on the observation that the number of contacts a person can have per unit of time is subjected to relevant sources of heterogeneity. Firstly, what we can call *geo-demographic, macroscopic* heterogeneity, in which the number of contacts depends on the population density in the region in which an individual inhabits. Secondly, at a more *individual, microscopic* level, the heterogeneity arises because the number of contacts depends, in a region of constant population density (i.e., a town or neighborhood in a city), on the daily activity pattern of the individual within that region. These two factors will define, ultimately, the function $P(k)$. Having that said, the assumption implicitly incorporated in the system of equations 7.19 does not hold. Note that this equation implies the connectivity of individuals is hereditary and therefore that the number of births within each k class equals the birth rate times the number of individuals within each k class, $N_k = P(k,t)N$.

The above situation would be equivalent to assume that the dynamics of the disease being studied is the only one that influences the demographic structure of a population, which is false, since it is clear that there are countless cultural, economic and social factors that ultimately define the above levels of heterogeneity. We therefore assume in what follows that the individuals of each generation are distributed among the k classes according to an invariant distribution function, which we further assume to be the initial degree distribution of the original network: $P(k, t_o)$. As we shall see, this assumption besides being more plausible, has the advantage that makes the connectivity distribution to be roughly stable, and so will be the critical value λ_c.

So, we have the following reformulation of the system of differential equations (7.11):

$$\begin{aligned}
\frac{dU_k(t)}{dt} &= bP(k,t_o)N - \lambda k\Theta(t)U_k(t) - \mu U_k(t),\\
\frac{dL_k(t)}{dt} &= (1-p)\lambda k\Theta(t)U_k(t) - (\mu+r)L_k(t),\\
\frac{dT_k(t)}{dt} &= p\lambda k\Theta(t)U_k(t) + rL_k(t) - (\mu+\mu_{tb})T_k(t),
\end{aligned} \tag{7.30}$$

with the definition of the number of individuals in each class of connectivity:

$$N_k(t) = N(t)P(k,t), \tag{7.31}$$

and inside each class:

$$\begin{aligned}
U_k(t) &= N_k(t)u_k(t),\\
L_k(t) &= N_k(t)l_k(t),\\
T_k(t) &= N_k(t)t_k(t).
\end{aligned} \tag{7.32}$$

Now the total population within each connectivity class verifies:

$$\frac{dN_k(t)}{dt} = bNP(k,t_o) - \mu NP(k,t) - \mu_{tb}T_k(t), \tag{7.33}$$

so that, if we add in k, the last modification has no effect on the variation of the total volume of the population. The temporal evolution of the degree distribution is now given as:

$$\frac{dP(k,t)}{dt} = b\left[P(k,t_o) - P(k,t)\right] - P(k,t)\mu_{tb}\left[t_k(t) - \langle t_k\rangle(t)\right]. \tag{7.34}$$

Finally, the equations for the densities read as:

$$\begin{aligned}
\frac{du_k(t)}{dt} &= b\frac{P(k,t_o)}{P(k,t)}\left(1-u_k(t)\right) - u_k(t)\left(\lambda k\Theta(t) - \mu_{tb}t_k(t)\right),\\
\frac{dl_k(t)}{dt} &= (1-p)\lambda k\Theta(t)u_k(t) - \left(b\frac{P(k,t_o)}{P(k,t)} + r\right)l_k(t) + \mu_{tb}l_k(t)t_k(t),\\
\frac{dt_k(t)}{dt} &= p\lambda k\Theta(t)u_k(t) + rl_k(t) - \left(b\frac{P(k,t_o)}{P(k,t)} + \mu_{tb}\right)t_k(t) + \mu_{tb}t_k(t)^2.
\end{aligned} \tag{7.35}$$

The question of the structure and evolution of the networks of contacts relevant for transmission of airborne diseases is a subtle matter of recent interest in the literature of infectious diseases modeling. One of the factors more relevantly affecting individuals' connectivities within these networks is age [12]. In this sense, more detailed models of airborne diseases' spreading usually associates connectivity classes with age groups, [391], fully describing age-to-age mixing patterns correlations via detailed contact matrixes. Even if we are getting rid of any age-structure of our populations in our model,

the hypothesis of considering the degree distribution quasi-stationary would translate -in the context of such age-degree association- into a constant demographic structure for the populations. Such association, however, mustn't be taken too far, as our model neither contemplates age-structure nor any aging dynamics. In further sections, the influence of these aspects -very relevant for TB spreading dynamics- will be thoroughly assessed in the context of more detailed data driven models.

7.4.2 Characterization of the equilibrium points.

The previous set of differential equations tells us how the different densities of interest evolves within each connectivity class. Their corresponding macroscopic quantities are defined as

$$\begin{aligned} \langle u\rangle(t) &= \sum kP(k,t)u_k(t), \\ \langle l\rangle(t) &= \sum kP(k,t)l_k(t), \\ \langle t\rangle(t) &= \sum kP(k,t)t_k(t), \end{aligned} \tag{7.36}$$

where $\langle u\rangle(t), \langle l\rangle(t)$ and $\langle t\rangle(t)$ are the mean densities of healthy, latent and sick individuals, respectively.

Let us now go one step further and characterize the equilibrium points. The magnitudes of interest are the average densities, so that an equilibrium point $(\langle u\rangle^*, \langle l\rangle^*, \langle t\rangle^*)$ must verify by definition:

$$\left(\frac{d\langle u\rangle^*}{dt}, \frac{d\langle l\rangle^*}{dt}, \frac{d\langle t\rangle^*}{dt}\right) = (0,0,0). \tag{7.37}$$

We also impose a further constraint which is that the degree distribution of the network is stationary, that is:

$$\frac{dP(k)^*}{dt} = 0 \qquad\qquad \forall k. \tag{7.38}$$

At this point one must ask whether macroscopic stability also implies stability within each connectivity class. The answer is yes, if we also demand stability of the degree distribution. Admittedly, if we equate expression (7.34) to zero and solve for the stationary $P(k,t)^*$ we get:

$$P(k,t)^* = \frac{bP(k,t_o)}{b + \mu_{tb}(t_k^* - \langle t\rangle^*)}, \tag{7.39}$$

which shows that this value depends on the microscopic scale t_k. Therefore, the stability of the degree distribution imposes a stationary condition on t_k for all k, which in its turns extends to the other densities l_k^* and u_k^*. Hence, we have:

$$\left(\frac{du_k^*}{dt}, \frac{dl_k^*}{dt}, \frac{dt_k^*}{dt}\right) = (0,0,0) \qquad \forall k. \tag{7.40}$$

The above condition is trivially satisfied for the solution $(u_k^*, l_k^*, t_k^*) = (1,0,0)\ \forall k$, which leads to a degree distribution exactly as the initial distribution. We next analyze the stability of this solution, which shall allow us to characterize the epidemic threshold.

7.4.3 Epidemic threshold

As we have said, in this section we will study the stability of the solution $(u_k^*, l_k^*, t_k^*) = (1, 0, 0)\ \forall k$. In this point, as we have already discussed, we find that, while no latent or infected individual is introduced in the network, the degree distribution does not change in time; so that $P(k)^* = P(k, t_o)$. This situation allows us to work with the system of differential equations given by (7.19) instead of working with system (7.36).

Case $p = 1$

For simplicity and to gain some preliminary insight into the problem, we first study the case $p = 1$. The latent subpopulation disappears ($l_k = 0\ \forall k$) and using $u_k + t_k = 1$ we get:

$$\frac{du_k}{dt} = b - u_k(b + \lambda k\Theta - \mu_{tb}) - \mu_{tb}t_k^2, \tag{7.41}$$

where we have omitted temporal dependences, as we will do from now on. Looking for the stationary solution, we have that the condition $\frac{du_k}{dt} = 0$ implies:

$$u_k = -\left(\frac{1}{2\mu_{tb}}\right)\left(b + \lambda k\Theta - \mu_{tb} \pm \sqrt{(b + \lambda k\Theta - \mu_{tb})^2 + 4b\mu_{tb}}\right), \tag{7.42}$$

from which the meaningful solution is the one with the negative sign. The previous expression is consistent with the meaning of u^* since we recover the expected result $u^* = 1$ when $\Theta = 0$. Moreover, if we calculate the derivative with respect to Θ we get:

$$\frac{du_k^*(\Theta)}{d\Theta} = \frac{\lambda k}{2\mu_{tb}}\left(-1 + \frac{b + \lambda k\Theta - \mu_{tb}}{\sqrt{(b + \lambda k\Theta - \mu_{tb})^2 + 4b\mu_{tb}}}\right) < 0, \tag{7.43}$$

which guarantees that u^* will always be less than unity and therefore is a real, valid solution. The study of the value of Θ in the steady state help us to identify the epidemic threshold. We write:

$$\Theta^* = \frac{1}{\langle k\rangle}\sum_k kP(k)t_k^* = 1 - \frac{1}{\langle k\rangle}\sum_k kP(k)u_k^*, \tag{7.44}$$

which, after substituting u_k^* for its value, leads to:

$$\Theta^* = f(\Theta) = \frac{1}{2} - \frac{b}{2\mu_t b} + \frac{\lambda\langle k^2\rangle}{2\mu_{tb}\langle k\rangle}\Theta -$$
$$- \frac{1}{2\mu_{tb}\langle k\rangle}\sum_k kP(k)\sqrt{(b + \lambda k\Theta - \mu_{tb})^2 + 4b\mu_{tb}}. \tag{7.45}$$

The graphical interpretation of the above equation indicates that the existence of an equilibrium point in which $\Theta^* > 0$ is equivalent to the existence of a point at which $f(\Theta)$ cross the bisector of the first quadrant. Evaluating the second derivative of $f(\Theta)$ one gets:

$$\frac{d^2f(\Theta)}{d\Theta^2} = \frac{-2b\lambda^2}{\langle k\rangle}\sum_k \frac{P(k)k^3}{[(b + \lambda k\Theta - \mu_{tb})^2 + 4b\mu_{tb}]} < 0, \tag{7.46}$$

which ensures that the condition for the existence of such intersection is reduced to:

$$\left(\frac{df(\Theta)}{d\Theta}\right)_{\Theta=0} = \frac{\lambda\langle k^2\rangle}{\langle k\rangle}\frac{1}{b+\mu_{tb}} = 1, \tag{7.47}$$

from which the epidemic threshold is derived as:

$$\lambda_c = \frac{(b+\mu_{tb})\langle k\rangle}{\langle k^2\rangle}. \tag{7.48}$$

Note that apart from the factor $(b+\mu_{tb})$, the previous result, coincides formally with the SIR model.

Case $p \neq 1$

This is a somewhat more involved case. For structured populations, the resolution of the system of differential equations (7.36) cannot be done explicitly. The system; in its homogenous version conduces to a of a cubic equation, finding the roots of which is too complicated to be a useful strategy to cope with the problem. We next find the epidemic threshold for the case $p \neq 1$ using two approaches. On one hand, we study the time derivative of Θ. On the other hand, we will also make use of the singularity of the Jacobian at the point $(u_k, l_k, t_k) = (1, 0, 0)$ to argue that the expression for the critical threshold is given by:

$$(\lambda)_c = \frac{\langle k\rangle}{\langle k^2\rangle}\frac{(r+b)(\mu_{tb}+b)}{pb+r}. \tag{7.49}$$

Time evolution of Θ in populations with a low number of sicks. A first approach to characterize the epidemic threshold in heterogeneous networks when $p \neq 1$ is to study the sign of the derivative of Θ at the onset of an epidemic outbreak. We consider an initially healthy population in which a small proportion of infectious individuals are introduced so that $t_k << 1\ \forall k$. The derivative of Θ is:

$$\left(\frac{d\Theta}{dt}\right)_{\Theta\sim 0} = \frac{\sum_k P(k)k\frac{dt_k}{dt}}{\langle k\rangle} + \frac{\sum_k t_k k\frac{dP(k)}{dt}}{\langle k\rangle} - \left[\frac{\sum_k P(k)kt_k}{\langle k\rangle}\right]\left[\frac{\sum_k k\frac{dP(k)}{dt}}{\langle k\rangle}\right] \tag{7.50}$$

which, after substitution of the values of the derivatives of $P(k,t)$ and $t_k(t)$ leads to:

$$\left(\frac{d\Theta}{dt}\right)_{\Theta\sim 0} = \frac{\sum_k P(k)kl_k}{\langle k\rangle} + p\lambda\Theta\frac{\sum_k P(k)k^2u_k}{\langle k\rangle} - (b+\mu_{tb})\,\Theta + \mu_{tb}\Theta^2. \tag{7.51}$$

At this point we make two simplifications. The first and most easily justifiable is to neglect the term Θ^2. The second is related to the presence of l_k in the above equation, that we have to transform in a dependency with respect to t_k. Specifically, we assume to be sufficiently close to the stationary point $(u_k, l_k, t_k) = (1, 0, 0)$ as to be able to

assume that the three derivatives vanishes. In other words, and focusing our attention on latent and sick classes, we assume that:

$$\left(\frac{dl_k}{dt}\right)_{\Theta\sim 0} = (1-p)\lambda k\Theta u_k - (b+r)l_k + \mu_{tb} l_k t_k \simeq 0, \tag{7.52}$$

$$\left(\frac{dt_k}{dt}\right)_{\Theta\sim 0} = p\lambda k\Theta u_k + r l_k - (b+\mu_{tb})t_k + \mu_{tb} t_k^2 \simeq 0, \tag{7.53}$$

from which:

$$l_k = \frac{(1-p)(b+\mu_{tb})\, t_k - (1-p)\mu_{tb} t_k^2}{r + pb - \mu_{tb} p t_k}. \tag{7.54}$$

After expanding the previous equation and taking the leading term one arrives to:

$$l_k = \frac{(1-p)(b+\mu_{tb})}{r+pb} t_k, \tag{7.55}$$

which allows to express the derivative of Θ as:

$$\left(\frac{d\Theta}{dt}\right)_{\Theta\sim 0} = \Theta\left[\frac{r\,(1-p)(b+\mu_{tb})}{r+pb} - (b+\mu_{tb}) + p\lambda\frac{\sum_k P(k)k^2 u_k}{\langle k\rangle}\right]. \tag{7.56}$$

In the limit $u_k \simeq 1\ \forall k$ the third term within brackets is the ratio $\langle k^2\rangle/\langle k\rangle$, from which the epidemic threshold condition may be derived as:

$$\left(\frac{d\Theta}{dt}\right)_{\Theta\sim 0} = \Theta\left[\frac{r\,(1-p)(b+\mu_{tb})}{r+pb} - (b+\mu_{tb}) + p\lambda_c\frac{\langle k^2\rangle}{\langle k\rangle}\right] = 0, \tag{7.57}$$

finally leading to the expected expression for the threshold:

$$\lambda_c = \frac{\langle k\rangle}{\langle k^2\rangle}\frac{(r+b)(\mu_{tb}+b)}{pb+r}. \tag{7.58}$$

Analysis of the Jacobian While for well-mixed populations the condition of singularity of the Jacobian allows to get the epidemic threshold in a straightforward way, for heterogeneous populations the analysis of the Jacobian is a difficult task because Θ is a function of each and every one of the t_kś which translates into the need of calculating a determinant whose order is three times the number of connectivity classes. What we can reasonably do is to verify is the threshold condition is verified for systems in which there are two or three different connectivity classes.

Population with two degree classes (Jacobian (6x6)) In the first case in which only two different classes of connectivity exist, the Jacobian is just a quite distasteful 6x6 determinant that, after some cumbersome and lengthy algebra can be reduced to the expression:

$$J = b^2(b+r)(b+\mu_{tb})\left[(b+r)(b+\mu_{tb}) + \frac{\lambda\langle k^2\rangle}{\langle k\rangle}(pb+r)\right], \tag{7.59}$$

which equated to zero leads again to the previously obtained expression for the epidemic threshold:

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \frac{(r+b)(\mu_{tb}+b)}{pb+r}. \tag{7.60}$$

Population with three degree classes (Jacobian (9x9)) If we consider a population with three degree classes, the algebraic complexity of the problem largely increases as we now have a determinant to solve of 9x9 degree. However, we can proceed as before getting the following expression for the Jacobian:

$$J = b^3(b+r)^2(b+\mu_{tb})^2 \left[(b+r)(b+\mu_{tb}) + \frac{\lambda \langle k^2 \rangle}{\langle k \rangle}(pb+r) \right], \tag{7.61}$$

which only differs from the previous case (two degree classes) in the first three factors, obviously leading to the same value for the threshold.

7.4.4 Numerical simulations

When designing numerical simulations to simulate the dynamics of the system under study, we have two difficulties not previously addressed in the literature. These numerical issues with which we have to deal are derived from the fact that we have a system that is simultaneously open and structured. As a result of dealing with an open system, new individuals are being added to the population at a rate given by the birth rate. Additionally, these new individuals must enter the network of contacts with a predefined connectivity. While deciding how many nodes our new individuals connect to is not a problem, it certainly is to decide what are those nodes the newcomers will be linked to, as this will impact the degree distribution in a nontrivial way. This is an unavoidable numerical complication that we should face relentlessly if the analytical calculations are to be compared with Monte Carlo simulations.

To this end, we have adapted a simulation method based on transition probabilities first proposed in [59] for SIR models in complex networks. The numerical approach consider all transitions between states that take place during the dynamical evolution of the subpopulations, defined by the system of differential equations (7.36). When dealing with structured populations, these transition rates depend, in general, on the connectivity class within which they occur. Moreover, within each k-class seven transitions are possible:

type 1 Birth of healthy individuals,

type 2 Natural death of healthy individuals,

type 3 Natural death of latently infected individuals,

type 4 Natural or disease-related death of sick individuals,

type 5 Transition from a healthy to the latent state,

$$\omega_{1,k} = bNP(k, t_o)$$
$$\omega_{2,k} = \mu NP(k, t)u_k$$
$$\omega_{3,k} = \mu NP(k, t)l_k$$
$$\omega_{4,k} = (\mu + \mu_{tb})NP(k, t)t_k$$
$$\omega_{5,k} = (1 - p)\lambda kNP(k, t)u_k\Theta$$
$$\omega_{6,k} = p\lambda kNP(k, t)u_k\Theta$$
$$\omega_{7,k} = rNP(k, t)l_k$$

TABLE 7.2: Transition frequencies

type 6 Transition from a healthy to the sick (infectious) state,

type 7 Transition from a latent to the sick (infectious) state.

Each of these transitions is characterized by a characteristic transition rate $\omega_{i,k}$ that can be directly derived from the system of equations that characterizes the rate at which they occur i ($i = 1, 2...7$) within the class k as:

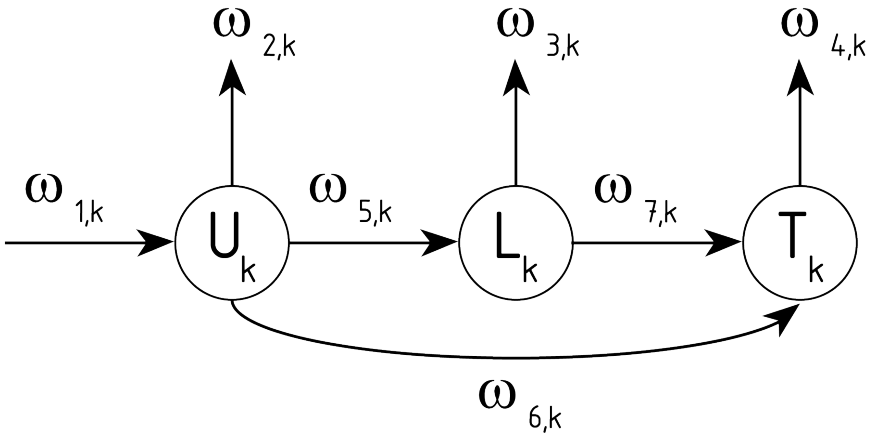


FIGURE 7.3: Set of allowed transitions in the epidemic model within each connectivity class

Similarly, we define the sum of all these transition rates as the average rate at which *one* transition (of any kind) occurs:

$$\Omega = \sum_{i,k} \omega_{i,k}. \tag{7.62}$$

This average transition rate in its turn defines the characteristic or average time τ elapsed between any two consecutive transitions, the latter being defined as the inverse of Ω:

$$\tau = \frac{1}{\Omega}. \tag{7.63}$$

$$\begin{gathered}\mu = 0.009 \quad \text{years}^{-1}, \\ b = 0.010 \quad \text{years}^{-1}, \\ \mu_{Tb} = 0.200 \quad \text{years}^{-1}, \\ r = 0.002 \quad \text{years}^{-1}, \\ p = 0.070,\end{gathered}$$

TABLE 7.3: Parameter values

Given the previous definitions, the Monte Carlo algorithm is implemented in such a way that at each MC step (of duration τ) one single transition takes place. Finally, the probability $\Pi_{i,k}$ that a given transition actually happens, is calculated as:

$$\Pi_{i,k} = \frac{\omega_{i,k}}{\Omega} = \tau\omega_{i,k}, \tag{7.64}$$

which determines what of all possible transitions is realized at each time step τ.

We have made extensive numerical simulations of the model starting from an initial population made up of $N \simeq 10^6$ individuals, whose network of contacts follows an initial degree distribution $P(k) \simeq k^{-3}$. Moreover, every newborn individual join the system with a degree that verifies the same connectivity distribution. As for the values of the parameters of the dynamics, and thinking of typical values for persistent diseases; we have set the following values:

Demographical parameters b and μ are roughly those of a country like Spain, while the parameters p, r are in the range of typical values for the case of TB. μ_{tb} has been chosen attending to numerical convenience (TB reaches a disease related mortality rate that reaches 0.8). In the other hand, we note that we must give a numerical criterion to define the stationary. In our simulations we first let the system evolve for 4500 years and later take averages in a window of 10000 Montecarlo steps (which corresponds, roughly, with a temporal lapse of 100 years), for the mean densities of healthy, latent and sick individuals defined in (7.36). This is a long enough time and ensures all of the outputs to the left of the threshold are stabilized to the state $(1, 0, 0)$ (this is achieved almost surely already for $t \leq 4000y$).

With the values for the parameters as specified above, the epidemic threshold is $\lambda_c = 0.305$. In the figure below, we have plotted the stationary proportion of sick individuals for values of the probability of transmission λ belonging to the interval $[0, 0.5]$:

In this case, we can observe that the difference between the predicted value for λ_c by the model and the value registered by the simulations reached not more than 15% of the analytical value. Being, however, the method not enough accurate to evaluate this error properly, we have been dealing with a variation of the algorithm focussed on improving the accuracy of λ_c register. In this variation we essentialy focus our attention in the surroundings of the critical value, instead of sweaping all the values for λ.

More specifically, we start by calculating the analytical value for λ_c and we start the simulation there. At each realization, we expect 6000 years at an eventual arrival of the

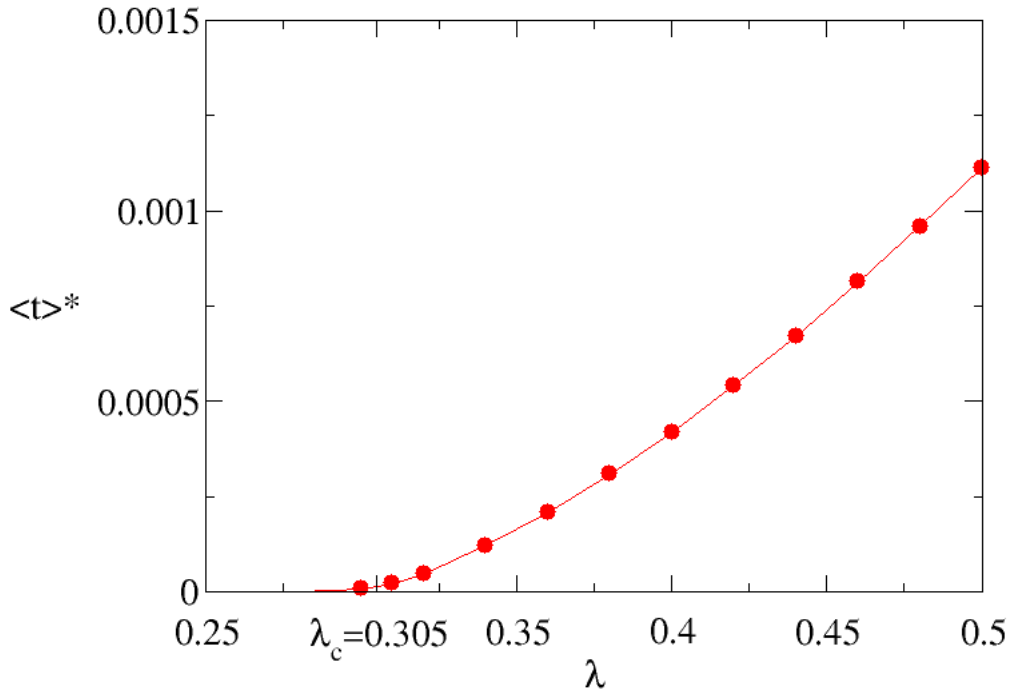


FIGURE 7.4: Stationary proportion of sick individuals for $\lambda \in [0, 0.5]$; with $p = 0.07$, $r = 0.002$, $\mu_{tb} = 0.2$, $\mu = 0.009$ and $b = 0.010$ ($\lambda_c = 0.305$). We can see the critical point at $\lambda \simeq \lambda_c$. Error bars are smaller than symbol size.

system to the state $(\langle u\rangle, \langle l\rangle, \langle t\rangle) = (1, 0, 0)$. In case of this state would not be reached in the 6000 years simulation, we consider that we are to the right of the critical point and so, we move to the left in λ just a little quantity $\delta\lambda = 0.01$. If, in the other case, we register the entire set of 10 realizations stabilized at state $(\langle u\rangle, \langle l\rangle, \langle t\rangle) = (1, 0, 0)$, we assume to be to the left of the critical point and, in consequence, we perform a $\delta\lambda$ switch to the right. Each time that the algorithm change direction, we divide by two the value of $\delta\lambda$ until we reach ten reversals; so as to converge to an enough accurate numerical λ_c. In order to compute the error for different initial population sizes, we have performed simulations for $N_o = 10^3, 10^4, 2 \cdot 10^4$ and 10^5; results are shown in table 7.4:

As we could expect of a finite size-effect, the larger is the size of the population, the smaller is the absolute error between numerical and analytical thresholds.

N_o	Analytical λ_{cA}	Numerical λ_{cN}	Absolute error $\lvert\lambda_{cA} - \lambda_{cN}\rvert$	Percentile relative error
1000	0.52903	0.46968	0.05935	11.2
10000	0.42559	0,37595	0.04964	11.7
50000	0,37620	0,33088	0,04532	12.0
100000	0,35854	0,31926	0,03929	11.0

TABLE 7.4: Numeric and analytic characterization of epidemic thresholds for different network sizes

7.5 Conclussions

We have discussed a model for the spreading of persistent infections in homogeneous and heterogeneous populations. This kind of approach is particularly suited for diseases like TB, where one can have at the same time new ingredients such as long latency periods and primoinfections.

As a general conclusion of this chapter, we have addressed the way in which the different dynamical features of TB affect its epidemic thresholds, and particularly, we have derived the kind of influence that the topology of the network exerts on the overall behavior of the epidemic spreading. In addition, we have outlined a rationale way to address the influence of simultaneously taking into account open and heterogeneous populations, which gives us insight for the development of more sophisticated models.

Connectivity fluctuations of the network play a major role by strongly enhancing the infection's incidence. This issue assumes a particular relevance in the case of scale-free (SF) networks that exhibit connectivity fluctuations diverging with the increasing size N of the network. SF networks are therefore very weak in face of infections presenting an effective epidemic threshold that is vanishing in the limit $N \to \infty$. In the case of an infinite population this corresponds to the absence of any epidemic threshold below which major epidemic outbreaks are impossible. These results strengthen the epidemiological framework for complex networks reported for the susceptible-infected-susceptible (SIS) model [8] and proposed as well for the SIR model [57].

Although for airborne transmitted diseases, contact networks' heterogeneity doesn't seem to reach the extreme levels observed for the networks of sexual contacts relevant for sexually-transmitted diseases [11] (i.e. scale-freedom [8]), the emerging picture after these studies stimulate the re-analysis of several concepts of standard epidemiology such as the "core group" or the characteristic number of contacts that appears to be ill-defined in SF networks. The high heterogeneity of SF networks finds signatures also in the different relative incidence within populations of varying connectivity k.

Chapter 8
Dynamics of interacting epidemics

8.1 Motivation and modeling framework

As it has been pointed in the introduction to this thesis, in the way towards the achievement of epidemic modeling platforms with quantitative description and forecasting abilities, an emergent research line in the field is the development of models for the description of the concurrent spreading of more than one disease over the same population [64, 65, 66, 67, 68]. These theoretical advances obeys to the need of improving our understanding of many relevant real examples of couples (or groups) of diseases that simultaneously transmit over human populations being able to modify each others' transmission patterns and burden levels. Some relevant examples of these systems of interacting epidemics are formed by the different influenza strains that compete each winter over the same host populations [71] and the syndemic system formed by the human immunodeficiency virus (HIV) and certain opportunistic pathogens that take advantage of the immune depression of HIV infected individuals to attack them [69, 70]. As a relevant case of these HIV-favoured infections, in the regions of the world in which HIV reaches highest prevalence levels –mainly sub-saharan Africa–, the radical increase of TB burden levels in the last two decades appears as a deadly side-effect of HIV irruption that constitutes a major concern for national and international public health authorities.

In this chapter, we propose a modeling framework, based on an HMF approximation, for the spreading of two concurrent diseases that interact with each other. By this way, we aim to explore epidemiological scenarios in which the dynamical parameters (i.e. infectiousnesses and recovery rates) of the diseases modeled depend, at the level of single individuals, on whether the agents involved are infected (or recovered from) only one of the diseases or both. In addition, we consider that neither the mechanisms behind each disease evolution nor the contact networks through which they spread have to be (in general) the same, and thus, these are considered independently for each disease. Let us then consider that we have two diseases (disease 1 and disease 2) which spread over two different networks of contacts: disease 1 propagates over network 1, which has a mean degree $\langle k \rangle = \sum_{k,l} P(k,l)k$; whilst disease 2 does so over network 2, whose mean connectivity is equal to $\langle l \rangle = \sum_{k,l} P(k,l)l$. The composed degree distribution $P(k,l)$ gives the proportion of nodes (individuals) having k and l links in networks 1 and 2, respectively [409].

Taking that in mind, we will analyze in detail the dynamics of such a two-layer system that represents the networks of contacts on top of which either two coupled SIS or SIR processes spread. Although the topological description we use has attracted a lot of attention lately, i.e., the so-called multilayer networks [230, 408], the main focus is

to study what are the effects of the complex interplay between the different interaction mechanisms considered and the temporal and topological scales onto the dynamical behavior of the diseases. Specifically, we explicitly derive the epidemic threshold of each disease in terms of the parameters that characterize their evolution and the topology of the networks of contacts. Remarkably, we find that the epidemic thresholds in the SIR case are different from those of the SIS case, which is a consequence of the emergence of new mechanisms of interaction and to the transitory nature of the epidemic outbreaks in the SIR case. Additionally, different scenarios and limiting cases of both theoretical and epidemiological interest are scrutinized. Finally, results from numerical simulations are presented to validate our analytical results. Our work thus shows how the different interaction mechanisms considered can give rise to new phenomenological insights regarding the dynamics of interacting diseases.

8.2 The SIS scenario

As a first step, we will consider the baseline scenario in which the isolated dynamics of each disease, when the second is absent, is described by a simple S-I-S scheme. Each individual belonging to a composed connectivity class (k, l) can be in four different dynamical states: susceptible with respect to both diseases $SS(k, l)$, infected of both $II(k, l)$, and infected with the first (second) one and still susceptible to catch the second (first) disease, $IS(k, l)$ $(SI(k, l))$, being these quantities the proportion of individuals at each disease state with composed degree class (k, l). Thus, we have that $SS(k, l) + IS(k, l) + SI(k, l) + II(k, l) = 1\ \forall(k, l)$. In addition, regardless of the connectivities of the nodes involved, we have eight possible contagion transitions after a contact (four for each disease).

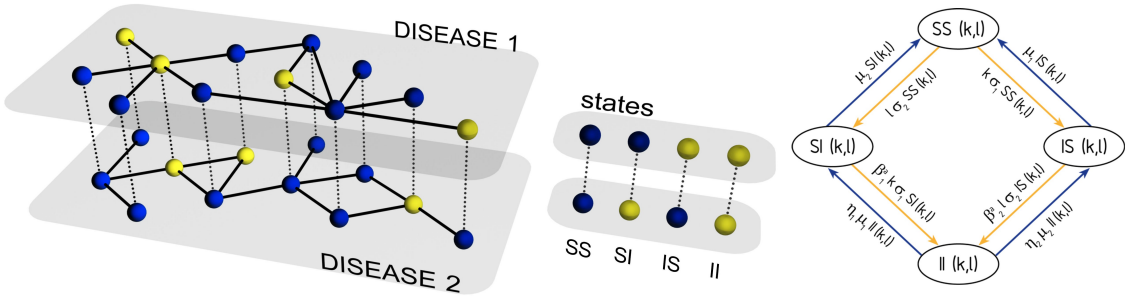


FIGURE 8.1: SIS-SIS interacting diseases model (a) Each of the diseases spreads over a different network. Each individual belongs simultaneously to each network and can be (or not) infected with each of the diseases. (b) Set of transitions allowed in the model. The variables $(SS(k, l), IS(k, l), SI(k, l), II(k, l))$ represent the densities of individuals of each type in the system having k neighbors in network 1 and l neighbors in network 2.

On the other hand, four other possible transitions correspond to the cases in which infected individuals go back to the susceptible class. This amounts to a total of twelve

Parameter	Dynamical meaning
λ_1	Baseline infectiousness of disease 1
λ_2	Baseline infectiousness of disease 2
μ_1	Baseline recovery rate of disease 1
μ_2	Baseline recovery rate of disease 2
β_1^a	Variation of disease 1 infectiousness due to the fact that the susceptible individual exposed to disease 1 is infected with disease 2
β_2^a	Variation of disease 2 infectiousness due to the fact that the susceptible individual exposed to disease 2 is infected with disease 1
β_1^b	Variation of disease 1 infectiousness due to the fact that the spreader is also infected with disease 2
β_2^b	Variation of disease 2 infectiousness due to the fact that the spreader is also infected with disease 1
η_1	Variation of disease 1 recovery rate for individuals also infected with disease 2
η_2	Variation of disease 2 recovery rate for individuals also infected with disease 1

TABLE 8.1: Definition of model parameters

elementary transitions, schematized as follows (see also Figure. 8.1):

$$
\begin{aligned}
IS + SS &\xrightarrow{\lambda_1} IS + IS \\
IS + SI &\xrightarrow{\beta_1^a \lambda_1} IS + II \\
II + SS &\xrightarrow{\beta_1^b \lambda_1} II + IS \\
II + SI &\xrightarrow{\beta_1^a \beta_1^b \lambda_1} II + II \\
SI + SS &\xrightarrow{\lambda_2} SI + SI \\
SI + IS &\xrightarrow{\beta_2^a \lambda_2} SI + II \\
II + SS &\xrightarrow{\beta_2^b \lambda_2} II + SI \\
II + IS &\xrightarrow{\beta_2^a \beta_2^b \lambda_2} II + II \\
IS &\xrightarrow{\mu_1} SS \\
SI &\xrightarrow{\mu_2} SS \\
II &\xrightarrow{\eta_1 \mu_1} SI \\
II &\xrightarrow{\eta_2 \mu_2} IS
\end{aligned}
$$

The model thus contains two basic infection probabilities –λ_1 and λ_2– as well as two basic recovery rates –μ_1 and μ_2–, one for each disease. In addition, infection

probabilities are affected by scaling factors –the four β's and combinations of them– and so are the recovery rates –by the η's–, as it is explained in detail in Table 8.4. These parameters describe the interaction of the diseases through three different effects that are concurrently taken into account. The first effect is the variation of the susceptibility of healthy individuals to get infected with one disease as a consequence of being infected with another. This mechanism is described by β_1^a for the variation of the infection risk of disease 1 caused by disease 2, and β_2^a, for the symmetric case. The second effect is the change of the spreading capabilities of double-infected individuals with respect to single-infected ones, which is described by the parameters β_1^b and β_2^b, for diseases 1 and 2, respectively. The last effect is the variation of the infectious periods of double-infected individuals also with respect to single-infected ones, described by η_1 and η_2.

Therefore, these parameters exhaustively describe all the ways in which two diseases can interact according to a SIS scheme, and allow us to isolate the different effects of one disease on the spreading of the other by making infectious individuals more efficient spreaders or by making susceptible individuals more prone to get sick. Once we have defined the whole set of parameters in table 8.4, we have all possible transitions between dynamical states well defined (see Figure 8.1).

According to the scheme depicted in the figure (in which simultaneous double contagions and recoveries from both diseases have been explicitly excluded), we can consider that all individuals within the same composed connectivity class (k, l) are dynamically equivalent, in order to get a composed HMF for the dynamical description of both diseases which neglects, as a first approximation, further correlations. In this way, the set of differential equations describing the evolution in time of the four densities of individuals $(SS(k,l), IS(k,l), SI(k,l), II(k,l))$ is as follows:

$$\begin{aligned}
\dot{SS}(k,l) &= -k\lambda_1\theta_1^{IS}SS(k,l) - l\lambda_2\theta_2^{SI}SS(k,l) - \\
&-k\beta_1^b\lambda_1\theta_1^{II}SS(k,l) - l\beta_2^b\lambda_2\theta_2^{II}SS(k,l) + \mu_1 IS(k,l) + \mu_2 SI(k,l) \\
\dot{IS}(k,l) &= k\lambda_1\theta_1^{IS}SS(k,l) + k\beta_1^b\lambda_1\theta_1^{II}SS(k,l) - \\
&-l\beta_2^a\lambda_2\theta_2^{SI}IS(k,l) - l\beta_2^a\beta_2^b\lambda_2\theta_2^{II}IS(k,l) - \mu_1 IS(k,l) + \eta_2\mu_2 II(k,l) \\
\dot{SI}(k,l) &= l\lambda_2\theta_2^{SI}SS(k,l) + l\beta_2^b\lambda_2\theta_2^{II}SS(k,l) - \\
&-k\beta_1^a\lambda_1\theta_1^{IS}SI(k,l) - k\beta_1^a\beta_1^b\lambda_1\theta_1^{II}SI(k,l) - \mu_2 SI(k,l) + \eta_1\mu_1 II(k,l) \\
\dot{II}(k,l) &= k\beta_1^a\lambda_1\theta_1^{IS}SI(k,l) + l\beta_2^a\lambda_2\theta_2^{SI}IS(k,l) + \\
&+k\beta_1^a\beta_1^b\lambda_1\theta_1^{II}SI(k,l) + l\beta_2^a\beta_2^b\lambda_2\theta_2^{II}IS(k,l) - (\eta_1\mu_1 + \eta_2\mu_2)II(k,l)
\end{aligned} \tag{8.1}$$

where the θ parameters are defined in Table 8.2.

Combining the probabilities θs and defining the following two new parameters $\sigma_1 = \lambda_1(\theta_1^{IS} + \beta_1^b\theta_1^{II})$ and $\sigma_2 = \lambda_2(\theta_2^{SI} + \beta_2^b\theta_2^{II})$ we obtain the average probabilities per link for SS-nodes to become infected with disease 1 (σ_1) or disease 2 (σ_2). This allows us

$\theta_1^{IS} = \frac{\sum_{k,l} P(k,l)kIS(k,l)}{\sum_{k,l} P(k,l)k}$	probability that a given link of network 1 points to an IS node
$\theta_1^{II} = \frac{\sum_{k,l} P(k,l)kII(k,l)}{\sum_{k,l} P(k,l)k}$	probability that a given link of network 1 points to an II node
$\theta_2^{SI} = \frac{\sum_{k,l} P(k,l)lSI(k,l)}{\sum_{k,l} P(k,l)l}$	probability that a given link of network 2 points to a SI node
$\theta_2^{II} = \frac{\sum_{k,l} P(k,l)lII(k,l)}{\sum_{k,l} P(k,l)l}$	probability that a given link of network 2 points to an II node

TABLE 8.2: Definition of parameters θ.

to rewrite the system of differential equations as:

$$\dot{SS}(k,l) = -(k\sigma_1 + l\sigma_2)SS(k,l) + \mu_1 IS(k,l) + \mu_2 SI(k,l) \tag{8.2}$$

$$\dot{IS}(k,l) = k\sigma_1 SS(k,l) - l\beta_2^a \sigma_2 IS(k,l) - \mu_1 IS(k,l) + \eta_2\mu_2 II(k,l) \tag{8.3}$$

$$\dot{SI}(k,l) = l\sigma_2 SS(k,l) - k\beta_1^a \sigma_1 SI(k,l) - \mu_2 SI(k,l) + \eta_1\mu_1 II(k,l) \tag{8.4}$$

$$\dot{II}(k,l) = k\beta_1^a \sigma_1 SI(k,l) + l\beta_2^a \sigma_2 IS(k,l) - (\eta_1\mu_1 + \eta_2\mu_2)II(k,l) \tag{8.5}$$

Due to the closure relationship $SS(k,l) + IS(k,l) + SI(k,l) + II(k,l) = 1\ \forall(k,l)$, only three of these four equations are linearly independent for each composed connectivity class (k,l). Taking this into account, we next analyze the time evolution of the vector $(IS(k,l), SI(k,l), II(k,l))$, since $SS(k,l) = 1 - IS(k,l) - SI(k,l) - II(k,l)$.

8.2.1 Epidemic thresholds

In order to analyze the most relevant dynamical properties of the system, we look for the values $(IS^*(k,l), SI^*(k,l), II^*(k,l))$ that define the stationary state:

$$(\dot{IS}^*(k,l), \dot{SI}^*(k,l), \dot{II}^*(k,l)) = (0,0,0)\forall(k,l) \tag{8.6}$$

for the system Eq. (8.5). In order to do that, we are forced to consider σ_1 and σ_2 as additional parameters, although these two quantities are linear combinations of the other variables and the ultimate responsible of the coupling among all connectivity classes. As it happens commonly in this kind of models [400, 395], there exists a trivial fixed point of the dynamics in which there are no infected individuals in the system: $(II_1^*(k,l), II_2^*(k,l), II^*(k,l)) = (0,0,0)\ \forall(k,l)$. This fixed point represents the absorbing state of our model. In addition, there are other possible fixed points for which the densities of infected individuals could be written as a function of the σ parameters: $(II_1^*(k,l,\sigma_1,\sigma_2), II_2^*(k,l,\sigma_1,\sigma_2), II^*(k,l,\sigma_1,\sigma_2)) \neq (0,0,0)$.

It is thus possible to get self-consistent equations for the variables σ as:

$$\sigma_1 = f_1(\sigma_1,\sigma_2) = \frac{\lambda_1}{\langle k\rangle}\sum_{k,l} P(k,l)k(IS(k,l,\sigma_1,\sigma_2) + \beta_1^b II(k,l,\sigma_1,\sigma_2)) \tag{8.7}$$

$$\sigma_2 = f_2(\sigma_1,\sigma_2) = \frac{\lambda_2}{\langle l\rangle}\sum_{k,l} P(k,l)l(SI(k,l,\sigma_1,\sigma_2) + \beta_2^b II(k,l,\sigma_1,\sigma_2)) \tag{8.8}$$

The condition $\sigma_1 = f_1(\sigma_1, \sigma_2) > 0$ implies the existence of a stable active state for the dynamics of disease 1, i.e., a state in which disease 1 becomes endemic in the population. For this situation to take place for disease 2, the condition $\sigma_2 = f_2(\sigma_1, \sigma_2) > 0$ must be fulfilled. Given the symmetry between the two expressions in Eq. (8.8), we will only focus on the analysis of the first equation. In fact, it can be shown that always $\frac{\partial^2 f(\sigma_1,\sigma_2)}{\partial \sigma_1^2} < 0$, which means that, if we think of the graphical solution of the equation $\sigma_1 = f_1(\sigma_1, \sigma_2)$, for this non trivial solution to exist –given that, obviously $f(\sigma_1 = 0, \sigma_2 = 0) = 0$– it must be verified that $\left[\frac{\partial f(\sigma_1,\sigma_2)}{\partial \sigma_1}\right]_{\sigma_1=0} > 1$, as in [400]. After some algebraic operations, this condition yields the following expression:

$$\frac{\lambda_1 \sum_{k,l} P(k,l) k^2 \frac{l^2\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + l\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)}{l^2\sigma_2^2\beta_2^a\eta_1 + l\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)}}{\mu_1\langle k\rangle} > 1 \tag{8.9}$$

that allows us to derive the epidemic threshold as:

$$\lambda_1^c(\sigma_2) = \mu_1 \frac{\langle k\rangle}{\sum_{k,l} P(k,l) k^2 \frac{l^2\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + l\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)}{l^2\sigma_2^2\beta_2^a\eta_1 + l\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)}} \tag{8.10}$$

Looking at the latter expression −which contains the underlying topologies in a more intricate way than for the uncoupled, classical case−, the threshold dependence on disease 2's prevalence via σ_2 becomes explicit. If we evaluate $\lambda_1^c(\sigma_2 = 0)$ we recover the classical result $\lambda_1^c = \mu_1\langle k\rangle/\langle k^2\rangle$ [8, 400]. Therefore, in the following we will refer to this baseline case as primary threshold, $\lambda_1^c(0)$, whereas the more general case will be referred to as the secondary threshold, $\lambda_1^c(\sigma_2)$ (with $\sigma_2 > 0$). Obviously, the same stands for the primary ($\lambda_2^c(0)$) and secondary thresholds ($\lambda_2^c(\sigma_1)$) of the second disease.

A particular case for the topologies on top of which both diseases are spreading corresponds to the homogeneous MF version of the system, i.e., $P(k,l) = \delta(k-k_o)\delta(l-l_o)$, for which the last expression can be rewritten as:

$$\lambda_1^c = \frac{\lambda_2 l_o}{k_o} \frac{\eta_1\mu_1(\beta_2^a(\lambda_2 l_o - \mu_2) + \mu_1) + \eta_2\mu_2\mu_1}{\mu_2(\eta_2\mu_2 + \eta_1\mu_1) + (\lambda_2 l_o - \mu_2)(\beta_2^a\beta_1^a\beta_1^b(\lambda_2 l_o - \mu_2) + \beta_1^a\beta_1^b\mu_1 + \eta_2\mu_2\beta_1^a + \beta_2^a\beta_1^b\mu_2)} \tag{8.11}$$

An independent derivation of this expression can be obtained by analyzing the Jacobian matrix of the homogeneous MF system analogous to Eq. (8.5) as it is shown in the Appendix A.

8.2.2 Numerical simulations

In order to explore the quality of the threshold prediction of our model, we have designed a Monte-Carlo simulation scheme in which a single state transition is allowed per individual per time step. First, infected individuals will eventually spread the disease(s) that they carry. As double events are not allowed at a single time step, forbidden double transitions from SS to II are resolved by choosing the disease that an individual will catch proportionally to the status of her infected neighbors: the more neighbors one individual has, say, carrying disease 1, the more likely she will become

infected with disease 1 rather than with disease 2. After the spreading loop is completed for both diseases, infected nodes who have not suffered any contagion at the present time step eventually get back to the susceptible state of the disease(s) they carry. To avoid forbidden double transitions from the II class to the SS class, in those cases the only disease the individual is going to recover from is also chosen stochastically, according to the probabilities $p_1 = \eta_1\mu_1/(\eta_1\mu_1 + \eta_2\mu_2)$ for the first disease, and $p_2 = 1 - p_1$ for the second one.

Let us then explore first a simple scenario in which we assume that the dynamical effects of one disease on the other are totally symmetric. In terms of the parameters of the model, this implies that $\beta_1^a = \beta_2^a = \beta_1^b = \beta_2^b \equiv \beta$ and $\eta_1 = \eta_2 \equiv \eta$. In this case, let's focus on two opposite scenarios: i) mutual enhancement: $\eta < 1$ and $\beta > 1$ and ii) mutual impairment $\eta > 1$ and $\beta < 1$. In the case of mutual enhancement, individuals who are infected with the second disease spread and become infected with disease 1 more easily than those who are not (this is because $\beta > 1$). In addition, since also $\eta < 1$, infected individuals of disease 1 remain so for longer times if they are also infected with the second disease. These two effects imply that the appearance of disease 2 in the system enhances the spreading capabilities of disease 1. The reciprocal situation is also true, as the interaction between both diseases is symmetric. Finally, in the case of mutual impairment, the effects on the infectiousness and recovery rates are the opposite, and so the appearance of one of the diseases at a certain prevalence impairs the spreading of the other disease. In Figures 8.2–8.3, we represent the prevalence of each disease, as a function of the baseline infectiousnesses (λ_1, λ_2) for a given set of parameters after the introduction of infection seeds in the order shown. The networks through which diseases spread are, for this first case, two uncorrelated Erdös-Renyi (ER) graphs.

For the case of mutual enhancement (Fig. 8.2), given our set of parameters, the analytically-obtained curves for the secondary threshold remain below the primary thresholds, leading to the appearance of two regions in the plane (λ_1, λ_2), for which it is verified that $\lambda_1^c(\sigma_2) < \lambda_1 < \lambda_1^c(0)$ and $\lambda_2^c(\sigma_1) < \lambda_2 < \lambda_2^c(0)$, respectively. The dynamical relevance of these regions is that within them, the appearance of an outbreak of one of the diseases is conditional to the previous installation of the other infection in the system. In this way we can observe that, after an initial seed of IS individuals, disease 1 does not become endemic in the region $\lambda_1^c(\sigma_2) < \lambda_1 < \lambda_1^c(0)$ (fig. 8.2, panel a), but then, after the outbreak of the second disease in the network, the same seed leads disease 1 to become endemic in that same region (fig. 8.2 panel C) as predicted by our model. Regarding the conjugate region in which $\lambda_2^c(\sigma_1) < \lambda_2 < \lambda_2^c(0)$, we can see in figure 8.2, panel e, that disease 2 directly becomes endemic after the introduction of an infection seed SI due to the fact that, previously, disease 1 was already introduced in the system.

The situation for the mutual impairment case is the opposite, and the secondary thresholds remain, in this case, above the primary ones. So, in this scenario, we have another couple of relevant regions in which $\lambda_1^c(0) < \lambda_1 < \lambda_1^c(\sigma_2)$ and $\lambda_2^c(0) < \lambda_2 < \lambda_2^c(\sigma_1)$, respectively. In fig. 8.3, it is represented the behavior of the system under these conditions. In panel a, we can see how disease 1 becomes endemic after the

introduction of an initial seed above its primary threshold. Then, after introducing a seed of disease 2, as shown in panel b for the area comprised between $\lambda_1^c(0)$ and $\lambda_1^c(\sigma_2)$, the prevalence of disease 1 vanishes. In other words, in that region, the introduction of disease 2 makes it possible for the system to recover from disease 1. If we look at the behavior of the second disease in the region in which $\lambda_2^c(0) < \lambda_2 < \lambda_2^c(\sigma_1)$, we see how the disease is unable to become endemic as a consequence of the fact that the first disease has already been introduced in the system. This situation suggests that, as it has already been addressed in the context of computational sciences [410], the introduction of an infectious agent designed to immunize its host with respect to another, more harmful infection, might be a conceptually feasible option to reduce the prevalence of the latter, or even to eradicate it. This has also been recently reported in the context of multi-strain diseases, in which more than one strain of the same disease compete for the host population [64, 74].

Once the dynamics of the model has been exhaustively characterized when the diseases spread over homogeneous networks, we move on and explore the influence of degree heterogeneity on the dynamics. To this end we also perform intensive numerical simulations in an analogous way, but using SF graphs of the same size as before with different exponents. In Figure 8.4 we represent the final prevalence for each disease in two configurations: reciprocally enhanced diseases −panels a and c− and impaired spreading −panels b and d−. In both cases, network 1 ($\gamma = 2.7$, panels *a* and *b*) has a greater power law exponent than network 2 ($\gamma = 2.5$, panels *c* and *d*). As it can be seen, for both diseases and for both configurations, secondary thresholds are closer to primary ones than in the case of homogeneous networks.

8.2.3 System sizes and epidemic thresholds: general case

In the previous sections we have described the baseline cases in which none of the dynamical parameters vanishes, and both networks are, in each case, of the same kind –ER or uncorrelated SF graphs–. A relevant theoretical question yet remains unanswered, i.e., how the epidemic thresholds behave when the system size grows and eventually reaches the thermodynamic limit. Regarding this question, we present an exhaustive analysis in section 8.7.2, in which it is shown what are the conditions that lead to have vanishing small secondary thresholds as a function of the underlying topologies and some of the dynamical parameters − note, however, that this question is of interest from a theoretical viewpoint, as strictly speaking all real systems are finite and thus an effective epidemic threshold exits. In addition, our approach is based on an HMF approximation, and that means that influences of dynamical correlations on the analysis are neglected.

The results show that the secondary epidemic threshold associated to any of the diseases that is spreading over a SF network with $2 < \gamma \leq 3$ vanishes at the thermodynamic limit, regardless of the topology of the network on top of which its conjugate disease propagates and of the values of the dynamical parameters. In an analogously robust way, power laws with $\gamma > 3$ −or homogeneous degree distributions− yield finite, non vanishing secondary thresholds at the thermodynamic limit, regardless of the

conjugate topologies or parameter values, with some exceptions.

The relevance of this result relies on the fact that the model predicts the same behavior for the epidemic thresholds in heterogeneous and homogeneous networks as compared with HMF models of uncoupled (single) diseases that spread over simple networks. In addition, our analysis shows that the eventual vanishing of the epidemic thresholds for infinite systems is only determined by the topology of the network under consideration rather than by any possible coupling with another disease that spreads over any other possible conjugate network within our model framework. However, there is an exception to this general behavior which is meaningful from an epidemiological viewpoint. This is the case when both diseases spread over two highly and positively correlated SF networks with composed degree distribution $P(k,l) = \delta(k-l)\alpha k^{-\gamma}$, where δ stands for the Kronecker δ-function and α is a normalization constant. In that situation, if we focus, for example, in disease 1, there exist two different interaction schemes of interest for which we recover finite epidemic thresholds $\lambda_1^c > 0$ at the thermodynamic limit even for SF graphs with $2 < \gamma \leq 3$:

- Case 1: Individuals infected with disease 2 become immune to infection by disease 1. $\beta_1^a = 0$; $(\beta_2^a, \beta_2^b, \beta_2^b)$ are free parameters.
- Case 2: If $\beta_1^a \neq 0$, individuals infected with both diseases can not cause contagion of disease 1: $\beta_1^b = 0$. In addition, disease 1 can not cause total immunity to disease 2: $\beta_2^a \neq 0$. β_2^b is a free parameter.

To illustrate this situation, we take as an example a particular case of the first scheme, a mutual cross-immunity scenario given by $\beta_1^a = \beta_2^a = 0$. In order to point out the role of inter-layer degree correlations on this effect, we can directly compare the expression for the threshold when both networks are totally correlated with the analogous expression derived from an uncorrelated combined degree distribution:

$$P(k,l) = \delta(k-l)\alpha k^{-\gamma} \rightarrow \lambda_1^c(\sigma_2) = \mu_1 \frac{\langle k \rangle}{\sum_k \alpha \frac{1}{k^{2-\gamma}} \frac{\mu_2}{k\sigma_2 + \mu_2}} \tag{8.12}$$

$$P(k,l) = \alpha k^{-\gamma} l^{-\Gamma} \rightarrow \lambda_1^c(\sigma_2) = \mu_1 \frac{\langle k \rangle}{\sum_{k,l} \alpha \frac{1}{k^{2-\gamma}} \frac{1}{l^{\Gamma}} \frac{\mu_2}{l\sigma_2 + \mu_2}} \tag{8.13}$$

In Figure 8.5, we represent the values predicted by these expressions for different network sizes, when $\gamma = \Gamma = 2.5$. As we can see in panel a, for uncorrelated networks, regardless of the value of σ_2, the threshold continuously decreases as we increase network sizes. The result thus shows that the existence of a coupling with another disease present in the system with a certain prevalence proportional to σ_2, does not play any role, since the degree heterogeneities are still the main reason leading to the vanishing of the threshold at the thermodynamic limit. This picture turns out to be remarkably different when we introduce positive, strong correlations between the two networks. In that case, as we can see in panel b, the appearance of the second disease, characterized by a certain prevalence level $\sigma_2 > 0$, implies a sudden change in the behavior of the threshold, that does not vanish anymore, even when $N \rightarrow \infty$.

The influence of degree correlations between networks for this case of full cross-immunity becomes evident also at finite sizes, since the differences between primary and secondary thresholds, as seen in Fig. 8.6, are also greatly amplified. Another eventually relevant effect that can be observed in the last figure is that the transition that takes place at the epidemic threshold is much sharper in the case of correlated networks. All the previous results point out that the worst scenario for the spreading of a disease when it interacts with a second one that confers immunity to the former corresponds to the case in which there is a correlation between heterogeneous networks of contacts. This finding is essentially equivalent to what was found previously in [66], what we have shown here is that the effect comes to revert the vanishing threshold at the thermodynamic limit for a disease spreading on top of SF networks. In addition, we have found that it is not needed for this effect to take place that the second disease confers full immunity to disease one.

Remarkably all these "pathological" cases can be identified without abandoning HMF descriptions. Beyond HMF, the behavior of epidemic thresholds vary with respect to MF predictions as a consequence of dynamical correlations [401, 419]. Here, however, we identify that the interactions between diseases can modify the size scaling of thresholds from the classical MF picture without the need to recurring to dynamical correlations, which, for certain cases, remarkably modify the whole picture at the thermodynamic limit, even in annealed networks in which adjacency matrices are only fixed on average, and so, dynamical correlations do not exist.

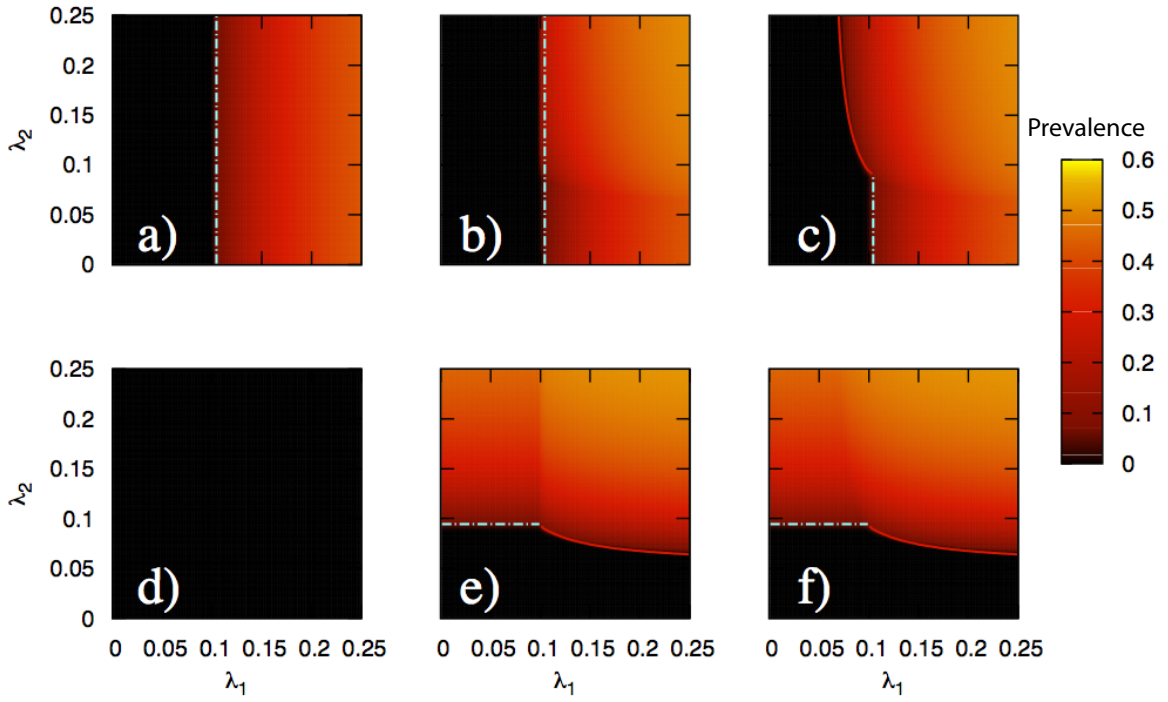


FIGURE 8.2: Reciprocally enhanced spreading. Dynamical parameters: $\mu_1 = \mu_2 = 0.75$, $\beta_1^a = \beta_1^b = \beta_2^a = \beta_2^b = 1.3$, $\eta_1 = \eta_2 = 0.8$. Networks: ER graphs: $N_1 = N_2 = 5000$ agents, $\langle k \rangle = 7$, $\langle l \rangle = 8$. The color maps represent the prevalence levels of diseases 1 – upper panels – and 2 – lower panels –, at different stages of the Monte-Carlo simulations. Step 1 (panels a,d): stationary levels after the initial introduction of an infection seed of disease 1 ($\epsilon_{IS} = 0.005$). Step 2 (panels b,e): once stage 1 is completed, an infection seed of disease 2 ($\epsilon_{SI} = 0.005$) is introduced, and stationarity is recovered. Step 3: (panels c,f): after stage 2, an additional seed of infection 1 is re-introduced ($\epsilon_{IS} = 0.005$), and the stationary prevalences plotted. Dashed and solid lines represent respectively primary and secondary thresholds predicted by the model.

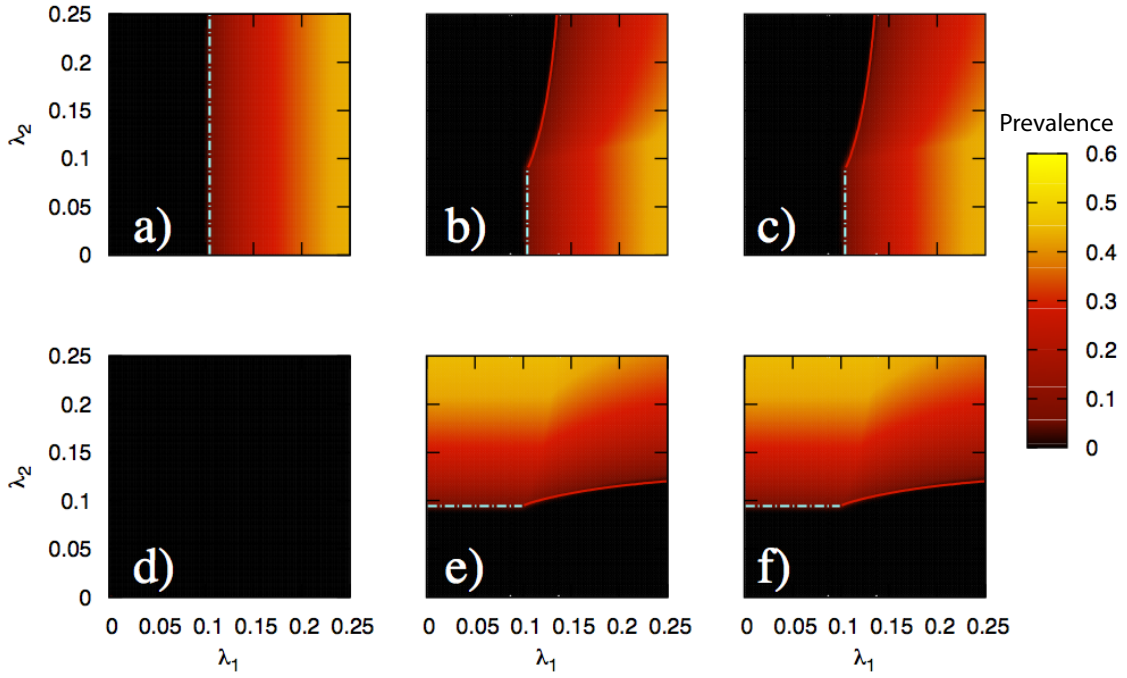


FIGURE 8.3: Reciprocally impaired spreading. Dynamical parameters: $\mu_1 = \mu_2 = 0.75$, $\beta_1^a = \beta_1^b = \beta_2^a = \beta_2^b = 0.8$, $\eta_1 = \eta_2 = 1.3$. Networks: ER graphs: $N_1 = N_2 = 5000$ agents, $\langle k \rangle = 7$, $\langle l \rangle = 8$. The color maps represent the prevalence levels of diseases 1 – upper panels – and 2 – lower panels –, at different stages of the Monte-Carlo simulations. As it is done in figure 8.2, we introduce successively three infectious seeds $(IS, SI, IS) = (0.005, 0.005, 0.005)$, and plot the stationary prevalences after each fluctuation in the three columns of the figure. As it can be seen, the reintroduction of the third seed of infection 1 in the system does not affect the prevalence levels, as global stability is reached after the second stage.

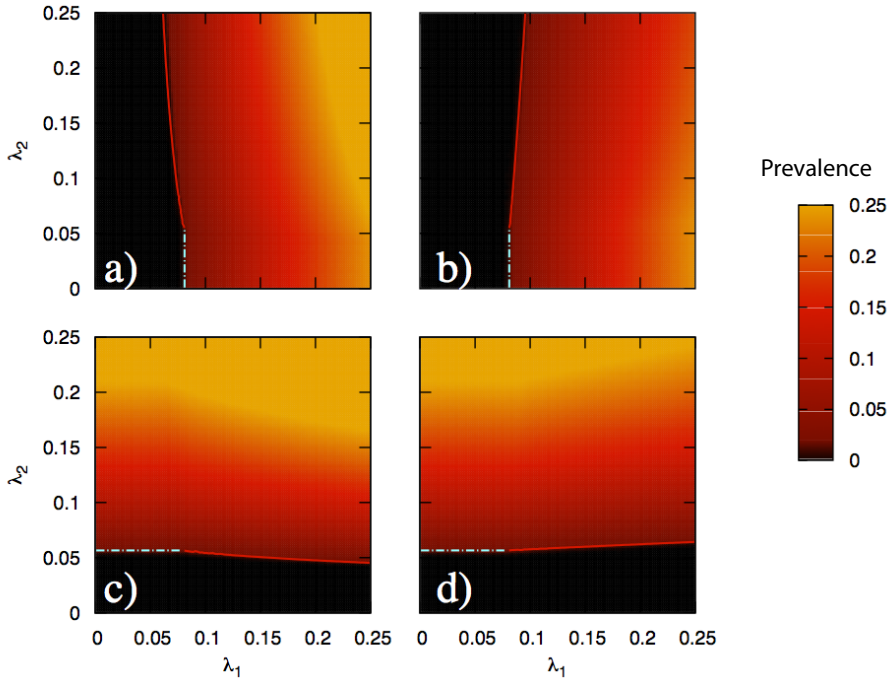


FIGURE 8.4: Panels a and c: reciprocally enhanced spreading with $\beta_1^a = \beta_1^b = \beta_2^a = \beta_2^b = 1.3$, $\eta_1 = \eta_2 = 0.8$, represented at the final stationary state. Panels b and d: reciprocally impaired spreading $\beta_1^a = \beta_1^b = \beta_2^a = \beta_2^b = 0.8$, $\eta_1 = \eta_2 = 1.3$, also at the final state. Scale-free networks are generated using the uncorrelated configuration model with $N_1 = N_2 = 5000$ agents, $\langle k \rangle = 4.00$, and $\langle l \rangle = 5.11$. The figure represents the final prevalence of diseases 1 (panels a and b) and 2 (panels c and d) in each case.

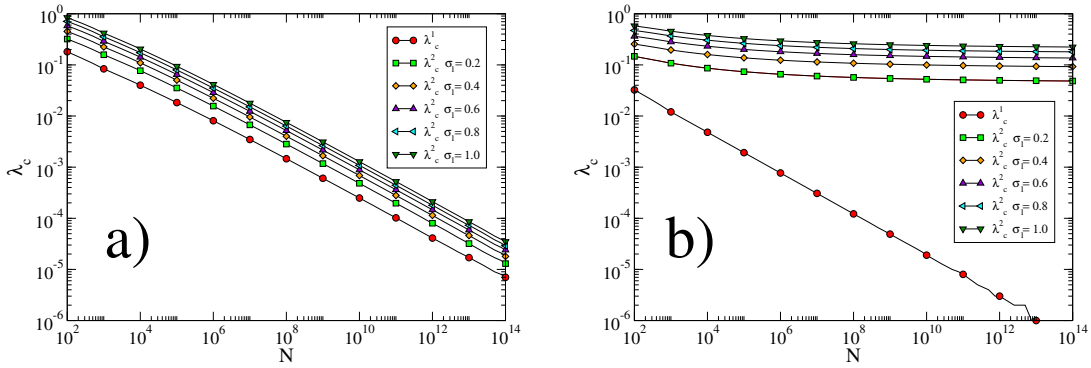


FIGURE 8.5: Primary and Secondary threshold values as a function of the size of the systems N for different values of σ_1: 0.0 (primary threshold), 0.2, 0.4, 0.6, 0.8 and 1.0. Networks are uncorrelated SF graphs (panel a) without correlations between them –panel a– or with them –panel b, fully correlated–. Diseases interact according to a full cross-immunity scheme $\beta_a^1 = \beta_a^2 = 0$ and $\gamma = \Gamma = 2.5$. The rest of the parameters are the same as in figure 8.2.

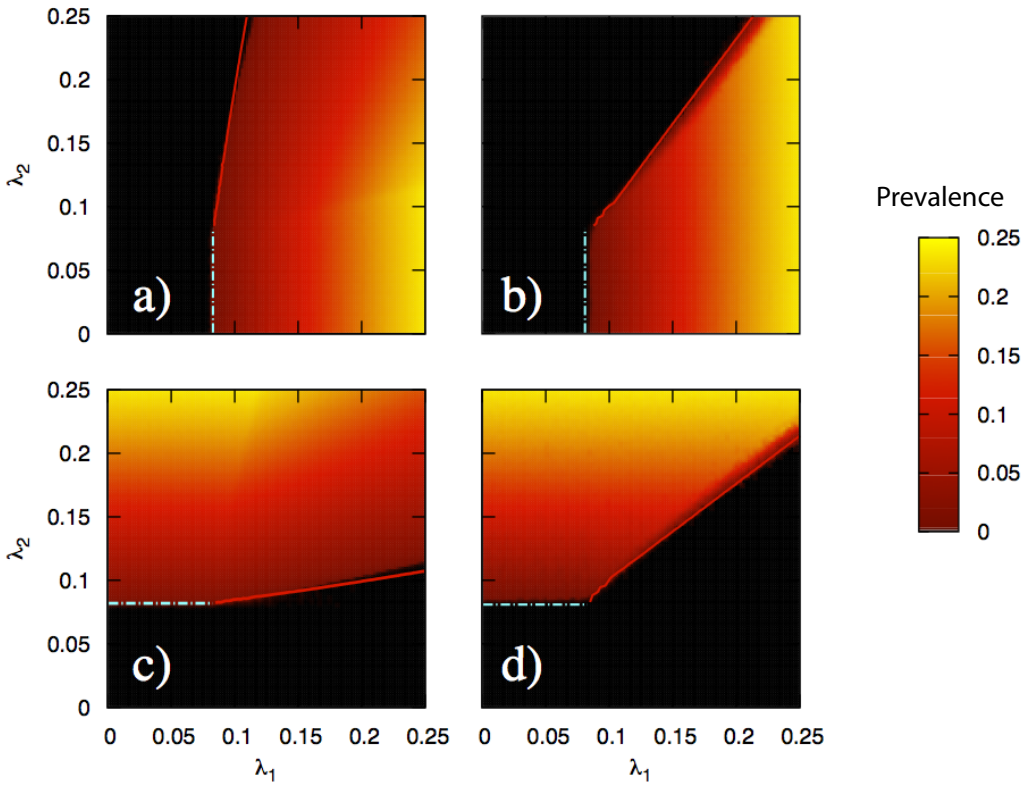


FIGURE 8.6: Effect of degree correlations between the two SF networks on the steady prevalence levels. Dynamical parameters: $\mu_1 = \mu_2 = 0.75$, $\beta_1^a = \beta_2^a = 0$. The rest of parameters are irrelevant, as no II individual will appear in the system. Panels a and c: final prevalence of diseases 1 and 2, respectively for uncorrelated SF networks: $P(k,l) = P(k)P(l) = \alpha k^{-\gamma} l^{-\gamma}$. Panels b and d: final prevalence of diseases 1 and 2, respectively for fully correlated SF networks: $P(k,l) = \alpha\delta(k-l)k^{-\gamma}$. In both cases $\gamma = 2.5$.

8.3 The SIR scenario

In the previous section we have studied the behavior of systems of interacting diseases that spread according to a SIS scheme, thus leading, above the epidemic threshold, to stationary, endemic states with a prevalence greater than zero. In the following, we will explore the case of transient, interacting epidemic phenomena.

In order to do so, we can extend the framework and describe the dynamics of two SIR epidemics interacting among them. In classical, non-interacting systems –either homogeneous or heterogeneous–, the resemblance of both types of models translates into a strong mathematical symmetry between them that yields identical expressions for the epidemic thresholds under MF descriptions (i.e. when neglecting the effects of dynamical correlations). In this section we will see the way in which part of this symmetry is broken as a consequence of the interacting nature of the epidemic processes.

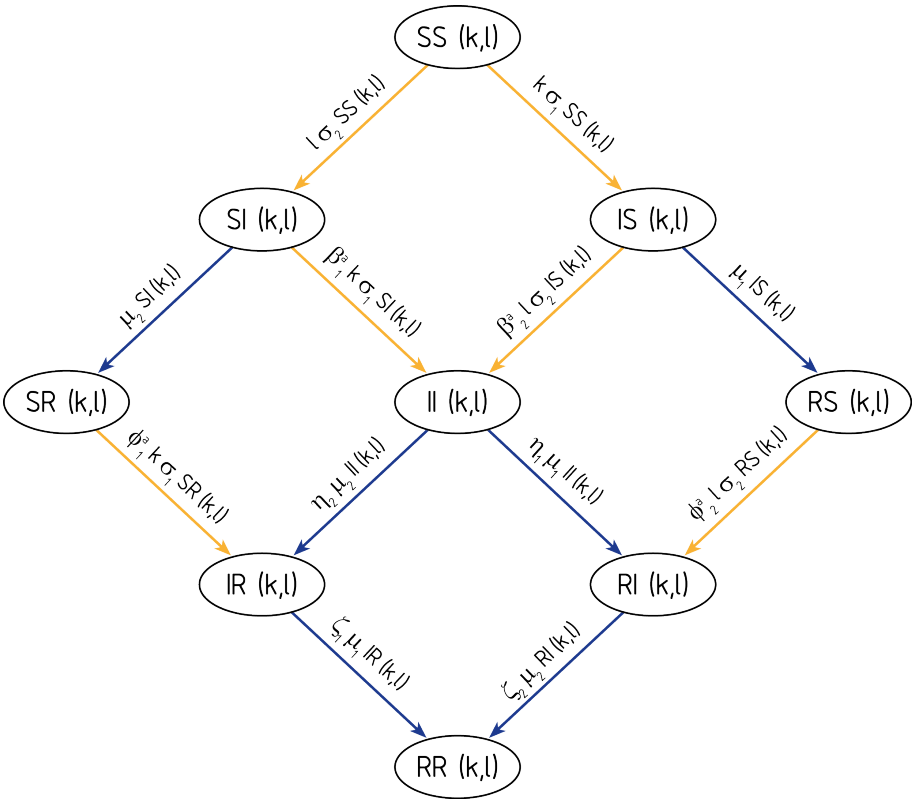


FIGURE 8.7: Set of transitions allowed in the double SIR-SIR model. Yellow: contagion processes. Blue: recovery processes

In this case, the first new aspect to note is that not just the condition of being infected with one disease can modify subjects' susceptibility to the conjugate infection, but also whether they have been infected and recovered. For instance, this might represent situations in which some kind of immunological memory is acquired after the first infection –for example, partial immunity in front of other strains is gained

New parameter	Dynamical meaning
ϕ_1^a	Variation of disease 1 infectiousness due to the fact that the susceptible individual exposed to disease 1 has been infected and recovered from disease 2
ϕ_2^a	Variation of disease 2 infectiousness due to the fact that the susceptible individual exposed to disease 2 has been infected and recovered from disease 1
ϕ_1^b	Variation of disease 1 infectiousness due to the fact that the spreader has been infected and recovered from disease 2
ϕ_2^b	Variation of disease 2 infectiousness due to the fact that the spreader has been infected and recovered from disease 1
ζ_1	Variation of disease 1 recovery rate for individuals that have been infected and recovered from disease 2
ζ_2	Variation of disease 2 recovery rate for individuals that have been infected and recovered from disease 1

TABLE 8.3: Parameters describing the influence of R classes on the conjugate infection

after suffering an influenza infection, specially if both are phylogenetically close enough [418]. This new phenomenology translates into the need of introducing new parameters (see Table 8.3) describing new eventual interactions and transitions, as shown in Fig 8.7. The system of equations describing the dynamics is now:

$$\dot{SS}(k,l) = -(k\sigma_1 + l\sigma_2)SS(k,l) \tag{8.14}$$
$$\dot{IS}(k,l) = k\sigma_1 SS(k,l) - \beta_2^a l\sigma_2 IS(k,l) - \mu_1 IS(k,l) \tag{8.15}$$
$$\dot{SI}(k,l) = l\sigma_2 SS(k,l) - \beta_1^a k\sigma_1 SI(k,l) - \mu_2 SI(k,l) \tag{8.16}$$
$$\dot{II}(k,l) = \beta_1^a k\sigma_1 SI(k,l) + \beta_2^a l\sigma_2 IS(k,l) - (\eta_1\mu_1 + \eta_2\mu_2)II(k,l) \tag{8.17}$$
$$\dot{RS}(k,l) = \mu_1 IS(k,l) - \phi_2^a l\sigma_2 RS(k,l) \tag{8.18}$$
$$\dot{SR}(k,l) = \mu_2 SI(k,l) - \phi_1^a k\sigma_1 SR(k,l) \tag{8.19}$$
$$\dot{RI}(k,l) = \phi_2^a l\sigma_2 RS(k,l) + \eta_1\mu_1 II(k,l) - \zeta_2\mu_2 RI(k,l) \tag{8.20}$$
$$\dot{IR}(k,l) = \phi_1^a k\sigma_1 SR(k,l) + \eta_2\mu_2 II(k,l) - \zeta_1\mu_1 IR(k,l) \tag{8.21}$$

where all the parameters and variables that were present in the SIS formulation retain their original meaning, with the nuance that now σ_1 and σ_2 have an extra term related to the appearance of classes IR and RI:

$$\begin{aligned}\sigma_1 &= \lambda_1(\theta_1^{IS} + \beta_1^b\theta_1^{II} + \phi_1^b\theta_1^{IR}) \\ \sigma_2 &= \lambda_2(\theta_2^{SI} + \beta_2^b\theta_2^{II} + \phi_2^b\theta_2^{RI})\end{aligned} \tag{8.22}$$

where θ_1^{IR} is the probability of a link of network 1 pointing to an IR node and θ_2^{RI} is the probability of a link of network 2 pointing to a node in the RI state.

8.3.1 Epidemic thresholds

In order to characterize the epidemic threshold of disease 2, we need to study the dynamics of the system around any point in which no infected individual of disease 2 has yet been introduced and so $(SI(k,l), II(k,l), RI(k,l)) = (0,0,0)\forall(k,l)$ as well as $\theta_2^{SI} = \theta_2^{II} = \theta_2^{RI} = 0$. Around such a disease free point -either fixed or not-, it is trivial to see that all the partial derivatives of classes $SI(k,l), II(k,l)$ and $RI(k,l)$ with respect to the rest of dynamic classes vanish. This makes the subset of variables $\{SI(k,l), II(k,l), RI(k,l)\}$ locally autonomous around that point, and so, its linearization might serve us to address the stability inversion yielding the emergence of the epidemic threshold.

In addition, the dimensionality of the system can be greatly reduced, if we write the equations driving the time evolution of the probabilities θ_2^{SI}, θ_2^{II}, and θ_2^{RI}, as follows:

$$\begin{aligned}\dot{\theta_2^{SI}} &= \frac{\sum_{k,l} P(k,l) l \dot{SI}}{\langle l\rangle} = \frac{\langle l^2 SS\rangle}{\langle l\rangle}\lambda_1(\theta_2^{SI} + \\ &+\beta_2^b\theta_2^{II} + \phi_2^b\theta_2^{RI}) - \beta_1^a\frac{\langle kl\rangle}{\langle l\rangle}\theta_2^{SI}\lambda_1(\theta_1^{IS} + \beta_1^b\theta_1^{II} + \phi_1^b\theta_1^{IR}) - \mu_2\theta_2^{SI} \\ \dot{\theta_2^{II}} &= \frac{\sum_{k,l} P(k,l) l \dot{II}}{\langle l\rangle} = \frac{\langle kl\rangle}{\langle l\rangle}\theta_2^{SI}\beta_1^a\lambda_1(\theta_1^{IS} + \beta_1^b\theta_1^{II} + \phi_1^b\theta_1^{IR}) + \\ &+\frac{\langle l^2 IS\rangle}{\langle l\rangle}\beta_2^a\lambda_2(\theta_2^{SI} + \beta_2^b\theta_2^{II} + \phi_2^b\theta_2^{RI}) - (\eta_1\mu_1 + \eta_2\mu_2)\theta_2^{II} \\ \dot{\theta_2^{RI}} &= \frac{\sum_{k,l} P(k,l) l \dot{RI}}{\langle l\rangle} = \frac{\langle l^2 RS\rangle}{\langle l\rangle}\phi_2^a\lambda_2(\theta_2^{SI} + \beta_2^b\theta_2^{II} + \phi_2^b\theta_2^{RI}) + \\ &+\eta_1\mu_1\theta_2^{II} - \zeta_2\mu_2\theta_2^{RI}\end{aligned} \tag{8.23}$$

where we have substituted $\langle klSI\rangle$ by $\langle kl\rangle\theta_2^{SI}$, approximation which is valid around the point $(SI(k,l), II(k,l), RI(k,l)) = (0,0,0)\forall(k,l)$. Obviously, as happened for $\{SI(k,l), II(k,l), RI(k,l)\}$, all of the partial derivatives of θ_2^{SI}, θ_2^{II}, and θ_2^{RI} with respect to variables unrelated to θ_2^{SI}, θ_2^{II}, and θ_2^{RI} vanish, which allows us to study the stability of the system with respect to disease 2 by linearizing the system: $(\dot{\theta_2^{SI}}, \dot{\theta_2^{II}}, \dot{\theta_2^{RI}}) = f(\theta_2^{SI}, \theta_2^{II}, \theta_2^{RI})$. The corresponding Jacobian matrix J can be calculated this way, and, by evaluating the condition $J = 0$ for stability shift, the epidemic threshold takes its final value:

$$\lambda_2^c(\langle l^2SS\rangle, \langle l^2IS\rangle, \langle l^2RS\rangle, \sigma_1) = \frac{(\eta_2\mu_2 + \eta_1\mu_1)(\mu_2\langle l\rangle + \beta_1^a\sigma_1\langle kl\rangle)\zeta_2\mu_2\langle l\rangle}{\Delta(\langle l^2SS\rangle, \langle l^2IS\rangle, \langle l^2RS\rangle, \sigma_1)} \tag{8.24}$$

where the denominator function Δ takes the form:

$$\begin{aligned}\Delta(\langle l^2SS\rangle, \langle l^2IS\rangle, \langle l^2RS\rangle, \sigma_1) = \\ = (\eta_2\mu_2 + \eta_1\mu_1)(\zeta_2\mu_2\langle l\rangle\langle l^2SS\rangle + \phi_2^a\phi_2^b\langle l^2RS\rangle(\mu_2\langle l\rangle + \beta_1^a\langle kl\rangle\sigma_1)) + \\ +(\beta_2^b\zeta_2\mu_2 + \phi_2^b\eta_1\mu_1)(\beta_1^a\langle kl\rangle\sigma_1\langle l^2SS\rangle + \beta_2^a(\mu_2\langle l\rangle + \beta_1^a\langle kl\rangle\sigma_1)\langle l^2IS\rangle)\end{aligned} \tag{8.25}$$

As we see, the threshold depends on the initial prevalence of disease 1 via σ_1, $\langle l^2IS\rangle$, $\langle l^2RS\rangle$ (and $\langle l^2SS\rangle$). In the case of having non concurrent outbreaks, we have that the second disease arrives to the system after the outbreak of the first disease has come to an end. In that case, we have that $IS(k,l) = 0$ and $RS(k,l) = R^1_{\infty,k,l}\ \forall(k,l)$, where $R^1_{\infty,k,l}$ is the fraction of recovered individuals at the end of an outbreak of disease one alone, in the composed degree class (k,l). In such a case, the problem is much simpler, as $\sigma_1 = \langle l^2IS\rangle = 0$, and $\langle l^2SS\rangle = \langle l^2\rangle - \langle l^2R^1_{\infty,k,l}\rangle$; and the threshold reads as:

$$\lambda_2^c(\langle l^2\rangle, \langle l^2R^1_{\infty,k,l}\rangle) = \frac{\mu_2\langle l\rangle}{\langle l^2\rangle + \frac{(\phi_2^a\phi_2^b - \zeta_2)}{\zeta_2}\langle l^2R^1_{\infty,k,l}\rangle}. \tag{8.26}$$

Obviously, if disease 1 has not yet appeared in the system, the threshold for disease 2 becomes $\lambda_2^c = \frac{\mu_2\langle l\rangle}{\langle l^2\rangle}$, as in the classical, non interacting HMF case [58]. As done for the SIS model, we refer to the latter expression as the primary threshold of disease 2, in counterposition to the secondary threshold of Eq. (8.25).

8.3.2 Numerical simulations

A remarkable property of the threshold for the SIR scenario is that its dependence on the dynamical state of the conjugate disease is more complex than in the SIS case. Specifically, once an outbreak of one influencing infection is unfolding, the epidemic thresholds of the other disease may vary with time in non trivial ways, depending on the nature and intensity of the different interaction mechanisms present. Figure 8.8 shows results from numerical simulations illustrating the previous phenomenology and the agreement with the analytical thresholds. Specifically, each panel represents the case in which the infection seed of the second disease is introduced in different moments for each topology, coinciding with different stages of the outbreak of disease 1: early phases (panel a, SF and panel c, ER), outbreak's peak (panel b, SF) and at the end of the outbreak (panel d, ER). As we can see, regardless of the topology, the parameters' values or the moment of appearance of the infection seed, our model adequately foresees the epidemic threshold and its variations with time. As in the SIS case, the influence of the interactions on the epidemic threshold is smaller in the case of SF networks, due to the smaller values of the primary thresholds in that case (in fact, primary thresholds have not been represented in panels a and b for the sake of clarity, because its values $\lambda_{2c}^1 = 0.00316$ virtually overlaps the secondary ones).

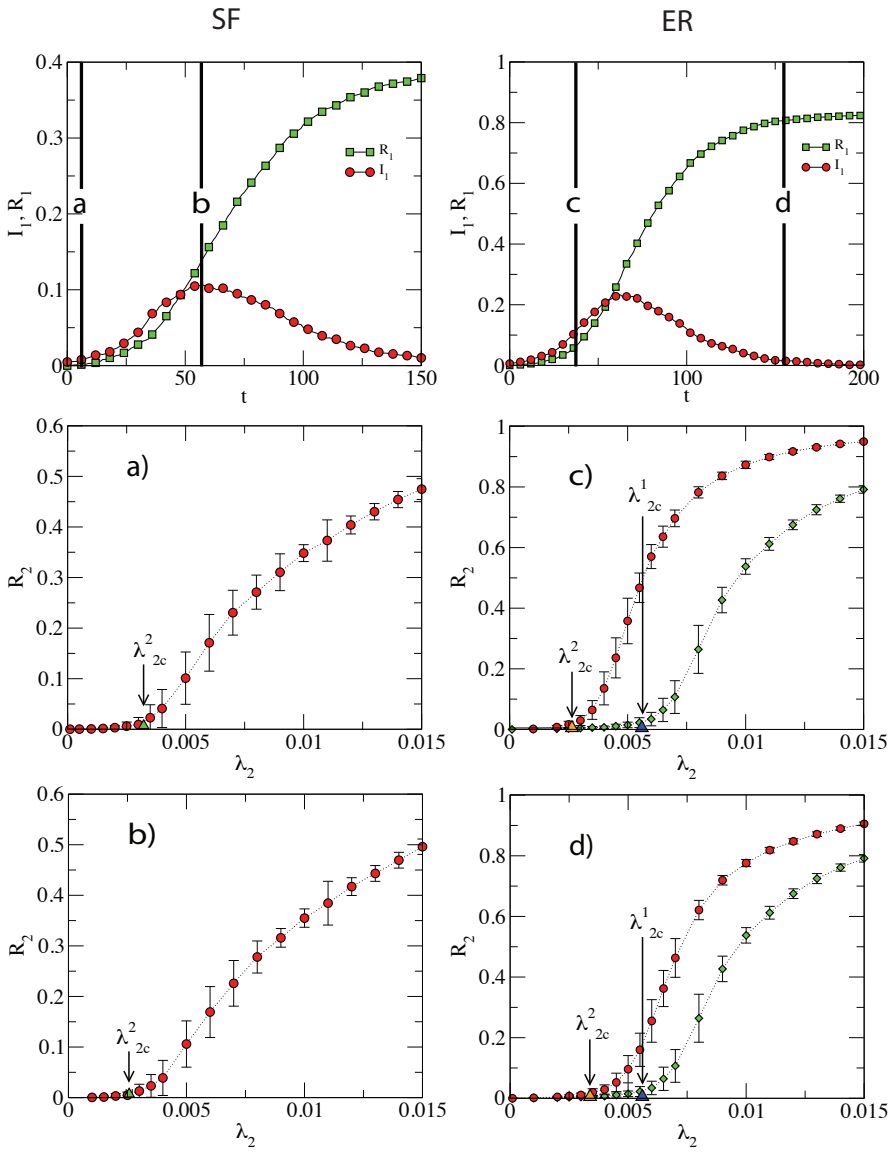


FIGURE 8.8: Epidemic thresholds in the SIR-SIR interacting epidemic model. Upper panels: temporal evolution of disease 1, in a SF (left) and in an ER network (ER, right): $I_1 = \sum_{k,l} P(k,l)(IS_{k,l} + II_{k,l} + IR_{k,l})(t)$, $R_1 = \sum_{k,l} P(k,l)(RS_{k,l} + RI_{k,l} + RR_{k,l})(t)$. Disease one enhances the spreading of disease two in both cases, but still is not influenced by the second infection. Lower panels: total number of affected individuals by disease 2 as a function of λ_2 in each network: $R_2^\infty = \lim_{t\to\infty} \sum_{k,l} P(k,l)(SR_{k,l} + IR_{k,l} + RR_{k,l})(t)$. Dynamical parameters: panels a, b: SF network: $N = 5000$, $\langle k\rangle = 4.00$,$\langle l\rangle = 5.11$, $\lambda_1 = 0.02$ $\mu_1 = \mu_2 = 0.05$, $\beta_2^a = \beta_2^b = \phi_2^a = \phi_2^b = 1.3$ $\eta_2 = \zeta_2 = 0.8$, panels c,d: ER network: $N = 5000$, $\langle k\rangle = 7$,$\langle l\rangle = 8$, $\lambda_1 = 0.02$ $\mu_1 = \mu_2 = 0.05$, $\beta_2^a = \beta_2^b = 3$, $\phi_2^a = \phi_2^b = 1.2$, $\eta_2 = 0.33$, $\zeta_2 = 0.8$. As said before, in both cases, disease 1 is not influenced by disease 2, and so: $\beta_1^a = \beta_1^b = \phi_1^a = \phi_1^a = \eta_1 = \zeta_1 = 1$.

8.4 Conditions for disease enhancement and impairment

Finally, it is interesting to analyze what combinations of parameters give raise to situations in which there is either enhancement or impairment of the diseases, as well as some other cases in which both effects are possible. This can be done for both the SIS and the SIR scenarios by studying the sign of the difference between the primary and the secondary thresholds.

For the SIS case, if we focus, for example, on the second disease, we have:

$$\lambda_2^c(\sigma_1) - \lambda_2^c(0) = \\ \sum_{k,l} C(k,l)\sigma_1 \left[\beta_1^a(\eta_2 - \beta_2^a\beta_2^b)k\sigma_1 + (\eta_2 - \beta_2^a\beta_2^b)\mu_2 + \left(\eta_1(1-\beta_2^a) + \beta_1^a(\eta_2 - \beta_2^b)\right)\right] \tag{8.27}$$

where $C(k,l)$ is always positive. In this sense, $\lambda_2^c(\sigma_1) - \lambda_2^c(0) > 0$ implies that disease 1 is reducing the epidemic threshold of disease 2, thus enhancing its spreading. In the opposite case $\lambda_2^c(\sigma_1) - \lambda_2^c(0) < 0$, disease 1 makes the secondary threshold for disease 2 to be bigger than the primary one, hence impairing its spreading. As we can see, the condition yielding one or another case involves a complex combination of the parameters. However, it is trivial to show that, provided that $\beta_2^a > 1$, $\beta_2^b > 1$ and $\eta_2 < 1$, disease 1 enhances disease 2 spreading for any value of σ_1 greater than zero. We call this scenario coherent enhancement, because all the interaction mechanisms contribute to enhance the spreading of disease 2. In a similar way, if $\beta_2^a < 1$, $\beta_2^b < 1$ and $\eta_2 > 1$, the interaction has the opposite sign regardless of σ_1, and we have a situation of coherent impairment. Noticeably, if none of these conditions is fulfilled, there exists the possibility that the sign of the influence that disease 1 exerts on the spreading of disease 2 depends on its prevalence via σ_1 (see below). In Fig. 8.9 we have represented $\lambda_2^c(\sigma_1) - \lambda_2^c(0)$ for different parameter combinations that cover all possible phenomenologies.

For the SIR case, the proliferation of dynamical classes and the possibility that other mechanisms (i.e., those including R individuals) carry the interaction between both disease makes the equivalent expression more complex, but still derivable as:

$$\begin{aligned} &\lambda_2^c(\langle l^2SS\rangle, \langle l^2IS\rangle, \langle l^2RS\rangle, \sigma_1) - \lambda_2^c(\langle l^2\rangle, 0, 0, 0) = C'[\mu_2\langle l\rangle(\eta_1\mu_1 + \eta_2\mu_2)\cdot \\ &\cdot\left(\zeta_2\beta_1^a\sigma_1\langle kl\rangle\langle l^2\rangle + \zeta_2\mu_2\langle l\rangle\langle l^2IS\rangle + ((\zeta_2 - \phi_2^a\phi_2^b)\mu_2\langle l\rangle - \beta_1^a\langle kl\rangle\sigma_1)\langle l^2RS\rangle\right) \quad - \\ &\quad -\langle l\rangle\mu_2(\phi_2^b\eta_1\mu_1 + \beta_2^b\zeta_2\mu_2)\left(\beta_1^a\sigma_1\langle kl\rangle\langle l^2SS\rangle + \beta_2^a(\mu_2\langle l\rangle + \sigma_1\langle kl\rangle)\langle l^2IS\rangle\right)] \end{aligned} \tag{8.28}$$

where C' is always positive too.

As we have already said, in the interacting SIR model, once an outbreak of one disease has started, the temporal dependence of the epidemic threshold of the second infection depends on the dynamic state of the system as well as on the parameter values that account for the mechanisms of interaction that are present in the system and their intensities. In such a case, if the interaction between the diseases is mediated by the class R, (i.e. being recovered of one disease is what makes subjects dynamics with

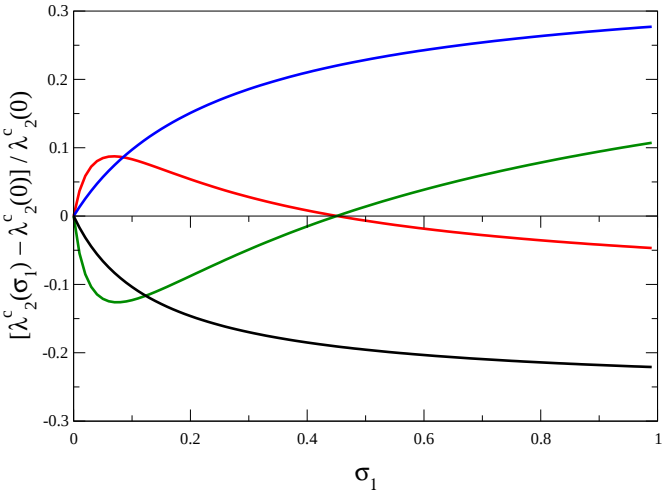


FIGURE 8.9: Relative variation of epidemic threshold of disease 2 as a function of σ_1 for different parameter sets. Black: (coherent enhancement of disease 1 over disease 2): $\lambda_1 = 0.2$, $\mu_1 = \mu_2 = 0.3$, $\beta_2^a = \beta_2^b = \beta_1^a = \beta_1^b = 1.1$, $\eta_1 = \eta_2 = 0.9$. Blue: (coherent impairment of disease 1 over disease 2): $\lambda_1 = 0.2$, $\mu_1 = \mu_2 = 0.3$, $\beta_2^a = \beta_2^b = \beta_1^a = \beta_1^b = 0.9$, $\eta_1 = \eta_2 = 1.1$. Red: $\lambda_1 = 0.2$, $\mu_1 = 0.4$, $\mu_2 = 0.2$, $\beta_2^a = \beta_2^b = 1.5$, $\beta_1^a = 4$, $\beta_1^b = 0.5$, $\eta_1 = 0.4$, $\eta_2 = 2$. Green: $\lambda_1 = 0.2$, $\mu_1 = 0.4$, $\mu_2 = 0.2$, $\beta_2^a = 0.5$, $\beta_2^b = 1.5$, $\beta_1^a = 4$, $\beta_1^b = 0.5$, $\eta_1 = 0.4$, $\eta_2 = 1$. The networks are two ER graphs: $N_1 = N_2 = 5000$ agents, $\langle k \rangle = 7$, $\langle l \rangle = 8$.

respect to the other infection to be altered, and so β and η parameters are equal to 1), the density of individuals belonging to the R class is a monotonically increasing function of time. This trivially implies that, in general, the closer to its end is the epidemic process for one disease, the deeper its impact on the conjugate one. However, if it is the I class the one that carries the interaction, it is easy to see that that interaction will take place if and only if the outbreak of disease two happens within a time window that -at least partially- overlaps with that characterizing the outbreak of disease 1. In such a situation, the interaction between the two diseases is a transitory property of the system, and the shift in the epidemic threshold crucially depends on the temporal co-occurrence of the outbreaks of both diseases.

A similar time-dependent shift of the epidemic thresholds can also take place in the SIS model, for example for the thresholds of disease 2, if the difference $\lambda_2^c(\sigma_1) - \lambda_2^c(0)$ changes its sign during an outbreak of disease 1, as a consequence of the evolution of σ_1. This is the scenario given by the green and red curves in Fig. 8.9. However, it is worth remarking that such combinations of parameters are somehow pathological, as they imply the appearance of diverse non-coherent mechanisms of interaction (i.e. one disease simultaneously enhances subjects susceptibility to disease but, at the same time, it reduces their infectiousness, etc.) that are epidemiologically less plausible. On the contrary, in the SIR framework, this transitoriness is an intrinsic property of the

system when the infected class is the responsible of influencing the dynamics of the conjugate disease. This transitory nature of the interactions constitutes an essential difference between the double SIR model with respect to the SIS one, which is at the root of the remarkable difference between their corresponding thresholds and the richer phenomenology of two interacting SIR-like processes.

8.5 Conclusions

In this chapter, we propose a composed HMF model to describe the simultaneous spreading of two diseases over the same population but driven by independent mechanisms of transmission taking place on different networks of contacts. Within this framework, we have thoroughly studied extensions of the SIS and SIR models, where the parameters defining the infectious and recovery transitions of one disease depend on the state of each node with respect to the conjugate disease, establishing in this way a coupling between both diffusion processes. Our modeling approach presents different advantages with respect to previous models [65, 66, 67, 68], as it simultaneously allows analytical derivations of the epidemic thresholds and an approximate description of the temporal evolution of the system, besides providing a way to isolate the effects on spreading dynamics of each possible interaction mechanism, such as variations of infectivity, susceptibility or infectious periods. In addition, it enables us to solve the two paradigmatic modeling scenarios (SIS and SIR), identifying relevant differences between the two cases that arise as a consequence of disease interactions.

The model here presented and analyzed, even if it is based on an HMF and so, constitutes a first approximation to the problem that neglects dynamical correlations among nodes, is able to identify novel phenomena qualitatively different to what is found on HMF descriptions of non interacting systems. First, our model foresees different thresholds for SIR and SIS models, which arises from the asymmetry between the interaction mechanisms that take place under both models. Additionally, in what regards size scaling of epidemic thresholds, at least in the SIS case, we have identified some relevant situations that yield threshold dependences on network's size which are different from what is found in HMF models for single diseases. At this particular, it is worth mentioning that asymmetries between SIS and SIR critical properties [419, 401], or a different behavior with respect to that predicted by the HMF for epidemic thresholds [405] have also been identified in the context of single diseases spreading over one network, if the HMF approach is abandoned.

The situation here is qualitatively different, as all these divergent results with respect to classical HMF models of single diseases do not arise as a consequence of considering any dynamical correlation but as a consequence of disease-disease interactions. This has two relevant implications. On the one hand, deviations from HMF results on epidemic thresholds on single diseases identified as a consequence of considering dynamical correlations are of quantitatively residual relevance, precisely because, at the epidemic threshold, those correlations tend to vanish and, there, HMF models perform consistently well [405]. Instead, in our case, differences between SIS and SIR

thresholds, for example, are totally different, as they even may depend on conceptually different interaction mechanisms. On the other hand, HMF is exact when dealing with annealed networks [401], and so, on these type of networks which lack, by definition, dynamical correlations, both SIS-SIR asymmetries and eventually anomalous threshold dependences with size that we have identified in this chapter constitute phenomena not found before.

In addition, for the first time, the modeling framework proposed here allows to isolate the independent effects of the different mechanisms of interaction that can determine the critical properties of the model. On the one hand, this allows us to foresee that, in a SIS model, certain interaction schemes may yield to effects of one disease on another of different sign as a function of the prevalence levels of the former. On the other hand, for the SIR model, we have also discussed how the interaction between the two diseases and the different dynamical classes give rise to a richer phenomenology. In particular, we have shown that due to the transitory nature of SIR spreading processes, the moment at which the interaction of the two diseases is made effective might greatly determine the values of the epidemic threshold of the disease whose course is modified by the other one.

Further advances in the field should address the influences of dynamical correlations between nodes on the spreading of interacting epidemics. This constitutes a truly conceptual challenge as, for example, pairwise descriptions of the disease dynamics [419], should transform into quadruplet-based descriptions, as, in this case, dynamical correlations go beyond the dynamical state of neighbors in a network, but comprises both dynamical states on the two networks involved, which multiplies the complexity of the description.

8.6 Confronting the model to data: the case of HIV-TB syndemics in South-Africa

To illustrate the usefulness of our approach, let us discuss a real and relevant scenario in which two diseases coexist in the same host population: that of the interaction between the acquired immunodeficiency syndrome (AIDS) and certain infections that spread from person to person and that are caused by opportunist pathogens usually associated to the immunodeficiency syndrome in advanced phases of the disease. Examples of these kind of diseases are respiratory disorders caused either by bacteria (like TB [251]) or fungi (like pneumonia caused by *Pneumocystis* [70]). Other examples of pathologies that appear commonly associated to advanced stages of AIDS are infections generated by the two variants of human herpes virus [69] and candidiasis [70], which is due to the fungal pathogen *Candida albicans*.

In some regions, the increase of the epidemiological risk related to some of the previous diseases after the irruption of HIV in the last decade of the XXth century

	TB prev. $\cdot 10^5$ (1990 - 2000)	HIV prev. (%) (1990 - 2000)	% of HIV+ of new TB cases(%) (1990 - 2000)
Malawi	352 - 356	6.08 - 9.80	51 - 69
Ethiopia	434 - 427	0.68 - 1.98	13 - 33
Tanzania	322 - 238	2.51 - 4.41	31 - 45
Kenya	243 - 262	1.19 - 5.12	17 - 52
Zimbabwe	239 - 328	5.54 - 15.19	69 - 92
Nigeria	287 - 307	0.41 - 1.86	2.2 - 25
South Africa	435 - 536	0.13 - 8.71	3.6 - 55

TABLE 8.4: Evolution of active TB and HIV infection prevalences in most-populated, high HIV-burden countries in sub-saharan Africa (more than 1 million people estimated to live with HIV). Data extracted from public databases for Tuberculosis (WHO TB database: [24]) and HIV (UNAIDS database curated by United Nations [411])

has reached the dimension of a major threat for public health. The paradigmatic case is the recent boost of TB burden in sub-Saharian Africa, which is tightly related to the dramatic prevalence levels of HIV in that region, as it can be seen in figure 8.10. Given that our main purpose is to model these variations in the endemic prevalence level of one disease due to the irruption of another infection, our SIS-based modeling framework is the simplest way to recover an endemic equilibrium of a disease. In our case, this endemic equilibrium can be altered as a consequence of the irruption of the conjugate infection, in the same way that the appearance of HIV supposed an increase of TB burden in certain countries like the Republic of South Africa.

Figure 8.11 shows a comparison of the results obtained with our model with real data for the case of the Republic of South Africa during the period 1990-2011. In particular, we represent the results of a run of the MF model with a set of parameters obtained through a Levenberg-Marquard [413] least-squares fitting approach (see Figure caption).

A first obvious conclusion extractable after a first glance on figure 8.11 is that, even if a simple SIS is a caricatural simplification of the much more complex natural histories of both diseases, our model is remarkably accurate at reproducing the disease burden evolution for both pathologies and their combinations, after an adequate fitting procedure. This shows that our conceptual modeling framework is essentially compatible with the observation of that the interaction between HIV and TB is the major force responsible of TB enhancement in South-Africa; something which is currently accepted as a dramatic, incontrovertible fact.

But even more relevant than the ability of the model to reproduce the burden curves (fitting procedures –specially those that are highly parametric– many times make wrong models reproduce real data), are the values of the parameters obtained after the fitting procedure. The values of the parameters obtained through the fitting

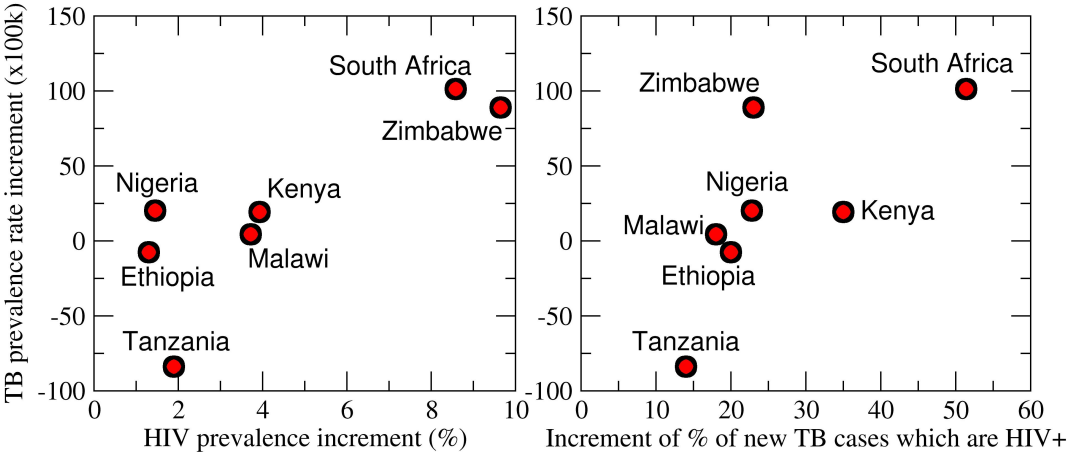


FIGURE 8.10: Scatter plot for the variation of active TB prevalence rates, HIV infection prevalence and TB-HIV coinfection in the countries listed in table 8.4. The increase of HIV infection and related immunodeficiency syndrome on the population in the last decade of the XX^{th} century has been identified as the main cause of the increase of TB incidence and prevalence rates [157], as is evidenced not only by the correlation between HIV-infection and active TB prevalences but also by the correlation between TB burden and HIV-TB association frequencies. Remarkably, there exist some countries (Tanzania, Ethiopia) in which HIV irruption has not been strong enough to prevent the decay of TB burden levels due to other causes -paradigmaticaly socio-economic improvements and effective implementation of specific public healths programs, generally speaking-, that is observed in most of the countries worldwide (see chapter 9). Data extracted from table 8.4.

present two relevant features perfectly compatible to what is actually observed for the syndemics HIV-TB.

On the one hand, the driving effect in the interaction between both epidemics, in terms of our fitted model, seems to be the increased susceptibility of individuals to get infected with one disease once infected with the other, rather than a significant increment in the spreading capabilities of double-infected individuals ($\beta_1^a > \beta_1^b$ and $\beta_2^a > \beta_2^b$). If we think about TB susceptibility of HIV-positive individuals, we have that they are much more prune to develop TB (either after infection or endogenous reactivation of disease) than HIV negative individuals [171], which is mapped, within our modeling framework with a high value for β_1^a. Instead, the lack of lymphocytes that doubly infected individuals suffer (because of HIV) imposes a great limitation to the magnitude of the immune response against TB bacilli, specially to its acquired component. This immune system underperformance translate into an inability for the host to adequately contain bacterial growth within lung granuloma, which have two consequences: on the one hand, doubly infected individuals have a greater tendency to develop extended forms of milliar TB, in which bacteria spread through the whole host body; causing a medical condition with a very serious medical prognosis. On the other hand, doubly infected individuals have great difficulties to contain bacterial growth

within granuloma, which, at the end, means that smaller and lesser tuberculous cavities are formed in the lungs than what is found for HIV negative, TB patients. This ultimately make doubly infected individuals much less infectious than HIV negative patients, as we remarkably attain after our fitting procedure ($\beta_1^b < 1$). These phenomena have been widely reported in medical literature [415, 416, 417], and even incorporated on models for TB-HIV syndemics spreading [171] on homogeneous populations.

The explanation of how TB influences HIV contagion dynamics is, however, more speculative. Although the enhancement of the infectiousness of doubly infected individuals (described by β_2^b) is the most modest effect among all infectiousness variations of the model, the fitting procedure evidences a very strong increase for TB sick individuals' susceptibility to HIV ($\beta_2^a > 30$). This extreme enhancement effect seems to be associated to an enhanced probability of receiving an HIV diagnosis corresponding to a medical condition previously unnoticed, after a severe health worsening associated to TB disease.

The second key feature of the parameter set we obtain as a result of the fitting procedure is a reduced infectious period for doubly infected individuals for both diseases, which translates in values for η parameters greater than one. This feature is also perfectly compatible with real observations about HIV-TB syndemics in South Africa, because of two reasons. The first of these is the heavy increase of the mortality rates associated to TB-HIV co-infection [414], with respect to any of the diseases separately. This crucial factor in the interaction between both diseases, although it is not explicitly described by our model, obviously contributes to reduce the infectious periods of both diseases for TB-HIV co-infected individuals. The second reason is the increased detection rate that is described for doubly infected individuals [171], which is related with a faster tendency to recur to the health services associated after the fast health decay that follows TB-HIV co-infection.

It is worth remarking that the actual interaction between both diseases is much more complex than the description provided by our model, as it involves many different phases of the natural history of the diseases, which are in turn much more complex than a simple SIS. Therefore, this makes the precise values of model parameters rather contingent. However, as we see in figure 8.11, our model remarkably captures the effects of one disease on the dynamics of the other. Therefore, this framework helps understanding what ingredients are basic and what others can be thrown out in a first approximation − it seems that for the real case here discussed, adding the coupling of the two diseases to simple models is effective enough to understand the coupled temporal evolution of both diseases. The previous comparison also indicates that modeling approaches like the one here introduced can be used to develop applied, data-driven models aimed at evaluating the increase in epidemiological risk due to AIDS high-prevalence levels on other diseases, like TB [22].

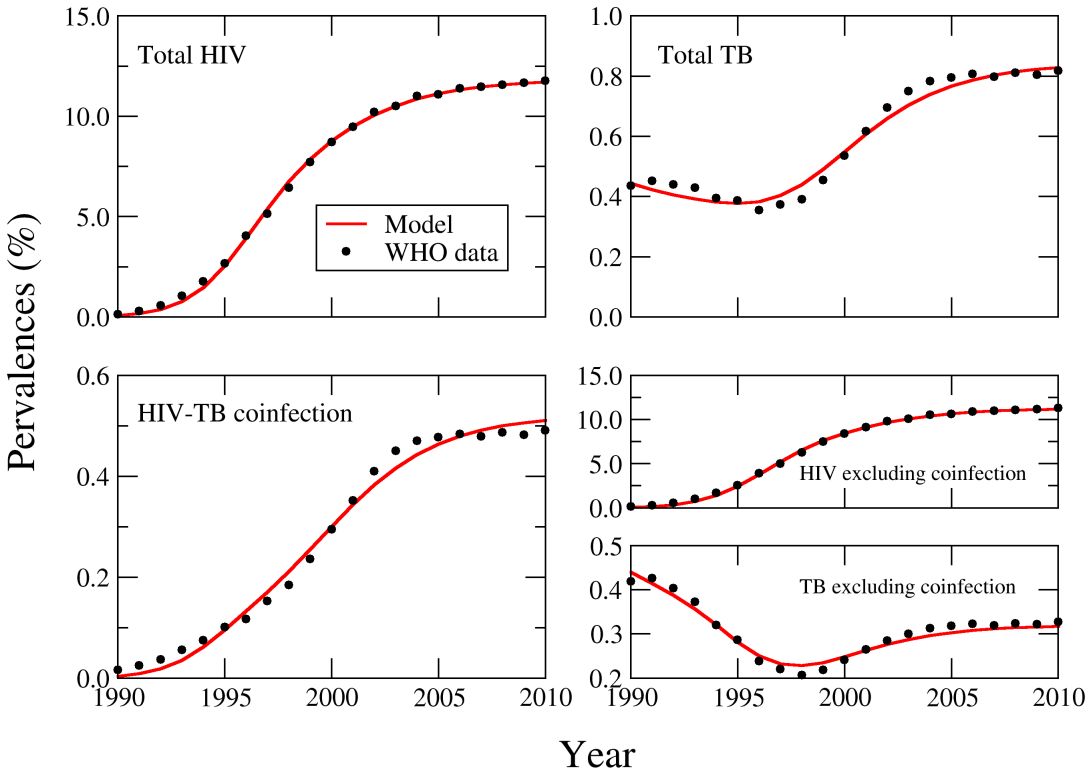


FIGURE 8.11: Prevalence percentages for HIV infection and active TB in the Republic of South Africa from 1990 to 2011. In black, we represent the prevalence levels of both diseases, as available online at public databases for Tuberculosis [24] and HIV surveillance [411]. Even if the natural history of both pathologies are much more complex than a naive SIS model [22, 412], our model is able to reproduce the interactions between both diseases, and more specifically, the increment in TB's prevalence as a consequence of the irruption of HIV, mostly associated to co-infection. In the figure, in red, we represent the results of a run of the MF model with the following parameters, obtained through a Levenberg-Marquard [413] least-squares fitting approach: $\langle k\rangle\lambda_1 = 6.011$, $\langle l\rangle\lambda_2 = 1.177$, $\beta_1^a = 21.56$, $\beta_2^a = 30.76$, $\beta_1^b = 0.03451$, $\beta_2^b = 1.719$, $\mu_1 = 29.97$, $\mu_2 = 5.734$, $\eta_1 = 1.476$, $\eta_2 = 2.548$. The first disease represents TB and the second one, HIV. The initial conditions are settled in 1988 for TB, (0.5% of prevalence of IS individuals) and, a year after, a 0.03% of SI individuals (HIV first carriers) is introduced. In this exercise the fraction of prevalent cases of TB being HIV+ has been approximated by the fraction of new TB cases being HIV+, available at the database [24]. This is a reasonable first approximation due to the fact that the order of magnitude of TB infectious period is of the order of 1 year (see chapter 9). See the text for further discussions.

8.7 Supplementary analysis

8.7.1 Epidemic thresholds on regular networks (SIS case)

In the particular case in which both networks are regular graphs, an independent derivation of the epidemic threshold can be obtained by a linear stability analysis. In that case, for which $P(k,l) = \delta(k-k_o)\delta(l-l_o)$, the dynamics can be described by the following system of four equations:

$$\begin{aligned}
\dot{SS} &= -k_o\lambda_1 SS\cdot IS - l_o\lambda_2 SS\cdot SI - k_o\beta_1^b\lambda_1 SS\cdot II - l_o\beta_2^b\lambda_2 SS\cdot II + \mu_1 IS + \mu_2 SI,\\
\dot{IS} &= k_o\lambda_1 SS\cdot IS + k_o\beta_1^b\lambda_1 SS\cdot II - l_o\beta_2^a\lambda_2 SI\cdot IS - l_o\beta_2^a\beta_2^b\lambda_2 IS\cdot II - \mu_1 IS + \eta_2\mu_2 II,\\
\dot{SI} &= l_o\lambda_2 SS\cdot SI + l_o\beta_2^b\lambda_2 SS\cdot II - k_o\beta_1^a\lambda_1 IS\cdot SI - k_o\beta_1^a\beta_1^b\lambda_1 SI\cdot II - \mu_2 SI + \eta_1\mu_1 II,\\
\dot{II} &= k_o\beta_1^a\lambda_1 IS\cdot SI + l_o\beta_2^a\lambda_2 IS\cdot SI + k_o\beta_1^a\beta_1^b\lambda_1 SI\cdot II + l_o\beta_2^a\beta_2^b\lambda_2 IS\cdot II - (\eta_1\mu_1+\eta_2\mu_2)II, \qquad (8.29)
\end{aligned}$$

one of which is linearly dependent of the rest. Thus, we will only analyze the system constituted by the three last equations and use $s = 1 - IS - SI - II$. In order to perform a linear stability analysis we first linearize the system around the equilibrium point and we calculate the Jacobian to get:

$$J = \begin{vmatrix} \lambda_1 k_o - \mu_1 & 0 & k_o\beta_1^b\lambda_1 + \eta_2\mu_2 \\ 0 & \lambda_2 l_o - \mu_2 & l_o\beta_2^b\lambda_2 + \eta_1\mu_1 \\ 0 & 0 & -(\eta_1\mu_1+\eta_2\mu_2) \end{vmatrix} \qquad (8.30)$$

which, taking advantage of the fact that the Jacobian itself is just the product of the elements in the diagonal, yields the stability conditions detailed in table 8.7.1.

If we look carefully at the results sketched in table 8.7.1, we firstly recognize the appearance of a couple of critical values for the infectiousnesses λ_{1i}^c and λ_{2i}^c which will be referred to in what follows as the primary thresholds of their respective diseases, just like in the general case described in the main text. These threshold values stand for the *minimum values of the infectiousnesses that yield epidemic outbreaks after the introduction of infinitesimal seeds of infected individuals of each disease in an initially healthy population.* Therefore, the condition $\lambda_1 > \lambda_{1i}^c$ must be verified in order to have an epidemic outbreak for the first disease once an infinitesimal seed

Eigen-value	Eigen-vector (IS, SI, II)
$\xi_1 = \lambda_1 k_o - \mu_1$	$\vec{\psi_1} = (1,0,0)$
$\xi_2 = \lambda_2 l_o - \mu_2$	$\vec{\psi_2} = (0,1,0)$
$\xi_3 = -(\eta_1\mu_1+\eta_2\mu_2)$	$\vec{\psi_3} = \left(-\frac{\eta_2\mu_2+\beta_1^b\lambda_1 k_o}{(\lambda_1 k_o-\mu_1)+(\eta_1\mu_1+\eta_2\mu_2)}, -\frac{\eta_1\mu_1+\beta_2^b\lambda_2 l_o}{(\lambda_2 l_o-\mu_2)+(\eta_1\mu_1+\eta_2\mu_2)}, 1\right)$

TABLE 8.5: Linear stability analysis of the fixed point $(IS, SI, II) = (0,0,0)$.Threshold values correspond to stability inversion points, and are reached for $\lambda_1 = \lambda_{1i}^c = \mu_1/k_o$ (stability inversion in the direction of $\vec{\psi_1}$) and $\lambda_2 = \lambda_{2i}^c = \mu_2/l_o$ (stability inversion in the direction of $\vec{\psi_2}$). In the direction of the third eigenvector, the system is always stable.

$(IS, SI, II) = (\epsilon, 0, 0)$ has been introduced on a system being at the disease-free fixed point $(IS, SI, II) = (0, 0, 0)$. On the other hand, the condition λ_{2i}^c plays an equivalent role for the second disease with respect to a seed $(IS, SI, II) = (0, \epsilon, 0)$.

The fixed point $(IS, SI, II) = (0, 0, 0)$ is, strictly speaking, not the only possible disease-free fixed point in our model. Other two partially disease-free fixed points can exist: a first fixed point in which disease 1 is installed in the system at a certain prevalence π_1 whilst disease 2 is absent $(IS, SI, II) = (\pi_1, 0, 0)$ and its cognate $(IS, SI, II) = (0, \pi_2, 0)$, for which there is no individual infected with disease 1. As we are going to show, the stability of these fixed points depends on the prevalence fractions π_1 and π_2, which are the stationary proportions of sick individuals of each disease:

$$\pi_1 = \frac{\lambda_1 k_o - \mu_1}{\lambda_1 k_o} \tag{8.31}$$

$$\pi_2 = \frac{\lambda_2 l_o - \mu_2}{\lambda_2 l_o} \tag{8.32}$$

Considering that, let us study the stability of the first fixed point $(IS, SI, II) = (\pi_1 = \frac{\lambda_1 k_o - \mu_1}{\lambda_1 k_o}, 0, 0)$ as a function of λ_1: the Jacobian, around this fixed point takes the following form:

$$J = \begin{vmatrix} (\mu_1 - \lambda_1 k_o) & (\mu_1 - \lambda_1 k_o)(1 + \beta_2^a \frac{\lambda_2 l_o}{\lambda_1 k_o}) & \beta_1^b \lambda_1 k_o + \eta_2 \mu_2 + (\mu_1 - \lambda_1 k_o)(\beta_1^b + 1 + \beta_2^a \beta_2^b \frac{\lambda_2 l_o}{\lambda_1 k_o}) \\ 0 & (\lambda_2 l_o - \mu_2) + (\mu_1 - \lambda_1 k_o)(\frac{\lambda_2 l_o}{\lambda_1 k_o} + \beta_1^a) & \beta_2^b \mu_1 \frac{\lambda_2 l_o}{\lambda_1 k_o} + \eta_1 \mu_1 \\ 0 & (\lambda_1 k_o - \mu_1)(\beta_2^a \frac{\lambda_2 l_o}{\lambda_1 k_o} + \beta_1^a) & (\lambda_1 k_o - \mu_1)\beta_2^a \beta_2^b \frac{\lambda_2 l_o}{\lambda_1 k_o} - (\eta_1 \mu_1 + \eta_2 \mu_2) \end{vmatrix} \tag{8.33}$$

By solving the equation derived from imposing that $J = 0$, we obtain the values of the parameters leading to stability inversion. As it can be shown at naked eye, $(\mu_1 - \lambda_1 k_o)$ is the eigenvalue ξ_1 associated to the eigenvector $\vec{\psi_1} = (1, 0, 0)$, and the condition $\xi_1 = 0$ yields again the same condition $\lambda_1 = \mu_1 / k_o$, thus defining the threshold of the classical SIS model. Regarding the other two eigenvalues ξ_2 and ξ_3, the result is more cumbersome. Despite of that, the vanishing of the 2x2 determinant of the right-inferior corner of the Jacobian matrix 8.33 yields:

$$\lambda_2 = \lambda_{2ii}^c = \frac{\lambda_1 k_o}{l_o} \frac{\eta_2 \mu_2 ((\beta_1^a (\lambda_1 k_o - \mu_1) + \mu_2) + \eta_1 \mu_1 \mu_2}{\mu_1 (\eta_1 \mu_1 + \eta_2 \mu_2) + (\lambda_1 k_o - \mu_1)(\beta_1^a \beta_2^a \beta_2^b (\lambda_1 k_o - \mu_1) + \beta_2^a \beta_2^b \mu_2 + \eta_1 \mu_1 \beta_2^a + \beta_1^a \beta_2^b \mu_1)}, \tag{8.34}$$

to which the general expression presented in the main text for the epidemic threshold reduces when $P(k, l) = \delta(k - k_o)\delta(l - l_o)$. The agreement between numerical simulations and the analytic expression of the threshold presented here is only accurate when $\lambda_1 k_o \simeq \mu_1$, and the reason is easily understandable. Let us compare the reaction of the system to the introduction of a small seed of SI individuals when both diseases are absent ($(IS, SI, II) = (0, 0, 0)$, case 1) or when the first disease was already installed in the system ($(IS, SI, II) = (\pi_1, 0, 0)$, case 2). As we have argued in the precedent sections, the epidemic threshold is different in each case, and the reason is simply the presence, in the second case, of a fraction π_1 of IS of individuals for which the infectiousness and recovery rates for disease 2 are different with respect to the rest of

individuals. As a consequence, the mean values of the dynamical parameters averaged over the whole population $\langle\lambda_2\rangle$ and $\langle\mu_2\rangle$, are different in both cases and will yield different values for the epidemic thresholds. Thus, in order to accurately evaluate the secondary threshold for disease 2 it is essential to know with enough precision the prevalence π_1 corresponding to a single SIS model, as a function of λ_1. The problem arises from the fact that, precisely, the derivation of 8.34 explicitly assumes that the bijection between π_1 and λ_1 is governed by the MF stationary expression 8.31 which is only precise when $\lambda_1 k_o \simeq \mu_1$ [75]. In fact, a way to rebuild a more accurate secondary threshold curve can be achieved if, from equation 8.31 we substitute λ_1 as a function of π_1, and then introduce the so obtained expression into 8.34 to get:

$$\lambda_{2ii}^{c}(\pi_1) = \frac{1}{l_o}\frac{\eta_2\mu_2(\beta_1^a\frac{\mu_1\pi_1}{(1-\pi_1)}+\mu_2)+\eta_1\mu_1\mu_2}{(1-\pi_1)(\eta_1\mu_1+\eta_2\mu_2)+\pi_1(\beta_1^a\beta_2^a\beta_2^b\frac{\mu_1\pi_1}{(1-\pi_1)}+\beta_2^a\beta_2^b\mu_2+\eta_1\mu_1\beta_2^a+\beta_1^a\beta_2^b\mu_1)} \tag{8.35}$$

which allows to evaluate directly the threshold as a function of π_1 rather than of λ_1. Thus, by introducing in equation 8.35 the π_1 values obtained from the simulations instead of the theoretical prediction of the MF (equation 8.31), we recover the curves for the secondary threshold represented with red lines in Figures 8.2 and 8.3 of the main text, in quantitative agreement with results from simulations.

Obviously, the same arguments stand for the secondary threshold of the first disease, which obeys the following expression:

$$\lambda_{1ii}^{c}(\pi_2) = \frac{1}{k_o}\frac{\eta_1\mu_1(\beta_2^a\frac{\mu_2\pi_2}{(1-\pi_2)}+\mu_1)+\eta_2\mu_2\mu_1}{(1-\pi_2)(\eta_2\mu_2+\eta_1\mu_1)+\pi_2(\beta_2^a\beta_1^a\beta_1^b\frac{\mu_2\pi_2}{(1-\pi_2)}+\beta_1^a\beta_1^b\mu_1+\eta_2\mu_2\beta_1^a+\beta_2^a\beta_1^b\mu_2)} \tag{8.36}$$

When the networks are heterogeneous, this reformulation of the threshold curves as a function of π_1 or π_2 can not be done straightforwardly as no analytical bijection $\lambda_1(\pi_1)$ or $\lambda_2(\pi_2)$ can be reached. The best we can do is to use a numerically built-up relationship $\lambda_1(\theta_1)$ (or $\lambda_2(\theta_2)$), and introduce it into the general expression for the threshold. Although the accuracy of the curves for the secondary threshold is very satisfactory in Fig. 8.4, we identify this effect to be the source of the slight divergence between the analytical and numerical secondary thresholds shown in Fig. 8.6.

8.7.2 Vanishing conditions for epidemic thresholds (SIS case)

Here we provide a systematic analysis of the behavior of the secondary threshold in our model for large SF networks. In particular, we inspect whether the coupling between the spreading of the two diseases in the terms described in our model can modify the classical, single-disease scheme, and, if so, under what conditions. The epidemic threshold for the first disease in our model reads as:

$$\lambda_1^c(\sigma_2) = \mu_1\frac{\langle k\rangle}{\sum_{k,l}P(k,l)\frac{k^2l^2\sigma_2^2\beta_2^a\beta_1^a\beta_1^b+lk^2\sigma_2(\eta_2\mu_2\beta_1^a+\beta_1^b(\beta_1^a\mu_1+\beta_2^a\mu_2))+k^2\mu_2(\eta_1\mu_1+\eta_2\mu_2)}{l^2\sigma_2^2\beta_2^a\eta_1+l\sigma_2(\eta_1\mu_1+\eta_2\mu_2+\beta_2^a\eta_1\mu_2)+\mu_2(\eta_1\mu_1+\eta_2\mu_2)}} \tag{8.37}$$

and we have to address its behavior in the limit $N \to \infty$. Substituting sums by integrals in Eq. (8.37), one gets:

$$\lim_{N\to\infty} \lambda_1^c(\sigma_2) = \lim_{(k_{max},l_{max})\to\infty} \mu_1 \frac{\int_{k_{min}}^{k_{max}} \int_{l_{min}}^{l_{max}} P(k,l)k \; dk \; dl}{\Delta'}. \tag{8.38}$$

where the denominator function Δ' takes the form of the following double integral:

$$\Delta' = \int_{k_{min}}^{k_{max}} \int_{l_{min}}^{l_{max}} P(k,l) \frac{k^2 l^2 \sigma_2^2 \beta_2^a \beta_1^a \beta_1^b + l k^2 \sigma_2 (\eta_2 \mu_2 \beta_1^a + \beta_1^b (\beta_1^a \mu_1 + \beta_2^a \mu_2)) + k^2 \mu_2 (\eta_1 \mu_1 + \eta_2 \mu_2)}{l^2 \sigma_2^2 \beta_2^a \eta_1 + l \sigma_2 (\eta_1 \mu_1 + \eta_2 \mu_2 + \beta_2^a \eta_1 \mu_2) + \mu_2 (\eta_1 \mu_1 + \eta_2 \mu_2)} dk \; dl \tag{8.39}$$

To study the behavior of the threshold in the thermodynamic limit, we present here an analysis which is essentially based on the following result, whose proof is an exercise of elementary algebra that we present for the sake of completeness in the last appendix. Given two polynomials $P(k)$, $Q(k)$, we have that:

$$\left| \lim_{k_{max}\to\infty} \int_{k_{min}}^{k_{max}} \frac{P(k)}{Q(k)} dk \right| = \infty \leftrightarrow deg(Q) - deg(P) \leq 1 \tag{8.40}$$

Focusing on SF connectivity distributions, we will distinguish two different scenarios in this section: uncorrelated and totally correlated layers. In the first case, both networks present a SF distribution in which the connectivity of a node in a layer is essentially independent of its degree on the other layer, in such a way that the composed connectivity distribution verifies $P(k,l) = C_o k^{-\gamma} l^{-\Gamma}$. Instead, in the second case, although both layers are also SF networks, the degree of any given node in both layers is forced to be the same. Therefore, nodes which are hubs in a layer are so in the other, and the composed degree distribution is $P(k,l) = C_o \delta(k-l) k^{-\gamma}$. In addition, we assume that both γ and Γ exponents are rational, hence, we can write $\gamma = w/x$ and $\Gamma = y/z$ with $(w,x,y,z) \in \mathbb{N}$. By addressing these two opposite scenarios, our aim is to characterize the difference, in terms of the spreading dynamics, between the coupling of two diseases that spread by independent means –which will give place to different networks of contacts– and the coupling of two related diseases –or variations of the same disease– that spread following the very same mechanisms and, thus, they do so over highly correlated networks of contacts.

Uncorrelated SF layers

If we substitute $P(k,l) = C_o k^{-\gamma} l^{-\Gamma}$ into eq. 8.38, we can factorize the double integrals into independent terms:

$$\lim_{N\to\infty} \lambda_1^c(\sigma_2) = \lim_{(k_{max},l_{max})\to\infty} \mu_1 \frac{\int_{k_{min}}^{k_{max}} k^{1-\gamma} dk \int_{l_{min}}^{l_{max}} l^{-\Gamma} dl}{\Delta''} \tag{8.41}$$

and this time the denominator Δ'' takes the following form:

$$\Delta'' = \int_{k_{min}}^{k_{max}} k^{2-\gamma} dk \int_{l_{min}}^{l_{max}} \frac{l^2 \sigma_2^2 \beta_2^a \beta_1^a \beta_1^b + l \sigma_2 (\eta_2 \mu_2 \beta_1^a + \beta_1^b (\beta_1^a \mu_1 + \beta_2^a \mu_2)) + \mu_2 (\eta_1 \mu_1 + \eta_2 \mu_2)}{l^\Gamma \left(l^2 \sigma_2^2 \beta_2^a \eta_1 + l \sigma_2 (\eta_1 \mu_1 + \eta_2 \mu_2 + \beta_2^a \eta_1 \mu_2) + \mu_2 (\eta_1 \mu_1 + \eta_2 \mu_2) \right)} dl \tag{8.42}$$

The condition for the numerator to diverge is $\gamma \leq 2 \vee \Gamma \leq 1$. These conditions are verified by networks whose mean connectivity diverges in the thermodynamic limit. For this reason, networks with those small exponents neither are reasonable systems to be used to describe any information diffusion process over them nor are found in real systems. Taking this into account, we will restrict our analysis to the more epidemiologically relevant scenario in which $\gamma > 2$ and $\Gamma > 2$, although our reasoning is generalizable to any value of the exponents.

Thus, in the scenario $\gamma > 2$ and $\Gamma > 2$, the numerator remains always finite and the only phenomenon of interest that could be found is an eventual vanishing of the threshold due to a divergence in the denominator Δ''. The double integral Δ'' can be factorized as the product of two integrals: one in k and the other in l, as it can be seen from Eq. 8.42: we will schematically refer this like follows: $\Delta'' = \Delta''_k \Delta''_l$. The integral in k, $\Delta''_k = \int_{k_{min}}^{k_{max}} k^{2-\gamma} dk$ will diverge if and only if $\gamma \leq 3$, but the integral in l, Δ''_l, might independently diverge under some conditions. That situation would suppose the vanishing of the threshold of the first disease as a consequence of its coupling to the second rather than to internal, dynamical or topological features. To find out the conditions for Δ''_l, lets make the following change of variable:

$$m = l^{1/z} \tag{8.43}$$

so as to change the dependence of the denominator $\Delta'' = \Delta''_k \Delta''_l$ in Eq. 8.42 to $\Delta'' = \Delta''_k \Delta''_m$:

$$\Delta'' = \int_{k_{min}}^{k_{max}} k^{2-\gamma} dk \int_{l_{min}^{1/z}}^{l_{max}^{1/z}} \frac{\left(m^{2z}\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + m^z\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right) z m^{z-1}}{m^y \left(m^{2z}\sigma_2^2\beta_2^a\eta_1 + m^z\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right)} dm \tag{8.44}$$

The last expression has the advantage that the argument of the integral in m, Δ''_m is just the quotient of two polynomials, and thus, its behavior in the limit $l_{max} \to \infty$ is governed by Eq. (8.40) and depends only on the difference of degrees of the denominator and the numerator. If

$$\Delta''_m = \int_{l_{min}^{1/z}}^{l_{max}^{1/z}} \frac{\left(m^{2z}\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + m^z\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right) z m^{z-1}}{m^y \left(m^{2z}\sigma_2^2\beta_2^a\eta_1 + m^z\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right)} dm = \\ = \int_{l_{min}^{1/z}}^{l_{max}^{1/z}} \frac{P(m)}{Q(m)} dm \tag{8.45}$$

we have that –in the general case in which none of the β and η parameters vanishes– $deg(Q) - deg(P) = 1 + y - z$, and hence, the conditions for Δ''_m to diverge are

$$\lim_{(k_{max}, l_{max}) \to \infty} \int_{l_{min}^{1/z}}^{l_{max}^{1/z}} \frac{P(m)}{Q(m)} dm = \infty \leftrightarrow y - z + 1 \leq 1 \leftrightarrow \Gamma = y/z \leq 1 \tag{8.46}$$

Therefore, in the region of interest $\gamma > 2 \wedge \Gamma > 2$, the factor Δ''_m can not diverge and make the threshold vanishes. It is worth noticing that the condition $\Gamma < 1$ does not guarantee the vanishing of the threshold, as it also makes the numerator to diverge. Expression 8.46 is valid for the case in which none of the β and η parameters of the model vanishes. If this is not the case, i.e., if some of the infectiousness variations β do

$(\beta_1^a, \beta_1^b, \beta_2^a)$ parameters	$deg(Q) - deg(P)$ in $\Delta_m'' = \int P(m)/Q(m)$	Divergence condition
$(\beta_1^a, \beta_1^b, \beta_2^a)$ $(\beta_1^a, \beta_1^b, 0)$ $(\beta_1^a, 0, 0)$	$y - z + 1$	$\gamma \leq 3 \vee \Gamma \leq 1$
$(0, \beta_1^b, \beta_2^a)$ $(\beta_1^a, 0, \beta_2^a)$ $(0, \beta_1^b, 0)$ $(0, 0, 0)$	$y + 1$	$\gamma \leq 3 \vee \Gamma \leq 0$
$(0, 0, \beta_2^a)$	$y + z + 1$	$\gamma \leq 3 \vee \Gamma \leq -1$

TABLE 8.6: Divergence conditions for Δ_m'' for disease 1. Case 1: uncorrelated SF layers $P(k,l) = C_o k^{-\gamma} l^{-\Gamma}$, $\gamma = (w/x)$, $\Gamma = (y/z)$.

vanish −the rest of the parameters can not do that within a realistic epidemiological framework−, the degrees of the numerator and the denominator in Eq. 8.45 could vary. In table 8.6 we systematically address all possible combinations of null parameters, and the composed conditions yielding a vanishing threshold for each case.

Once again, we see that, whatever the interacting scheme between both diseases is, the denominator remains always finite provided that $\Gamma > 2$. In conclusion, if no degree correlation is introduced between layers, and for realistic systems characterized by double power laws verifying $\gamma > 2 \wedge \Gamma > 2$, the behavior of the thresholds in the thermodynamic limit is essentially the same of the uncoupled systems: the threshold associated to the first disease vanishes if and only if the exponent of its own network verifies $\gamma \leq 3$, whatever the exponent of the second network. The coupling will introduce, in general, only a finite pre-factor. The symmetric situation obviously stands for the threshold of the second disease.

Totally correlated SF layers

If we consider the case in which $P(k,l) = C_o \delta(k-l) k^{-\gamma}$, where $\gamma = w/x$ with $(w,x) \in \mathbb{N}$, the epidemic threshold reads as:

$$\lim_{N\to\infty} \lambda_1^c(\sigma_2) = \lim_{(k_{max},l_{max})\to\infty} \mu_1 \frac{\int_{k_{min}}^{k_{max}} k^{1-\gamma} dk}{\Delta'''} \tag{8.47}$$

where the denominator Δ''' adopts now the following form:

$$\Delta''' = \int_{k_{min}}^{k_{max}} \frac{k^2\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + k\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)}{k^{\gamma-2}\left(k^2\sigma_2^2\beta_2^a\eta_1 + k\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right)} dk \tag{8.48}$$

hence, we do not recover the factorization of the denominator previously observed. Instead, after changing the variable to $m = k^{1/x}$, we transform Δ''' to express it as an

$(\beta_1^a, \beta_1^b, \beta_2^a)$ parameters	$deg(Q) - deg(P)$ in $\Delta_m''' = \int P(m)/Q(m)$	Divergence condition
$(\beta_1^a, \beta_1^b, \beta_2^a)$ $(\beta_1^a, \beta_1^b, 0)$ $(\beta_1^a, 0, 0)$	$w - 3x + 1$	$\gamma \leq 3$
$(0, \beta_1^b, \beta_2^a)$ $(\beta_1^a, 0, \beta_2^a)$ $(0, \beta_1^b, 0)$ $(0, 0, 0)$	$w - 2x + 1$	$\gamma \leq 2$
$(0, 0, \beta_2^a)$	$w - x + 1$	$\gamma \leq 1$

TABLE 8.7: Divergence conditions for Δ_m''' for disease 1. Case 2: totally correlated SF layers $P(k,l) = C_o\delta(k-l)k^{-\gamma}$, $\gamma = (w/x)$.

integral in m (in what follows Δ_m'''):

$$\Delta''' = \int_{k_{min}^{1/x}}^{k_{max}^{1/x}} \frac{\left(m^{2x}\sigma_2^2\beta_2^a\beta_1^a\beta_1^b + m^x\sigma_2(\eta_2\mu_2\beta_1^a + \beta_1^b(\beta_1^a\mu_1 + \beta_2^a\mu_2)) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right) x m^{x-1}}{m^{w-2x}\left(m^{2x}\sigma_2^2\beta_2^a\eta_1 + m^x\sigma_2(\eta_1\mu_1 + \eta_2\mu_2 + \beta_2^a\eta_1\mu_2) + \mu_2(\eta_1\mu_1 + \eta_2\mu_2)\right)} dm \tag{8.49}$$

In expression 8.47, the numerator diverges for $\gamma \leq 2$. In turn, for Δ_m''', we have that it has the following form:

$$\Delta_m''' = \int_{k_{min}^{1/x}}^{k_{max}^{1/x}} \frac{P(m)}{Q(m)} dm \tag{8.50}$$

which diverges, according to equation 8.40, if and only if $deg(Q) - deg(P) \leq 1$. In the general case in which none of the β parameters vanishes, we have that $deg(Q) - deg(P) = w - 3x + 1$, and thus:

$$\lim_{N\to\infty} \lambda_1^c(\sigma_2) = 0 \leftrightarrow \gamma \leq 3 \tag{8.51}$$

As in the previous case, different combinations of null β parameters can change this result, as it can be seen in Table 8.7. Here, the phenomenology is remarkable different for all the cases in which at least one of the two variations of disease 1 infectiousness, β_1^a or β_1^b, cancel (except for the case $\beta_2^a = \beta_1^b = 0$ but $\beta_1^a \neq 0$, for which eq. 8.51 also stands). In those cases, $\theta > 0$ guarantees that, provided that $\gamma > 2$, no threshold vanishing is observed in the thermodynamic limit, even when the exponent is within the interval $2 < \gamma \leq 3$. In the case in which $\beta_1^a = \beta_1^b = \beta_2^a = 0$, the situation will be reciprocal, and thus, the epidemic threshold of the second disease will not vanish either for $\gamma > 2$.

8.7.3 Proof of eq. 8.40

Given two polynomials $P(k)$, $Q(k)$, we have to prove that

$$\left| \lim_{k_{max}\to\infty} \int_{k_{min}}^{k_{max}} \frac{P(k)}{Q(k)} dk \right| = \infty \leftrightarrow deg(Q) - deg(P) \leq 1 \tag{8.52}$$

First, we have that, if $deg(P) \geq deg(Q)$ the integral diverges, as the argument of the integral can be expressed in that case as:

$$\frac{P(k)}{Q(k)} = C(k) + \frac{P'(k)}{Q'(k)} \tag{8.53}$$

with $deg(P') < deg(Q')$ and $deg(C) \geq 0$. When substituting the last expression into eq.8.40, the integral in $C(k)$ automatically diverges. Therefore, to prove this result for the case in which $deg(P) < deg(Q)$, let us consider the general decomposition of $Q(k)$:

$$Q(k) = Q_o \prod_{i=1}^{i^*} (k - a_i)^{n_i} \prod_{j=1}^{j^*} (k^2 + b_j k + c_j)^{m_j} \tag{8.54}$$

where Q_o is a constant, $k_{min} > 0$ and $b_j^2 - 4c_j < 0 \ \forall j$. The values i^* and j^* stand for the number of different elemental factors of first and second order, respectively. So, the degree of the polynomial reads as follows:

$$deg(Q) = \sum_{i=1}^{i^*} n_i + 2 \sum_{j=1}^{j^*} m_j, \tag{8.55}$$

where n_i and m_j denote the multiplicity of each of the factors of first and second order, respectively. The factorization of $Q(k)$ yields the following decomposition of the quotient $P(k)/Q(k)$ into partial fractions:

$$\begin{aligned} \int_{k_{min}}^{k_{max}} \frac{P(k)}{Q(k)} dk = \sum_{i=1}^{i^*} \sum_{n=1}^{n_i} \int_{k_{min}}^{k_{max}} \frac{A_{i,n}}{(k + a_i)^n} dk + \\ \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \int_{k_{min}}^{k_{max}} \frac{B_{j,m} k + C_{j,m}}{(k^2 + b_j k + c_j)^m} dk \end{aligned} \tag{8.56}$$

where the coefficients $A_{i,n}, B_{j,m}, C_{j,m}$ have to be calculated by expanding this expression into a single fraction, and comparing the coefficients of the numerator obtained with those of $P'(k) = P(k)/Q_o$. With the aim of understanding what are the conditions that make these partial integrals to diverge, let us analyze them term by term. The first sum yields logarithmic and rational functions:

$$\begin{aligned} \sum_{i=1}^{i^*} \sum_{n=1}^{n_i} \int_{k_{min}}^{k_{max}} \frac{A_{i,n}}{(k + a_i)^n} dk = \sum_{i=1}^{i^*} A_{i,1} \ln \left(\frac{k_{max} + a_i}{k_{min} + a_i} \right) + \\ + \sum_{i=1|n_i>1}^{i^*} \sum_{n=2}^{n_i} \frac{-A_{i,n}}{(n-1)} \left(\frac{1}{(k_{min} + a_i)^{n-1}} - \frac{1}{(k_{max} + a_i)^{n-1}} \right) \end{aligned} \tag{8.57}$$

and, in the limit $k_{max} \to \infty$, only the logarithmic term diverges:

$$\lim_{k_{max}\to\infty} \sum_{i=1}^{i^*} \sum_{n=1}^{n_i} \int_{k_{min}}^{k_{max}} \frac{A_{i,n}}{(k+a_i)^n} dk = \sum_{i=1}^{i^*} A_{i,1} \ln(k_{max}) + \vartheta \tag{8.58}$$

where ϑ stands for a finite term, negligible when compared to $\ln(k_{max})$. On the other hand, the second sum in eq. 8.56 can be rewritten as:

$$\begin{aligned} \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \int_{k_{min}}^{k_{max}} \frac{B_{j,m}k + C_{j,m}}{(k^2+b_jk+c_j)^m} dk = \\ \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \frac{B_{j,m}}{2} \int_{k_{min}}^{k_{max}} \frac{2k+b_j}{(k^2+b_jk+c_j)^m} dk + \\ + \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \left(C_{j,m} - \frac{B_{j,m}}{2} \right) \int_{k_{min}}^{k_{max}} \frac{dk}{(k^2+b_jk+c_j)^m} dk \end{aligned} \tag{8.59}$$

In turn, the integrals in the first sum of last equation can be easily solved:

$$\begin{aligned} \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \frac{B_{j,m}}{2} \int_{k_{min}}^{k_{max}} \frac{2k+b_j}{(k^2+b_jk+c_j)^m} dk = \\ \sum_{j=1}^{j^*} \frac{B_{j,m}}{2} \ln \frac{(k_{max}^2 + b_jk_{max} + c_j)}{(k_{min}^2 + b_jk_{min} + c_j)} + \\ + \sum_{j=1|m_j>1}^{j^*} \sum_{m=1}^{m_j} \frac{B_{j,m}}{2} \left(\frac{1}{k_{min}^2 + b_jk_{min} + c_j} - \frac{1}{k_{max}^2 + b_jk_{max} + c_j} \right) \end{aligned} \tag{8.60}$$

and, again, when taking the limit $k_{max} \to \infty$, only the logarithmic terms diverge:

$$\begin{aligned} \sum_{j=1}^{j^*} \sum_{m=1}^{m_j} \frac{B_{j,m}}{2} \int_{k_{min}}^{k_{max}} \frac{2k+b_j}{(k^2+b_jk+c_j)^m} dk = \\ \sum_{j=1}^{j^*} \frac{B_{j,m}}{2} \ln(k_{max}^2) + \vartheta = \sum_{j=1}^{j^*} B_{j,m} \ln(k_{max}) + \vartheta \end{aligned} \tag{8.61}$$

Finally, we have to analyze the last integrals in eq. 8.59. After the following linear transformation:

$$v = \frac{2k+b_j}{\sqrt{4c_j - b_j^2}} \tag{8.62}$$

we can rewrite:

$$\begin{aligned} \left(C_{j,m} - \frac{B_{j,m}}{2} \right) \int_{k_{min}}^{k_{max}} \frac{dk}{(k^2+b_jk+c_j)^m} = \\ = \left(C_{j,m} - \frac{B_{j,m}}{2} \right) \left(c_j - \left(\frac{b_j}{2} \right)^2 \right)^{1/2-m} \int_{\frac{2k_{min}+b_j}{\sqrt{4c_j-b_j^2}}}^{\frac{2k_{max}+b_j}{\sqrt{4c_j-b_j^2}}} \frac{1}{(v^2+1)^m} dv \end{aligned} \tag{8.63}$$

and, though it is not possible to write an explicit solution for the integrals of the form $\int 1/(v^2+1)^m$, integrating by parts we get:

$$\int \frac{1}{(v^2+1)^m} dv = \frac{v}{(2m-2)(v^2+1)^{m-1}} + \frac{2m-3}{2m-2} \int \frac{dv}{(v^2+1)^{m-1}} \tag{8.64}$$

which allows to solve the integrals, yielding the appearance of rational and arctangent terms none of which diverges in the limit $k_{max} \to \infty$. Thus, only logarithmic terms from equations 8.58 and 8.61 contribute to the divergence of the initial limit of Eq. 8.40, that can be finally rewritten as follows:

$$\left| \lim_{k_{max} \to \infty} \int_{k_{min}}^{k_{max}} \frac{P(k)}{Q(k)} dk \right| = \ln\left(k_{max}\right) \left| \sum_{i=1}^{i^*} A_{i,1} + \sum_{j=1}^{j^*} B_{j,m} \right| + \vartheta \tag{8.65}$$

Thus, recalling that $|\vartheta| << \ln(k_{max})$ stands for finite terms, the only requisite for the limit to diverge is that:

$$\left(\sum_{i=1}^{i^*} A_{i,1} + \sum_{j=1}^{j^*} B_{j,m} \right) \neq 0 \tag{8.66}$$

which is precisely the coefficient of the monomial of degree equal to $deg(Q) - 1$ in the numerator $P'(k) = P(k)/Q_o$, as can be easily shown after grouping the partial fractions in Eq. 8.56 into a single one. Therefore, we have demonstrated the initial statement: the condition for the integral in Eq. (8.52) to diverge is that $deg(P') = deg(P) \geq deg(Q) - 1$.

Part V
Novel models of tuberculosis spreading

One of the historic ironies of tuberculosis research is that it has always been assumed that the current interventions would eliminate this disease as a major public health problem. BCG, an attenuated bovine tuberculosis strain, was discovered in 1908, and was thought to be the vaccine for tuberculosis. Streptomycin in the 1940s was hailed as the wonder drug for tuberculosis. Yet even with better antibiotics, tuberculosis remains a major global health problem. Concomitant with these historically shortsighted miscalculations were reductions in support for research on new tools and strategies, based on the assumption that with existing interventions the disease would disappear. It has not.

Barry Bloom
First Blueprint for Tuberculosis Vaccine Development, 1998.

Chapter 9

Novel models for the description of TB spreading

9.1 Introduction

Among the vaccine candidates currently under development, there are very diverse projects that find their effectiveness on disparate biochemical and immunological principles. One of the aspects that could exert a greater influence on the vaccine impacts is the age of the vaccination target populations, which eventually could be different for some of the novel vaccines [428]. Actually, these drugs can be grouped into two classes: immunogenic booster vaccines -aimed at being administrated on individuals previously vaccinated with the bacillus Calmette-Guérin vaccine (BCG) so as to enhance its effects– and substitutive vaccines –in principle to be applied on non-vaccinated newborns instead of BCG–. Thus, in principle, booster vaccines could be implemented on massive vaccination campaigns over all ages while substitutive vaccines would be limited to newborns, which would limit and slow down their impact in terms of deaths and cases prevented [97].

However, some eventualities may modify that scheme. For example, some teams involved in the development of substitutive vaccines are outlining the possibility of testing this kind of drugs on individuals previously vaccinated with BCG, which could extent their application range from newborns to individuals of any age [421]. In addition, the first, disappointing results about immunogenic efficacy of booster vaccines on children [213] suggest that, eventually, the application of this kind of drugs may have to be restricted to older individuals. Moreover, any vaccine should demonstrate its safety and efficacy when applied on seropositive individuals: in case some problems appear related to the interaction between TB and HIV, any vaccination campaign involving adults would be seriously compromised, specially in areas of high HIV endemicity like sub-Saharan Africa.

All the uncertainties mentioned above will be solved in the following years, once the clinical trials come to an end. When that moment comes, computational tools and spreading models should be able to offer a vaccine impact forecast as reliable as possible, taking into account both the vaccine efficacy and the age of the individuals on which its application is safe, feasible and effective. The precision of these forecasting tools, at that point, may be a crucial resource for the health authorities to count with in order to make the proper decisions about the design of the final vaccination campaign.

Hence, if we want to address how the impact of a vaccine depends on the age of the vaccination target population, it is more necessary than ever a precise understanding of the dependence of the disease dynamics on the age of single individuals and on the

demographic structure of the whole population. In order to accomplish that goal, in this chapter we develop an epidemiological model of TB spreading based on previous works [422, 22] in which several classical simplifications regarding the dependence of the spreading dynamics on age structure have been eliminated, capitalizing recent works and public data sources [12, 23] that provides some of the information needed to adopt more realistic modeling hypothesis.

These new ingredients consists of contemplating the explicit evolution of the age structure of the population analyzed, as well as the heterogeneity of the contacts driving the infections among the different age groups and the possibility of that the system lies in a dynamical state arbitrarily far away from the stationary. As we will see in the following sections, the introduction of these new ingredients relevantly modifies the model outcomes, making evident that these ingredients have to be taken into account if any age-dependent measure is going to be done or any age-focused intervention has to be analyzed.

9.2 Modeling framework overview

In order to undertake the impact evaluations mentioned, we have developed an age-structure model which is initially based on previous works on this topic [422, 22]. The model essentially obeys to an HMF scheme according to which all individuals within the same age group are dynamically equivalent. in this sense, this type of age-structured models differs from those developed in chapters 8 and 7 in the implementation of an aging dynamics that allows individuals to transit among the different dynamically distinct groups (i.e. age-groups) of the model. Its natural History scheme can be checked on figure 9.1, while a detailed description, as well as the proper model uncertainty and sensitivity analysis can be found in section 9.3.

Besides the age-structure of the population, the model contemplates the different kinds of TB disease: pulmonary –either smear positive and negative– and non-pulmonary. Fast and slow progression are contemplated, as well as smear progression, exogenous reinfections, endogenous reactivations, mother-child transmission, different treatment outcomes and natural recovery [96, 422, 22, 423, 424].

Model parameters have been obtained from bibliographical sources with three exceptions: the infectiousness, the diagnosis rate and the initial distance from the stationary, which have been fitted [413] so as to reproduce incidence and mortality series registered by health systems and available at the world health organization database for TB; from 2001 to 2011 [167, 24]. Very remarkably, the diagnosis rate $d(t)$ and the infectiousness $\beta(t)$ are assumed to vary over time, so as to describe the temporal evolution of public health systems, overall socioeconomic conditions, etc, which defines at the end the response that human societies are able to offer to the disease. This time variation is obtained trough two annual rates of variation α_d and α_β that are independently fitted to the initial values $d(2001) = d_o$ and $\beta(2001) = \beta_o$. Additional details about the meaning of all parameters and the way in which they are obtained, are available at section 9.3.

In what regards the age-structure of the population, the model takes as an explicit input the expected evolution of the demographic pyramid during the period under analysis: in this way, the disease dynamics is simulated over a population whose demographic structure evolution is known a priori. This allows us to introduce in our model the demographic predictions the the UN population division [23] makes for the different countries for the next century, and so we can explicitly evaluate the influence of that evolution on the disease burden without making any simplification or assuming that the structure of the population is different from the one expected, as it is usually done in previous models [422, 22].

In a similar way, we abandon in our model other classical hypothesis of TB modeling that suppose a strong simplification of reality, i.e., the homogeneity of the contact patterns among the different age groups that composes a society. Indeed, since some years ago, the remarkable heterogeneity of the number and intensity of the contacts between different age groups is a well known fact [12, 425] that has had a powerful impact in the spreading modeling of other diseases like influenza [426, 427]. After these works, we know that contact patterns among individuals of different ages are strongly assortative, more intense in young individuals and relatively robust in different countries. Capitalizing the results of one of these works, we are able to recover a plausible and more realistic heterogeneous contact structure on top of which we can model the spreading of the disease.

These two new modeling ingredients fall upon the age structure of the population, which is crucial for the measure of any age-dependent quantity or the impact and cost-effectiveness evaluation of any epidemiological intervention focused on specific age-groups, as an eventual vaccine. As a third relevant novelty, initial conditions are not forced to lie on an attractor of the dynamics, but the initial distance to stationary is determined as an additional tunable parameter.

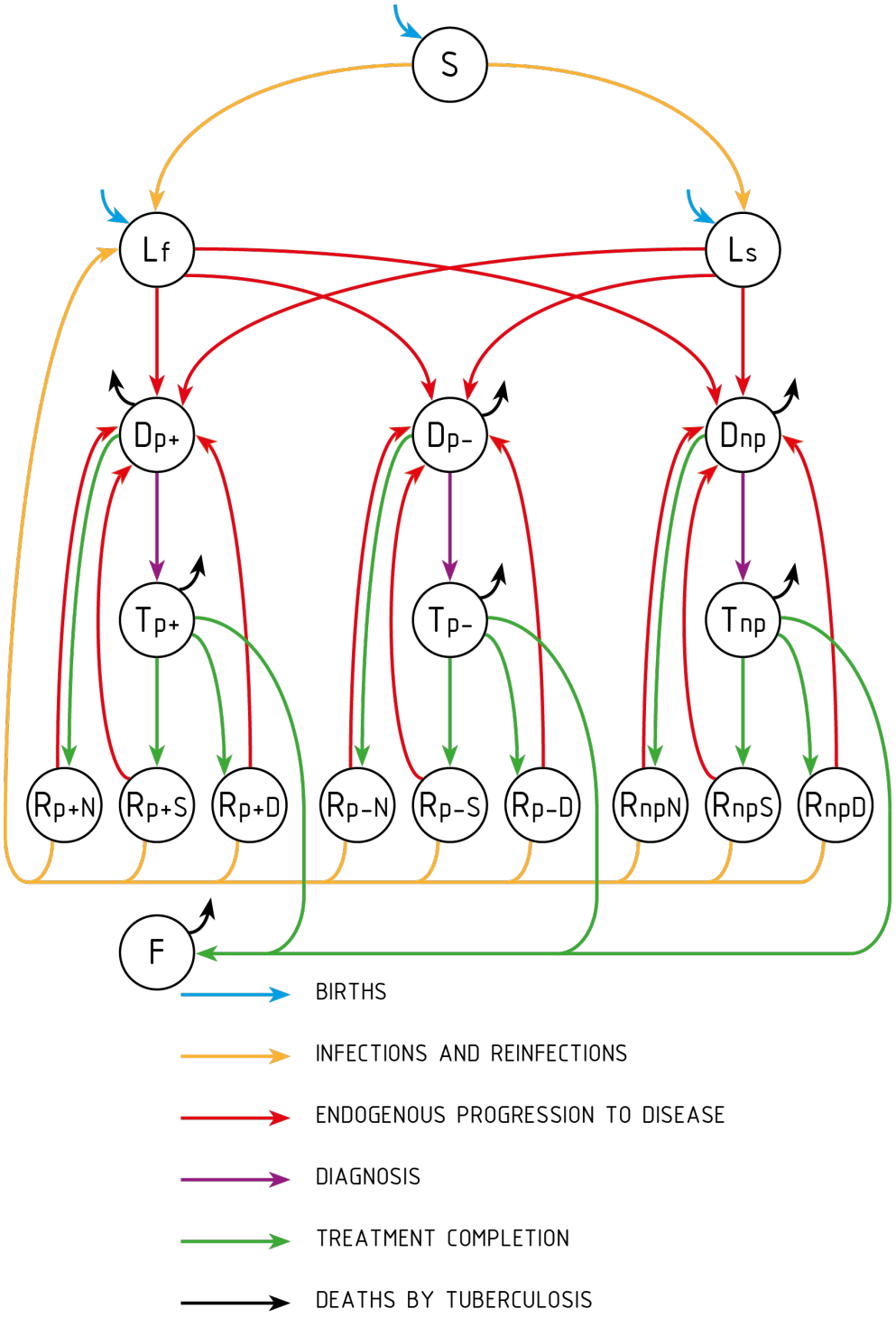


FIGURE 9.1: Schematic representation of the non immune branch of the natural History of the disease as described by our model. S: susceptible L: latent, either fast or slow. D: undiagnosed disease (either pulmonary, smear positive D_{p+}, pulmonary smear negative D_{p-} or non pulmonary D_{np}). T: treated disease. R: recovered individuals, either naturally (R_N), after successful treatment (R_S), or after treatment default (R_D). F: treatment failure. The figure represents the dynamical states and transitions of the non-immune branch of a single age group. Further details are available at the Supplementary information appendix.

9.3 Model description

9.3.1 Model dynamics

Natural history of the disease

Our work is essentially based on previous models by C. Dye and colleagues [422, 22], on which new ingredients –heterogeneous contact patterns [12] and an explicit coupling of the demographical evolution and the disease dynamics– have been incorporated in order to avoid systematic biases that affect certain model predictions and outcomes critically dependent on the age structure of the populations. The natural history scheme has also been refined so as to render it more suited to the definitions by WHO, mostly in what regards to treatment outcomes.

Summarizing, we deal with an ordinary differential equations based, age-structured model of TB infection in which we consider three different classes of unexposed individuals, –susceptible and vaccinated, either keeping immunity or having lost it–, two different latency paths to disease –fast and slow– and six different kinds of disease, depending on its etiology: -non pulmonary, pulmonary (smear positive) and pulmonary (smear negative)–, and depending on whether it is untreated or treated. After the disease phase, we consider explicitly the main treatment outcomes contemplated by WHO data schemes: treatment completion, default, failure and, of course, death.

The model presents two branches: the non vaccinated branch and the immune branch, to which individuals retaining the immunogenic effect of the vaccine belong to. Each state X in the unvaccinated branch has its homologue X^v in the immune branch. For example, at time step t, non vaccinated susceptible individuals, unexposed to TB infection are denoted by $S(a,t)$, while vaccinated, immune susceptible individuals are $S^v(a,t)$. Finally, susceptible vaccinated individuals whose induced immunity to the disease due to vaccination has waned are $S^w(a,t)$. The integer $a \in [0,13]$, is the index representing any of the fourteen age groups which individuals belong to, each of them covering $\tau = 5$ years; which yields the description of a demographic pyramid up to 70 years old.

In the following, we detail the natural history ingredients and transitions between states that we have considered to build up our model; whose natural history is schematized in figure 9.2

Primary Tuberculosis infection We call primary the infection of an individual who was not previously exposed to the bacterium: i.e. individuals of classes S,S^v and S^w. If we denote the force of infection $\lambda(a,t)$ as the probability per unit time of any unexposed individual of age group a of being infected, then, for example, the total number of unvaccinated susceptible individuals getting infected per unit time will be equal to $\lambda(a,t)S(a,t)$. We will address the explicit form of $\lambda(a,t)$ in the following sections.

Of these newly infected individuals, a fraction $p(a) \in [0,1]$ will develop the so-called primo-infection, i.e. a quick development of the disease after a short course latency

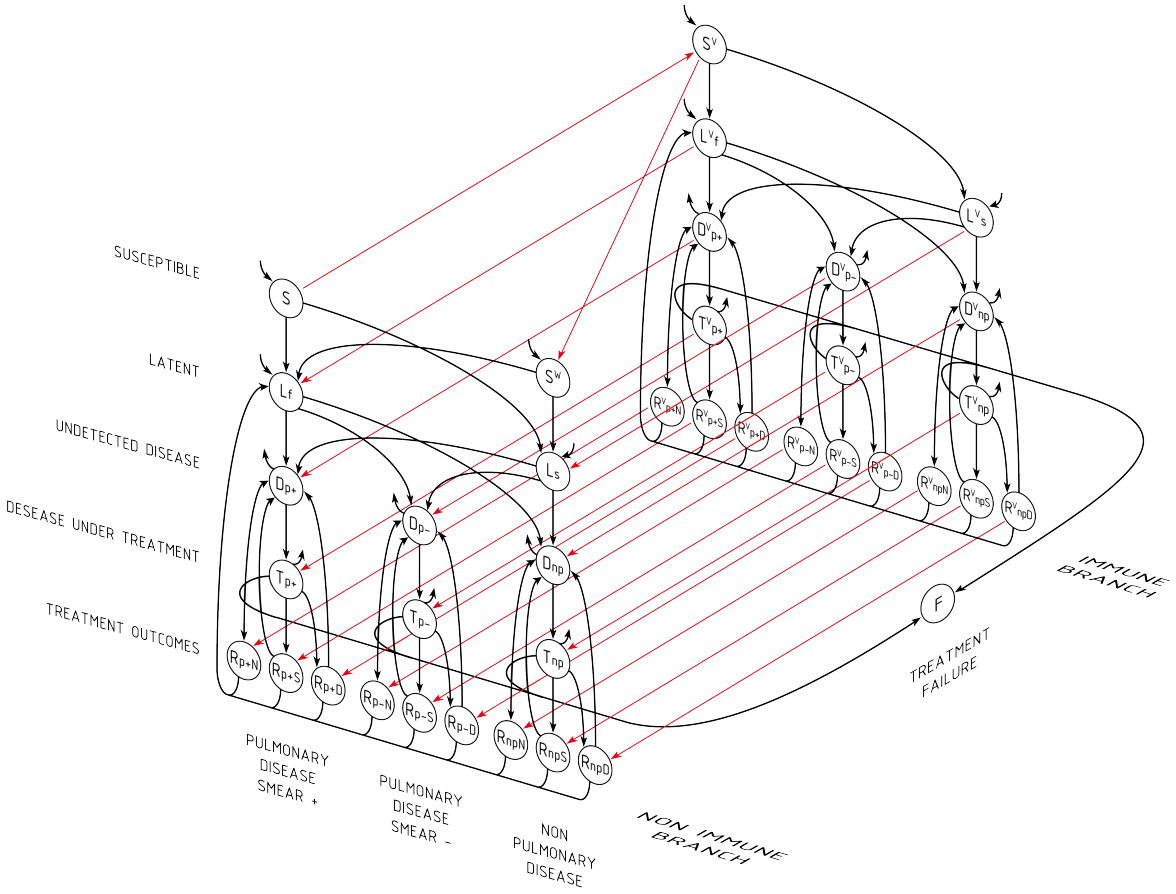


FIGURE 9.2: Global scheme of the TB spreading model: S: susceptible. L: latent. D: (untreated) disease, T (treated) disease, R recovered, F: failed recovery W: waned immunity.

period (fast latency L_f in what follows) shorter than a year and characterized by the inability of the host's immune system to restrain mycobacterial growth. In the rest of the cases, newly infected individuals' immune system succeeds at containing bacterial proliferation so establishing a host-pathogen dynamic equilibrium that is characterized by a asymptomatic latency state –slow latency L_s in what follows– that can last for the rest of the host's life, or be broken even decades after the infection, typically after an episode of immune-depression. In conclusion, the primary infection of unvaccinated individuals is described as follows:

- Primary infection of unvaccinated individuals (primo-infection): transition from $S(a,t)$ to $L_f(a,t)$: $p(a)\lambda(a,t)S(a,t)$ individuals/unit time.
- Primary infection of unvaccinated individuals (to slow latency): transition from $S(a,t)$ to $L_s(a,t)$: $(1-p(a))\,\lambda(a,t)S(a,t)$ individuals/unit time.

In what regards the infection of susceptible, immune individuals S^v, the vaccine benefits consist in reducing the probability of an individual to get infected (i.e. the force

of infection). We model this as the only effect provided by vaccination, by introducing the *observed efficacy coefficient* $\epsilon \in [0, 1]$ of the vaccine, which stands for the fraction of the probability of infection that an immunized individual presents with respect to a non vaccinated subject. Therefore, the transitions describing primary infections of immune individuals in age group a are described like this:

- Primary infection of immune individuals (primo-infection): transition from $S^v(a,t)$ to $L_f^v(a,t)$: $\epsilon p(a)\lambda(a,t)S^v(a,t)$ individuals/unit time.
- Primary infection of immune individuals (to slow latency): transition from $S^v(a,t)$ to $L_s^v(a,t)$: $\epsilon\left(1-p(a)\right)\lambda(a,t)S^v(a,t)$ individuals/unit time.

At this point, it is worth noticing that our model aims to evaluate the impacts generated by novel TB vaccines. This implies that our non-immunized branch corresponds to the population group not receiving the novel vaccine, and so, BCG immunized people (who are the vast majority in most countries in the world, in which BCG vaccination is mandatory) belong to this branch. This implies that BCG is not explicitly considered in our model, and its influence is embedded as a background effect in the dynamics of the non-immunized branch. Thus, ϵ reflects the protection provided by the novel vaccine that is observed in a comparison to BCG. -induced protection levels at each age. In chapter 11 we deepen in the implications of this definition, and, remarkably, how ϵ is affected as a function of the background performance of BCG.

Finally, regarding the class S^w, we consider that it is dynamically equivalent to class S, except for the fact that individuals in S^w can not be vaccinated again. So, we have:

- Primary infection of vaccinated individuals who have lost immunity (primo-infection): transition from $S^w(a,t)$ to $L_f(a,t)$: $p(a)\lambda(a,t)S^w(a,t)$ individuals/unit time.
- Primary infection of vaccinated individuals who have lost immunity (to slow latency): transition from $S^w(a,t)$ to $L_s(a,t)$: $(1-p(a))\lambda(a,t)S^w(a,t)$ individuals/unit time.

It worths remarking that individuals in latency classes do not have TB disease: they do not develop any disease symptom and they are not infectious at all. Indeed, as we will see in the following sections, they can even suffer ulterior re-infections (to which, for example, individuals in L^v classes will be protected with respect to individuals in L classes, as we will see, being this the reason because of which L^v classes must be differentiated from L classes).

Progression from latency to (untreated) disease Either from fast or slow latency, infected individuals can fall sick and so, progress to three different active forms of the disease. In the first of these forms, the non-pulmonary disease D_{np}, the pathogen can grow in disparate parts of the host body, including the nervous system, bones, kidneys and other organs foreign to lungs. The main characteristic of this kind of TB is

that, since the bacilli can not reach the respiratory tract, the individuals are considered unable to transmit the disease. However, if the pathogens proliferate in the lungs, they can eventually reach the upper respiratory tract making its host able to transmit the disease. According to the presence of viable bacilli in the sputum, we have the other variants of TB: pulmonary disease, smear negative D_{p-}, or pulmonary disease smear positive D_{p+}; being the latter more infectious than the former.

This scheme allows six different transitions from the two latency classes to the three untreated TB disease classes:

- Progression from $L_f(a,t)$ to $D_{np}(a,t)$: $\omega_{fnp}(a)L_f(a,t)$ individuals/unit time.
- Progression from $L_f(a,t)$ to $D_{p-}(a,t)$: $\omega_{fp+}(a)L_f(a,t)$ individuals/unit time.
- Progression from $L_f(a,t)$ to $D_{p+}(a,t)$: $\omega_{fp-}(a)L_f(a,t)$ individuals/unit time.
- Progression from $L_s(a,t)$ to $D_{np}(a,t)$: $\omega_{snp}(a)L_s(a,t)$ individuals/unit time.
- Progression from $L_s(a,t)$ to $D_{p-}(a,t)$: $\omega_{sp-}(a)L_s(a,t)$ individuals/unit time.
- Progression from $L_s(a,t)$ to $D_{p+}(a,t)$: $\omega_{sp+}(a)L_s(a,t)$ individuals/unit time.

where the ω parameters represent the different rates at which the progressions to disease take place.

In the immune branch, the transitions are the same, not influenced by the vaccine; which only protects from infection:

- Progression from $L_f^v(a,t)$ to $D_{np}^v(a,t)$: $\omega_{fnp}(a)L_f^v(a,t)$ individuals/unit time.
- Progression from $L_f^v(a,t)$ to $D_{p-}^v(a,t)$: $\omega_{fp+}(a)L_f^v(a,t)$ individuals/unit time.
- Progression from $L_f^v(a,t)$ to $D_{p+}^v(a,t)$: $\omega_{fp-}(a)L_f^v(a,t)$ individuals/unit time.
- Progression from $L_s^v(a,t)$ to $D_{np}^v(a,t)$: $\omega_{snp}(a)L_s^v(a,t)$ individuals/unit time.
- Progression from $L_s^v(a,t)$ to $D_{p-}^v(a,t)$: $\omega_{sp-}(a)L_s^v(a,t)$ individuals/unit time.
- Progression from $L_s^v(a,t)$ to $D_{p+}^v(a,t)$: $\omega_{sp+}(a)L_s^v(a,t)$ individuals/unit time.

These progression rates, as well as their respective confidence intervals have been obtained from the product of the overall rates of progression from latency to whatever kind of disease and the fractions of individuals that develop each type of disease, as available at [22].

Death of untreated individuals Individuals in D states suffer the effects of the disease by three ways: they develop disease symptoms; they –except the individuals in D_{np}– infect other individuals and some of them die because of the disease. In the model, we consider that any of the three kinds of disease has a specific mortality rate, so deaths of D individuals are modeled by introducing three independent fluxes:

- Deaths of untreated non pulmonary disease: $\mu_{np} D_{np}(a,t)$ individuals/unit time.
- Deaths of untreated smear negative pulmonary disease: $\mu_{p-} D_{p-}(a,t)$ individuals/unit time.
- Deaths of untreated smear positive pulmonary disease: $\mu_{p+} D_{p+}(a,t)$ individuals/unit time.

where μ_{np}, μ_{p-} and μ_{p+} are the TB-related death rates of D_{np}, D_{p-} and D_{p+} individuals, respectively.

Individuals who got infected despite being immune and develop disease can also die because of TB; as likely as anyone else:

- Deaths of untreated non pulmonary disease (immune branch): $\mu_{np} D^v_{np}(a,t)$ individuals/unit time.
- Deaths of untreated smear negative pulmonary disease (immune branch): $\mu_{p-} D^v_{p-}(a,t)$ individuals/unit time.
- Deaths of untreated smear positive pulmonary disease (immune branch): $\mu_{p+} D^v_{p+}(a,t)$ individuals/unit time.

Tb diagnosis and treatment For what regards our model, we consider that an individual belongs to D classes until she receives his diagnosis, moment in which it joins the corresponding treated TB class T. That corresponds to the following set of three transitions:

- Diagnosis of non pulmonar TB: transition from $D_{np}(a,t)$ to $T_{np}(a,t)$: $\eta d(t) D_{np}(a,t)$ individuals/unit time
- Diagnosis of smear negative pulmonar TB: transition from $D_{p-}(a,t)$ to $T_{p-}(a,t)$: $\eta d(t) D_{p-}(a,t)$ individuals/unit time
- Diagnosis of smear positive pulmonar TB: transition from $D_{p+}(a,t)$ to $T_{p+}(a,t)$: $d(t) D_{p+}(a,t)$ individuals/unit time

This means that the diagnosis rate $d(t)$ is the inverse of the "mean life time" of D_{p+}, mentioned before. This corresponds, essentially, to the time that sick individuals pass ignoring that they have TB disease, either because they suffer symptoms but they have not visited any doctor yet, or because, after recurring to medical services, a TB diagnosis has not been yet provided. This diagnosis refractory times are region

Regions	χ_{p+}	χ_{p-}	η
AFRH	0.51	0.43	0.843 (0.717-0.97 c.i.)
AFRL	0.49	0.25	0.510 (0.434-0.587 c.i.)
EMR	0.45	0.53	1.178 (1.001-1.354 c.i.)
SEAR	0.64	0.51	0.797 (0.677-0.916 c.i.)
WPR	0.78	0.50	0.641 (0.545-0.737 c.i.)

TABLE 9.1: Relative variation of the diagnosis rate for different types of TB in each region.

dependent, as they depend, among other factors, on the agility of the diagnosis services of public health systems. Indeed, it is a well known fact that, in countries with lower TB burden, diagnosis times tend to be higher, because of that pulmonary TB tends to be confounded, at its early stages, with more venial respiratory pathologies.

In the other hand, this time needed for TB diagnosis, is known to vary depending the type of disease, essentially because the diagnosis tools used in each type are different too. In our model η represents the variation for the diagnosis rate that is observed for the detection and diagnosis of non smear positives types of disease.

Our estimations for the parameter η are based upon the case detection ratios χ for each type of disease –D_{p+}, D_{p-} and D_{np}– reported in [22]. The case detection ratio is commonly defined as the ratio of the number of notified cases of TB to the number of incident TB cases in a given year. In [22], estimations for the case detection ratios are provided for each type of disease and regions: χ_{p+}, χ_{p-} and χ_{np}; and it turns out that according to that source $\chi_{np} \simeq \chi_{p-}$ in all regions. Therefore, if we compare the case detection ratios of non smear positive and smear positive types of the disease we can obtain an estimation for the parameter η for each region:

$$\eta = \frac{\chi_{p+}}{\chi_{p-}} (\simeq \frac{\chi_{p+}}{\chi_{np}}) \tag{9.1}$$

In table 9.3.1, the values of η so calculated are listed for the different regions under analysis. The errors have been estimated by considering a 15% as the typical uncertainty of both χ_{p+} and χ_{p-}[22], and propagating the error from there:

It is noticeable that the diagnosis rate is allowed to vary in time, as it has been done in other previous models. By doing this, case detection rates are allowed to vary independently to the total volume of classes T and the incidence rates. This is achieved through the introduction of an exponential variation term: $d(t) = d_o e^{(\alpha_d t)}$, as we will explain in detail in the subsection devoted to the fitting scheme.

Once again, the transitions are the same in the immune branch:

- Diagnosis of non pulmonary TB (immune branch): from $D^v_{np}(a,t)$ to $T^v_{np}(a,t)$: $\eta d(t) D^v_{np}(a,t)$ individuals/unit time

- Diagnosis of smear negative pulmonary TB (immune branch): from $D^v_{p-}(a,t)$ to $T^v_{p-}(a,t)$: $\eta d(t) D^v_{p-}(a,t)$ individuals/unit time

- Diagnosis of smear positive pulmonary TB (immune branch): from $D^v_{p+}(a,t)$ to $T^v_{p+}(a,t)$: $d(t)D^v_{p+}(a,t)$ individuals/unit time

Treatment outcomes Right after diagnosis, and supposing that anti-biotic treatments are available immediately, sick individuals start their treatment. In terms of our model, individuals under current treatment lie into T_{np}, T_{p-} or T_{p+}, depending on the type of disease they receive treatment to be cured from. During their stage at T classes, either by the effect of treatment or by the common quarantine measures that use to follow a TB diagnosis, individuals are not considered able to spread the disease.

Typical anti-biotic series last six months; let Ψ be the rate associated to that treatment time. Once the treatment is completed, different results are possible, and the World Health Organization groups these treatment outcomes into four main groups:

- Success: the treatment has been completed and bacilli are not present in the sputum.
- Default: the treatment has been abandoned before completion.
- Death.
- Failure: bacilli persist -or appear- in the sputum at the end of the treatment (month five or later).

plus an additional treatment outcome for individuals whose treatment outcome is not well known, either because they have been geographically transferred during the treatment ("transferred") or because their treatment outcome has not been properly evaluated ("not evaluated").

Therefore, let us denote as f_S^{p+}, f_D^{p+}, f_F^{p+} and f_μ^{p+}, the normalized fraction of pulmonary, smear positive TB sick individuals who finish their treatments belonging respectively to success, default, failure and death groups, as they are available, on a yearly basis, at [24], once normalized so as to discard the unknown outcomes to get $f_S^{p+} + f_D^{p+} + f_F^{p+} + f_\mu^{p+} = 1$ what allows us to substitute $f_S^{p+} = 1 - (f_D^{p+} + f_F^{p+} + f_\mu^{p+})$ so as to work just with these three fractions of unsuccessful treatment outcomes. For pulmonary, smear positive and non pulmonary TB cases, [24] does not differentiate the fractions of treatment outcomes, and so we have f_S^{p-}, f_D^{p-}, f_F^{p-} and f_μ^{p-} standing for the fraction of individuals undertaking each outcome both from pulmonary, smear negative and from non pulmonary classes of TB. Again, we have the closure relationship $f_S^{p-} + f_D^{p-} + f_F^{p-} + f_\mu^{p-} = 1$ that yields the substitution $f_S^{p-} = 1 - (f_D^{p-} + f_F^{p-} + f_\mu^{p})$. The values of the fractions of non successful outcomes have been averaged during the fitting time window (from 2000 to 2011), their values are provided in table 9.3.3, where confidence intervals correspond to two typical deviations of the time average of each parameter.

Therefore, we can enumerate all the possible treatment outcomes from all the different kinds of unvaccinated patients to get:

- Early treatment abandon (default) of smear positive TB: transition from $T_{p+}(a,t)$ to $R_{p+D}(a,t)$: $\Psi f_D^{p+} T_{p+}(a,t)$ individuals/unit time.
- Failed treatment completion of smear positive TB: transition from $T_{p+}(a,t)$ to $F(a,t)$: $\Psi f_F^{p+} T_{p+}(a,t)$ individuals/unit time.
- Death during treatment of smear positive TB: $\Psi f_\mu^{p+} T_{p+}(a,t)$ individuals/unit time.
- Successful treatment completion of smear positive TB: transition from $T_{p+}(a,t)$ to $R_{p+S}(a,t)$: $\Psi(1 - f_D^{p+} - f_F^{p+} - f_\mu^{p+}) T_{p+}(a,t)$ individuals/unit time.
- Early treatment abandon (default) of smear negative TB: transition from $T_{p-}(a,t)$ to $R_{p-D}(a,t)$: $\Psi f_D^{p-} T_{p-}(a,t)$ individuals/unit time.
- Failed treatment completion of smear negative TB: transition from $T_{p-}(a,t)$ to $F(a,t)$: $\Psi f_F^{p-} T_{p-}(a,t)$ individuals/unit time.
- Death during treatment of smear negative TB: $\Psi f_\mu^{p-} T_{p-}(a,t)$ individuals/unit time.
- Successful treatment completion of smear negative TB: transition from $T_{p-}(a,t)$ to $R_{p-S}(a,t)$: $\Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-}) T_{p-}(a,t)$ individuals/unit time.
- Early treatment abandon (default) of non pulmonary TB: transition from $T_{np}(a,t)$ to $R_{npD}(a,t)$: $\Psi f_D^{p-} T_{np}(a,t)$ individuals/unit time.
- Failed treatment completion of non pulmonary TB: transition from $T_{np}(a,t)$ to $F(a,t)$: $\Psi f_F^{p-} T_{np}(a,t)$ individuals/unit time.
- Death during treatment of non pulmonary TB: $\Psi f_\mu^{p-} T_{np}(a,t)$ individuals/unit time.
- Successful treatment completion of non pulmonary TB: transition from $T_{np}(a,t)$ to $R_{npS}(a,t)$: $\Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-}) T_{np}(a,t)$ individuals/unit time.

where the different R_{xy} variables stand for the groups of individuals that have completed their treatment for disease of type x (pulmonary smear positive $p+$ or negative, $p-$ or non-pulmonary np) with an outcome denoted by y (Success, S, default D, fail F or death μ).

Immunity, once the infection (and the progression to disease and the diagnosis) took place, does not affect the treatment outcomes transitions, which are equivalent in both branches of the model:

- Early treatment abandon (default) of smear positive TB: transition from $T_{p+}^v(a,t)$ to $R_{p+D}^v(a,t)$: $\Psi f_D^{p+} T_{p+}^v(a,t)$ individuals/unit time.
- Failed treatment completion of smear positive TB: transition from $T_{p+}^v(a,t)$ to $F(a,t)$: $\Psi f_F^{p+} T_{p+}^v(a,t)$ individuals/unit time.

- Death during treatment of smear positive TB:
  $\Psi f_{\mu}^{p+} T_{p+}^{v}(a,t)$ individuals/unit time.
- Successful treatment completion of smear positive TB: transition from $T_{p+}^{v}(a,t)$ to $R_{p+S}^{v}(a,t)$: $\Psi(1-f_{D}^{p+}-f_{F}^{p+}-f_{\mu}^{p+})T_{p+}^{v}(a,t)$ individuals/unit time.
- Early treatment abandon (default) of smear negative TB: transition from $T_{p-}^{v}(a,t)$ to $R_{p-D}^{v}(a,t)$: $\Psi f_{D}^{p-} T_{p-}^{v}(a,t)$ individuals/unit time.
- Failed treatment completion of smear negative TB: transition from $T_{p-}^{v}(a,t)$ to $F(a,t)$: $\Psi f_{F}^{p-} T_{p-}^{v}(a,t)$ individuals/unit time.
- Death during treatment of smear negative TB:
  $\Psi f_{\mu}^{p-} T_{p-}^{v}(a,t)$ individuals/unit time.
- Successful treatment completion of smear negative TB: transition from $T_{p-}^{v}(a,t)$ to $R_{p-S}^{v}(a,t)$: $\Psi(1-f_{D}^{p-}-f_{F}^{p-}-f_{\mu}^{p-})T_{p-}^{v}(a,t)$ individuals/unit time.
- Early treatment abandon (default) of non pulmonary TB: transition from $T_{np}^{v}(a,t)$ to $R_{npD}^{v}(a,t)$: $\Psi f_{D}^{p-} T_{np}^{v}(a,t)$ individuals/unit time.
- Failed treatment completion of non pulmonary TB: transition from $T_{np}^{v}(a,t)$ to $F(a,t)$: $\Psi f_{F}^{p-} T_{np}^{v}(a,t)$ individuals/unit time.
- Death during treatment of non pulmonary TB:
  $\Psi f_{\mu}^{p-} T_{np}^{v}(a,t)$ individuals/unit time.
- Successful treatment completion of non pulmonary TB: transition from $T_{np}^{v}(a,t)$ to $R_{npS}^{v}(a,t)$: $\Psi(1-f_{D}^{p-}-f_{F}^{p-}f_{\mu}^{p-})T_{np}^{v}(a,t)$ individuals/unit time.

It is worth noticing that recovered individuals are still susceptible to suffer further reinfection, from which, as it happens with L individuals, a vaccine might protect them. Because of this reason, it is necessary to differentiate R classes from R^{v} classes; as well as D and T classes from D^{v} and T^{v}.

Natural recovery In certain occasions, natural recovery from TB is possible without medical intervention or treatment. This is modeled by introducing three new classes of naturally recovered individuals in the first branch: $R_{npN}(a,t)$, $R_{p-N}(a,t)$ and $R_{p+N}(a,t)$, which undiagnosed, sick individuals of each type of TB join after natural recovery as follows:

- Natural recovery of non pulmonary TB: transition from $D_{np}(a,t)$ to $R_{npN}(a,t)$: $\nu D_{np}(a,t)$ individuals/unit time.
- Natural recovery of smear negative pulmonary TB: transition from $D_{p-}(a,t)$ to $R_{p-N}(a,t)$: $\nu D_{p-}(a,t)$ individuals/unit time.

- Natural recovery of smear positive pulmonary TB: transition from $D_{p+}(a,t)$ to $R_{p+N}(a,t)$: $\nu D_{p+}(a,t)$ individuals/unit time.

Once again, the same stands for the immune branch:

- Natural recovery of non pulmonary TB: transition from $D^v_{np}(a,t)$ to $R^v_{npN}(a,t)$: $\nu D^v_{np}(a,t)$ individuals/unit time.
- Natural recovery of smear negative pulmonary TB: transition from $D^v_{p-}(a,t)$ to $R^v_{p-N}(a,t)$: $\nu D^v_{p-}(a,t)$ individuals/unit time.
- Natural recovery of smear positive pulmonary TB: transition from $D^v_{p+}(a,t)$ to $R^v_{p+N}(a,t)$: $\nu D^v_{p+}(a,t)$ individuals/unit time.

Endogenous reactivations after treatment or natural recovery Nonetheless, naturally recovered individuals may experience an endogenous reactivation of the disease, since disease recovery does not suppose the total elimination of the bacilli from the host organism, generally speaking. If we denote as r_N the endogenous relapse rate of naturally recovered individuals we have:

- Endogenous reactivation of non pulmonary TB after natural recovery: transition from $R_{npN}(a,t)$ to $D_{np}(a,t)$: $r_N R_{npN}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after natural recovery: transition from $R_{p-N}(a,t)$ to $D_{p-}(a,t)$: $r_N R_{p-N}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after natural recovery: transition from $R_{p+N}(a,t)$ to $D_{p+}(a,t)$: $r_N R_{p+N}(a,t)$ individuals/unit time.

In the other hand, endogenous relapse is also possible after anti-biotic treatment. Once the treatment has finished the probabilities of experiencing an endogenous reactivation of the disease are related to the treatment outcome of the initial disease episode.

Individuals who have experienced a failed treatment ($F(a,t)$ class), regardless the type of TB they originally had are considered as infectious as smear positive untreated individuals, (because they present viable bacilli in the sputum at the end of the treatment), and their mortality risk due to TB is also the same of an smear positive untreated individual. In consequence, eventual relapses from F to D_{p+} class do not make sense, and they experiences a TB-related mortality described by $\mu_{p+}F(a,t)$ individuals/unit time. Within our modelling framework, ulterior re-diagnosis, re-infections or re-treatments for $F(a,t)$ individuals are not considered, and so, once an individual joins this class, its dynamics does not depend any more on the type of disease she previously had or on whether she was vaccinated or not.

Instead, recovered individuals after successful completion of treatment are considered functionally cured (i.e. they neither present an specific mortality risk due to TB nor are infectious). However, they may undergo ulterior endogenous reactivations of

the disease, caused by the proliferation of the very same bacilli of the original episode, not completely eliminated from the host organism. In that case, we have the following transitions:

- Endogenous reactivation of non pulmonary TB after successful treatment: transition from $R_{npS}(a,t)$ to $D_{np}(a,t)$: $r_S R_{npS}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after successful treatment: transition from $R_{p-S}(a,t)$ to $D_{p-}(a,t)$: $r_S R_{p-S}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after successful treatment: transition from $R_{p+S}(a,t)$ to $D_{p+}(a,t)$: $r_S R_{p+S}(a,t)$ individuals/unit time.

Where r_S is the endogenous relapse rate after successful treatment completion. In what regards its estimation, there exist many epidemiological studies based on the surveillance of cohorts of TB patients after treatment completion during defined follow up periods, which are aimed at determining the relapse rates, as well as the main risk factors associated to its increasing.

In the exhaustive meta-analysis by Korenrop and colleagues [96], an ensemble of such studies is considered. In that work, it is reported that, in all the works reanalized, an average of 4.2% $(3.1 - 5.3$ c.i.) of HIV uninfected subjects have a TB relapse episode during the follow up period of the study, of which, the 73% $(63 - 91$ c.i.) is due to endogenous reactivation. This means that the fraction of population that do not develop a relapse is the 96.77% $(95.73 - 97.80)$.

Another relevant result of the meta-analysis is the finding that, for HIV uninfected TB patients, the risk for TB relapse after treatment decreases with time. This can be seen from the fact that the relapse rates calculated in the different studies considered tend to be lower as the follow-up period of the trials is higher. This would imply that most of patients that experiments a relapse after treatment, do it within the first years after the initial episode.

This second result allows us to assume that the risk of developing a relapse during the follow up period of an epidemic surveillance study ($100 - 97.77 = 2.23\%$ of the population) can be associated to the total risk of developing such relapse during the entire life of an individual. So, our task is to calculate an annual risk of relapse such as, when applied over the whole period of life expectancy of a recovered individual, yields the same 2.23% of relapse cases. In order to do so, we estimate that the average life expectancy of individuals within classes R is equal to 35 years, estimation to which we come after assuming that infection and further recovery are events that occur uniformly in all ages.

Therefore, and as we are assuming that the relapse rate is constant in age and time, we have an exponential decay describing the relapse of R_{xS} individuals (R_{p+S}, R_{p-S} or R_{npS}) of the form: $R_{xS}(t) \sim e^{-r_S t}$. Therefore, after a period $t = 35$ years assimilable to the average life expectancy of a individual in R, from an initial fraction of $R_{xS} = 1$, there remains $R_{xS}(t = 35) \sim e^{-r_S 35} = 0.9677$. This calculation yields the actual value of r_S used in this chapter, $r_S = 9.4 \cdot 10^{-4}$ 1/year $(6.4 \cdot 10^{-4} - 1.3 \cdot 10^{-3})$. The confidence

interval of r_S has been obtained after the propagation of the fraction of not relapsing population as the main source of uncertainty.

Finally, recovered individuals after treatment default are considered partially infectious, although it is assumed that they do not have an explicit mortality risk due to TB. However, their endogenous relapse risk is higher, which can be modeled by introducing a parameter $r_D > r_S$ as follows:

- Endogenous reactivation of non pulmonary TB after treatment default: transition from $R_{npD}(a,t)$ to $D_{np}(a,t)$: $r_D R_{npD}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after treatment default: transition from $R_{p-D}(a,t)$ to $D_{p-}(a,t)$: $r_D R_{p-D}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after treatment default: transition from $R_{p+D}(a,t)$ to $D_{p+}(a,t)$: $r_D R_{p+D}(a,t)$ individuals/unit time.

r_D stands for the endogenous relapse rate after treatment default, which has been calculated as the product of r_S and the relative risk factor for endogenous relapse related to treatment noncompliance, –4.02 (1.79-9.01 c.i.)– taken from [423], which yields the final value of $r_D = 3.8 \cdot 10^{-3}$ 1/year ($1.4 \cdot 10^{-3} - 8.6 \cdot 10^{-3}$).

Finally, endogenous reactivation of disease transitions in the immune branch are equivalent to those of the unvaccinated branch:

- Endogenous reactivation of non pulmonary TB after natural recovery (immune branch): from $R^v_{npN}(a,t)$ to $D^v_{np}(a,t)$: $r_N R^v_{npN}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after natural recovery (immune branch): from $R^v_{p-N}(a,t)$ to $D^v_{p-}(a,t)$: $r_N R^v_{p-N}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after natural recovery (immune branch): from $R^v_{p+N}(a,t)$ to $D^v_{p+}(a,t)$: $r_N R^v_{p+N}(a,t)$ individuals/unit time.
- Endogenous reactivation of non pulmonary TB after successful treatment (immune branch): from $R^v_{npS}(a,t)$ to $D^v_{np}(a,t)$: $r_S R^v_{npS}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after successful treatment (immune branch): from $R^v_{p-S}(a,t)$ to $D^v_{p-}(a,t)$: $r_S R^v_{p-S}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after successful treatment (immune branch): from $R^v_{p+S}(a,t)$ to $D^v_{p+}(a,t)$: $r_S R^v_{p+S}(a,t)$ individuals/unit time.
- Endogenous reactivation of non pulmonary TB after treatment default (immune branch): from $R^v_{npD}(a,t)$ to $D^v_{np}(a,t)$: $r_D R^v_{npD}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear negative TB after treatment default (immune branch): from $R^v_{p-D}(a,t)$ to $D^v_{p-}(a,t)$: $r_D R^v_{p-D}(a,t)$ individuals/unit time.
- Endogenous reactivation of smear positive TB after treatment default (immune branch): from $R^v_{p+D}(a,t)$ to $D^v_{p+}(a,t)$: $r_D R^v_{p+D}(a,t)$ individuals/unit time.

Exogenous reinfection of infected individuals Individuals belonging to classes L_s and R classes have been previously exposed to TB bacilli, although they are not sick while within these classes. In addition, their rates of progression to disease due to eventual endogenous reactivations are slower than the rates ω_{fnp}, ω_{fp-} an ω_{fp+} of fast progression to disease from L_f. For those reasons, an eventual exogenous re-infection of an individual in classes L_s, R or N may cause faster progression to disease, if primo-infection takes place, than endogenous reactivation. This can be modeled by introducing the following transitions:

- Exogenous re-infection of $L_s(a,t)$ individuals yielding primo-infection: from $L_s(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)L_s(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{npN}(a,t)$ individuals yielding primo-infection: from $R_{npN}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{npN}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p-N}(a,t)$ individuals yielding primo-infection: from $R_{p-N}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p-N}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p+N}(a,t)$ individuals yielding primo-infection: from $R_{p+N}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p+N}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{npS}(a,t)$ individuals yielding primo-infection: from $R_{npS}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{npS}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p-S}(a,t)$ individuals yielding primo-infection: from $R_{p-S}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p-S}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p+S}(a,t)$ individuals yielding primo-infection: from $R_{p+S}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p+S}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{npD}(a,t)$ individuals yielding primo-infection: from $R_{npD}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{npD}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p-D}(a,t)$ individuals yielding primo-infection: from $R_{p-D}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p-D}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R_{p+D}(a,t)$ individuals yielding primo-infection: from $R_{p+D}(a,t)$ to $L_f(a,t)$: $p(a)q(a)\lambda(a,t)R_{p+D}(a,t)$ individuals/unit time.

where $q(a)$ stands for the coefficient of variation of the infection risk of individuals who has been previously infected in a previous episode.

In what regards the immunized branch, the individuals in classes L_s^v and R^v are supposed to have a lower risk of being reinfected, as a consequence of the effect of the vaccine, that is modeled by the presence of the ϵ factor in the following transitions:

- Exogenous re-infection of $L_s^v(a,t)$ individuals yielding primo-infection: from $L_s^v(a,t)$ to $L_f^v(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)L_s^v(a,t)$ individuals/unit time.

- Exogenous re-infection of $R^v_{npN}(a,t)$ individuals yielding primo-infection: from $R^v_{npN}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)N_{npN}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p-N}(a,t)$ individuals yielding primo-infection: from $R^v_{p-N}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p-N}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p+N}(a,t)$ individuals yielding primo-infection: from $R^v_{p+N}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p+N}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{npS}(a,t)$ individuals yielding primo-infection: from $R^v_{npS}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{npS}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p-S}(a,t)$ individuals yielding primo-infection: from $R^v_{p-S}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p-S}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p+S}(a,t)$ individuals yielding primo-infection: from $R^v_{p+S}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p+S}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{npD}(a,t)$ individuals yielding primo-infection: from $R^v_{npD}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{npD}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p-D}(a,t)$ individuals yielding primo-infection: from $R^v_{p-D}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p-D}(a,t)$ individuals/unit time.
- Exogenous re-infection of $R^v_{p+D}(a,t)$ individuals yielding primo-infection: from $R^v_{p+D}(a,t)$ to $L^v_f(a,t)$: $p(a)q(a)\epsilon\lambda(a,t)R^v_{p+D}(a,t)$ individuals/unit time.

In the other hand, if primo-infection after the secondary infection does not take place, even if the initial state of the individual is one of the possible R states, the rule is that no transition must be considered from these states to L_s, because it is always more likely an endogenous reactivation from those initial states to disease than from L_s, as either r_N, r_S or r_D are greater than any of the rates of transition from slow latency to disease ω_{snp}, ω_{sp-} and ω_{sp+}, (or even greater than the sum $\omega_{snp}+\omega_{sp-}+\omega_{sp+}$).

Smear progression In certain cases, it is documented that TB patients of smear negative pulmonary TB progress to smear positive, even after being treated. In order to describe this phenomenon, we introduce the smear progression by considering the following two transitions in the non immune branch of the model:

- Smear progression of untreated individuals: transition from $D_{p-}(a,t)$ to $D_{p+}(a,t)$: $\theta(a,t)D_{p-}(a,t)$ individuals/unit time.
- Smear progression of individuals under treatment: transition from $T_{p-}(a,t)$ to $T_{p+}(a,t)$: $\theta(a,t)T_{p-}(a,t)$ individuals/unit time.

where $\theta(a)$ stands for the smear progression rate. In the immune branch we have the same behavior:

- Smear progression of untreated individuals (immune branch): from $D^v_{p-}(a,t)$ to $D^v_{p+}(a,t)$: $\theta(a,t)D^v_{p-}(a,t)$ individuals/unit time.
- Smear progression of individuals under treatment (immune branch): from $T^v_{p-}(a,t)$ to $T^v_{p+}(a,t)$: $\theta(a,t)T^v_{p-}(a,t)$ individuals/unit time.

General and newborn vaccination. Mother-child disease transmission The vaccine described in the model has the only effect of diminishing the probability of infection –or reinfection– of immunized individuals. At a certain time t, the flux of susceptible individuals of age a $S(a,t)$ per unit time which gain immunity after having been vaccinated is described by the following transition from the unvaccinated branch to the immune one:

- Immunization of susceptible individuals: transition from $S(a,t)$ to $S^v(a,t)$: $f(a)S(a,t)$ individuals/unit time.

Under this point of view, the rate $f(a)$ defines the rhythm at which the population is immunized: $S(t) \simeq e^{-f(a)t}$. For example, if $f(a)$ =1/year, within the 5 years during which individuals belong to age group a, a fraction equal to $1 - e^{-5} = 0.993$ of the initial amount of susceptible individuals is immunized.

In addition, our model contemplates the description of newborn focused vaccination campaigns, like that required to introduce a novel BCG-substitutive vaccine on a population. In order to do so, if we have a number of $\Delta_N(a = 0, t)$ newborns at time-step t, we start by splitting them between the non immune branch and the immune branch: f_{neo} represents the fraction of newborns that are immunized. Among the fraction $(1 - f_{neo})$ of children who are not vaccinated, a percentage of them will be infected right after birth by their mothers, as it is a well known fact that a m_t fraction of sick women who are pregnant transmit the disease to their children within the first weeks of their lives [424]. This fraction of infected newborns obviously depend on the fraction of mothers who have the disease and are able to transmit it at time step t, but this opens the question of how is the relative risk for women in each of the infectious classes contemplated in our model to transmit the pathogen to their offspring. Since it is not a trivial question at all (note that mother child transmission does not even have to be necessarily airborne), we adopted the hypothesis of considering the total number of newborn infections proportional to the fraction m_d:

$$m_d(t) = \frac{\sum_{a=3}^{a=7} D_{p+}(a,t) + D_{p-}(a,t) + D_{np}(a,t) + R_{p+D}(a,t) + R_{p-D}(a,t)}{\sum_{a=3}^{a=7} N(a,t)} + \quad (9.2)$$
$$+ \frac{\sum_{a=3}^{a=7} R_{npD}(a,t) + F(a,t) + D^v_{p+}(a,t) + D^v_{p-}(a,t)}{\sum_{a=3}^{a=7} N(a,t)} +$$
$$+ \frac{\sum_{a=3}^{a=7} D^v_{np}(a,t) + R^v_{p+D}(a,t) + R^v_{p-D}(a,t) + R^v_{npD}(a,t)}{\sum_{a=3}^{a=7} N(a,t)}$$

that accounts to the fraction of infected individuals present in the age groups $a \in [3, 7]$ associated to women fertility (between 15 and 40 years old), regardless their relative infectiousnesses, that would act when talking about airborne disease transmission, which is generally not the case, since we hypothesize that other mechanisms besides airborne transmission may act in this case.

Therefore, this yields following the distribution of the $(1 - f_{neo})\Delta_N(a = 0, t)$ unvaccinated newborns among S and L classes:

- Birth of $S(0, t)$ individuals (susceptibles who will not be vaccinated right after birth): $(1 - m_t m_d(t))(1 - f_{neo})\Delta_N(a = 0, t)$
- Birth of $L_f(0, t)$ individuals (infected after birth who will develop primo-infection): $m_t m_d(t)p(0)(1 - f_{neo})\Delta_N(a = 0, t)$
- Birth of $L_s(0, t)$ individuals (infected after birth who will not develop primo-infection): $m_t m_d(t)(1 - p(0))(1 - f_{neo})\Delta_N(a = 0, t)$

In what regards the immune branch of the system, if we suppose that vaccination has also a prophylactic effect –i.e. it diminishes the probability for a sick mother to transmit the disease to her child from m_t to ϵm_t– on impairing the mechanisms of transmission between mother and child, we have that the $f_{neo}\Delta_N(a = 0, t)$ newborns that are immunized at time step t are distributed among $S^v$and L^v classes like follows:

- Birth of $S^v(0, t)$ individuals (susceptible, vaccinated right after birth): $(1 - \epsilon m_t m_d(t))f_{neo}\Delta_N(a = 0, t)$
- Birth of $L^v_f(0, t)$ individuals (vaccinated, yet infected after birth who will develop primo-infection): $\epsilon m_t m_d(t)p(0)f_{neo}\Delta_N(a = 0, t)$
- Birth of $L^v_s(0, t)$ individuals (vaccinates, yet infected after birth who will not develop primo-infection): $\epsilon m_t m_d(t)(1 - p(0))f_{neo}\Delta_N(a = 0, t)$

Finally, individuals of all ages in the immune branch have, according to our model, a reduced probability of getting infected (or re-infected) with TB: i.e. a partial (or absolute) immunity to infection or reinfection. But the immunogenic action of actual vaccines is not always a lifelong effect. In order to take this phenomenology into consideration, we have introduced the following set of transitions from any state X^v –except S^v– of the immune branch to any state X of the unvaccinated branch:

- Loss of immunity of vaccinated groups X^v (except S^v): transition from $X^v(a, t)$ to $X(a, t)$: $\gamma(a)X^v(a, t)$ individuals/unit time.

where $\gamma(a)$ is the immunity waning rate (i.e. the inverse of the vaccine immunity duration). In addition, we have an additional class $S^w(a, t)$ that does not belong to the unvaccinated or the immune branch; that represents the susceptible individuals who have lost their immunity without ever being infected with TB. Therefore, the immunity waning of S^v individuals is described like this:

- Loss of immunity of vaccinated susceptible individuals: transition from $S^v(a,t)$ to $S^w(a,t)$: $\gamma(a)S^v(a,t)$ individuals/unit time.

The reason for the need of this additional class is that, in case you describe the immunity waning of susceptible individuals by making them return to the unvaccinated branch, –as we do with the rest of the classes– then, the model would allow them to be vaccinated again, which eventually would cause a looped, unrealistic succession of immunization gain and loss transitions depending on the parameters values. Summarizing, the class $S^w(a,t)$ suffer the same risk of infection of $S(a,t)$, but individuals within it cannot be vaccinated again. With the rest of classes, instead, we can describe the immunity loss as a return to the first branch, because after being infected once; the model do not allow individuals to be vaccinated, and so, they will never be allowed to be vaccinated again.

Force of infection

The force of infection $\lambda(a,t)$ is, as it has been said before, the probability of being infected at time step t of a susceptible individual. This magnitude is calculated according to the following expression:

$$\lambda(a,t) = \beta(t) \sum_{a'} \xi(a,a') \Upsilon(a',t) \tag{9.3}$$

being $\Upsilon(a',t)$ the weighted sum of all the infectious individuals within age-group a' at time step t:

$$\begin{aligned}\Upsilon(a',t) = D_{p+}(a',t) + F(a',t) + \phi_{p-} D_{p-}(a',t) + \phi_D R_{p+D}(a',t) + \phi_{p-}\phi_D R_{p-D} + \\ + D^v_{p+}(a',t) + \phi_{p-} D^v_{p-}(a',t) + \phi_D R^v_{p+D}(a',t) + \phi_{p-}\phi_D R^v_{p-D}\end{aligned} \tag{9.4}$$

where, in the one hand, $\phi_{p-} \in [0,1]$ is the coefficient of infectivity reduction of smear negative sick individuals with respect to smear positive ones, and $\phi_D \in [0,1]$ the infectivity reduction of individuals who have defaulted the treatment, $-R_{p+D}$ individuals with respect to smear positive, undiagnosed individuals $D_{p+}-$. Diagnosed patients of smear negative TB who had defaulted their treatment have a infectivity reduction that is the product of the two terms $\phi_{p-}\phi_D$.

In the other hand, $\xi(a,a')$ represents the number of contacts per year that an individual within the age group a maintains with individuals in age group a'. In addition, $\beta(t)$ represents the scaled based infectiousness, that is considered to exponentially vary as follows:

$$\beta(t) = \beta_o e^{\alpha_\beta t} \tag{9.5}$$

where β_o is the initial value of $\beta(t)$ at the beginning of the period studied ($t = 2001$) and α_β represents the annual rate of variation of $\beta(t)$. The reason for which we call $\beta(t)$ as the "scaled" infectivity is that it is can be interpreted as the actual infectivity

–i.e. the probability of disease transmission per D_{p+}-S contact, which we can call λ_o– per a certain scaling factor σ_ξ which, under this conception would represent an average scaling factor of the contact matrix $\xi(a, a')$. Under this assumption, $\beta(t) = \lambda_o(t)\sigma_\xi(t)$. However, as it is by definition indistinguishable the contribution of both factors λ_o and σ_ξ, what it is used in the equations is $\beta(t)$ and the parameters that are explicitly fitted by the model are β_o and α_β.

Aging

The model considers 14 different age groups, one of each of a duration T of 5 years, in such a way that individuals aging is described up to 70 years old; moment in which people is assumed to die because of natural causes. The relevance of such an age structured description of the system comes from the fact that some of the most relevant dynamic parameters take different values in the different age groups. As time goes by, individuals get older and pass trough the different age groups; this situation is modelled by the introduction on the system of equations of the following aging transitions, that stands for the promotion of individuals within whatever dynamic class of the model $X(a, t)$ to the next age class $X(a + 1, t)$

- Aging of individuals belonging to class $X(a, t)$ transition from $X(a, t)$ to $X(a + 1, t)$: $X(a, t)/\Delta_t$ individuals/unit time.

Obviously, each class $X(a, t)$ receive people from $X(a - 1, t)$ and sends people to $X(a + 1, t)$, except $X(0, t)$, that only receives newborns and $X(13, t)$, from which the exit flow represent the death of the eldest individuals in the system.

Demographic evolution

Once all the transitions among the different dynamical states of the system have been described, as well as the aging fluxes, it is necessary to provide a global description of the evolution of the demographic pyramid, given by the evolution of the set of variables $N(a, t)$, defined as the total amount of individuals within age group a in the population, no matter their states regarding TB dynamics:

$$\begin{aligned}
N(a,t) = S(a,t) + L_f(a,t) + L_s(a,t) + D_{p+}(a,t) + \\
+D_{p-}(a,t) + D_{np}(a,t) + T_{p+}(a,t) + T_{p-}(a,t) + T_{np}(a,t) + \\
+F(a,t) + R_{p+N}(a,t) + R_{p-N}(a,t) + R_{npN}(a,t) + R_{p+S}(a,t) + \\
+R_{p-S}(a,t) + R_{npS}(a,t) + R_{p+D}(a,t) + R_{p-D}(a,t) + R_{npD}(a,t) + \\
+S^w(a,t) + S^v(a,t) + L_f^v(a,t) + L_s^v(a,t) + D_{p+}^v(a,t) + \\
+D_{p-}^v(a,t) + D_{np}^v(a,t) + T_{p+}^v(a,t) + T_{p-}^v(a,t) + T_{np}^v(a,t) + \\
+F^v(a,t) + R_{p+N}^v(a,t) + R_{p-N}^v(a,t) + R_{npN}^v(a,t) + R_{p+S}^v(a,t) + \\
+R_{p-S}^v(a,t) + R_{npS}^v(a,t) + R_{p+D}^v(a,t) + R_{p-D}^v(a,t) + R_{npD}^v(a,t)
\end{aligned} \tag{9.6}$$

the evolution of $N(a,t)$ -in addition to aging and death by TB– is subject to other driving forces related to aspects like vegetative variation of population –births and non-TB deaths– as well as migration. In order to provide a description of the temporal evolution of the demographic pyramid, previous models have turned to different simplifying hypothesis to describe the system. One of them consists of forcing the system to preserve at any time the total number of individuals $\mathcal{N}(t)$ [22]:

$$\mathcal{N}(t) = \sum_a N(a,t) \tag{9.7}$$

by imposing that $\mathcal{N}(t) = \mathcal{N}(t=0)\forall(t)$. Another one [422] is based on forcing the system to preserve the initial age structure of the population by imposing that, in each age group: $N(a,t) = N(a,t=t_o)\forall(t)$. Our approach, however is based on forcing the temporal evolution of the variables $N(a,t)$ to follow the predictions that are made for them by the United Nations Population Division, available at their on-line databases [23]: $N(a,t) = N_{UN}(a,t)$, which are represented in figure 9.3.

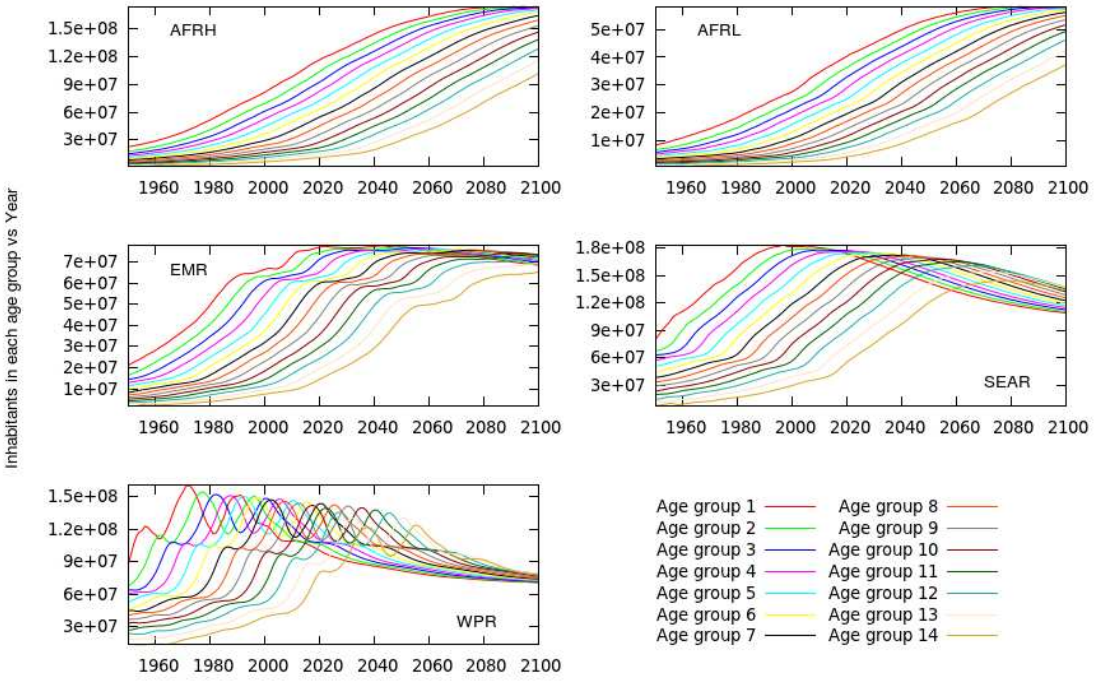


FIGURE 9.3: Population pyramids evolution projected by UN population division [23] for the next decades in each of the region under study.

As we can see from figure 9.3, population ageing is a common feature in all the regions under study.

In the following sub-sections, we detail the three different schemes, and, after that, we proceed to quantify the biases that are due to the adoption of either of the two simplifying hypothesis (approaches 1 and 2) with respect to the last, more realistic modeling strategy.

Approach 1: populations of constant volume: $\mathcal{N}(t) = \mathcal{N}(t=0)\forall(t)$ The way to implement this type of coupling is to calculate the total number of individuals dead because of TB (and abandon of the system of the last age class $a = 13$) at time step t:

$$\begin{aligned}\dot{N}_o(t) = -N(a=13,t)/\tau - \\ \sum_a [\Psi f_\mu^{p+}(T_{p+}(a,t) + T_{p+}^v(a,t)) + \\ \Psi f_\mu^{p-}(T_{p-}(a,t) + T_{p-}^v(a,t) + T_{np}(a,t) + T_{np}^v(a,t)) + \\ +\mu_{p+}(D_{p+}(a,t) + F(a,t) + D_{p+}^v(a,t)) + \\ +\mu_{p-}(D_{p-}(a,t) + D_{p-}^v(a,t)) + \mu_{np}(D_{np}(a,t) + D_{np}^v(a,t))]\end{aligned}$$

and balance it by introducing the same number of individuals at age group $a = 0$ as newborns:

$$\Delta_N(a=0,t) = -\dot{\mathcal{N}}_o(t) \tag{9.8}$$

so as to get:

$$\dot{\mathcal{N}}(t) = \dot{\mathcal{N}}_o(t) + \Delta_N(a=0,t) = 0 \tag{9.9}$$

All the individuals introduced can be introduced as susceptible as it is done in [22], or, according to our model, they can be distributed among S, L, S^v and L^v classes.

The problem with this kind of approach is that, during the dynamic process, the demographic pyramid is distorted in a way that is not under control. This situation, given that the dynamic parameters are dependent on age ultimately affects the model forecasts. In addition, by neglecting the population growth, without further information, no measures –i.e. incidence, mortality or even disease prevalence– can be done in terms of raw number of individuals, but only in terms of rates; at least, without further information about the actual volume of the population under study.

Approach 2: Constant demographic pyramid A typical alternative approach consists in imposing that, not just the total volume of population, but the demographic pyramid itself has to remain constant during the dynamical process: i.e. $N(a,t) = N(a,t_o)\forall t$. This is what is done in some previous works, like [422], in which, the dynamical states indicates densities rather than numbers of individuals. In that work, the force of infection calculated as an average of the densities of sick individuals in each age group, weighted by the number of individuals within each age class of the demographic pyramid. This way, authors of [422] provide a way for calculating infection and mortality rates that takes into account the initial demographic structure of the

population; and these rates can be eventually transformed into numbers by using data about the evolution of the total population under consideration.

In order to provide an equivalent description based upon our model –in which states represent number of individuals rather than densities–, we start by calculating the variation of population due to TB and aging in each age group:

$$\begin{aligned}\dot{N}_o(a,t) = ((1-\delta(a))N(a-1,t) - N(a,t))/\Delta_t - \\ -\mu_{p+}(D_{p+}(a,t) + F(a,t) + D^v_{p+}(a,t)) - \\ -\mu_{p-}(D_{p-}(a,t) + D^v_{p-}(a,t)) - \mu_{np}(D_{np}(a,t) + D^v_{np}(a,t)) - \\ -\Psi f^{p+}_{\mu}(T_{p+}(a,t) + T^v_{p+}(a,t)) - \Psi f^{p-}_{\mu}(T_{p-}(a,t) + T^v_{p-}(a,t) + T_{np}(a,t) + T^v_{np}(a,t))\end{aligned} \tag{9.10}$$

being $\delta(a)$ the Kronecker delta function. In order to preserve the number of individuals within each age group at any moment, we simply introduce a term $\Delta_N(a,t)$ that is intended to balance $\dot{N}_o(a,t)$ within each age group: $\Delta_N(a,t) = -\dot{N}_o(a,t)$, yielding:

$$\dot{N}(a,t) = \dot{N}_o(a,t) + \Delta_N(a,t) = 0 \forall (a,t) \tag{9.11}$$

The key question is how do we distribute these correction terms $\Delta_N(a,t)$ among the different dynamical states $X(a,t)$ within age-class a: these increments have to be distributed among $X(a,t)$ dynamical states respecting the relative volume of these states within the age group so as not to introduce external, undesired biases on states densities. If we call $\Delta_X(a,t)$ the fraction of $\Delta_N(a,t)$ that is introducing in state $X(a,t)$, we have:

$$\Delta_X(a,t) = \Delta_N(a,t)\frac{X(a,t)}{N(a,t)} \tag{9.12}$$

and obviously:

$$\Delta_N(a,t) = \sum_X \Delta_X(a,t) \tag{9.13}$$

This scheme has the advantage, with respect to consider $\mathcal{N}(t) = \mathcal{N}(t=0) \forall (t)$, that, at least, the initial structure of the population is conserved. However, as in the first case, population growth is not explicitly considered, and further information about population volume is required so as to scale rates into numbers, as done in [422]. Additionally, the main problem with this approach comes from the fact that no variation of the age structure of the population can be considered by proceeding this way, which might introduce significant biases, specially when studying regions subjected to strong processes of demographic aging or rejuvenation.

Approach 3: Demographic pyramid as an external constraint Starting from the last scheme for modeling the demographic evolution, it is easy to obtain a final approach that explicitly takes into account not only the influence of the age structure into the spreading but also the population growth and the variation of the demographic

pyramid itself. In order to do that, obviously, it is necessary to know the actual –or projected– evolution of the demographic pyramid of the population during the period under analysis. In our case, we are modeling TB dynamics from 2001 to 2050; and the official annual projections for the population per age group of any country are available, up to 2100, at the UN population division database [23]. After adequately grouping the countries, we obtain the actual annual population series expected by the UN for the populations at each age group, that can be trivially fitted to a continuos function $N_{UN}(a,t)$, from which we can derive an analytical derivative $\dot{N}_{UN}(a,t)$ at any moment. For the purpose of this chapter, a polynomial of degree 10 is more than enough for building the continuous function $N_{UN}(a,t)$ from the annual data series from [23] during the period under study.

So, if we recover the variation of population due to TB and aging in each age group, initially provided by the model:

$$\begin{aligned}\dot{N}_o(a,t) = ((1-\delta(a))N(a-1,t) - N(a,t))/\Delta_t - \\ -\mu_{p+}(D_{p+}(a,t) + F(a,t) + D^v_{p+}(a,t)) - \\ -\mu_{p-}(D_{p-}(a,t) + D^v_{p-}(a,t)) - \mu_{np}(D_{np}(a,t) + D^v_{np}(a,t)) - \\ -\Psi f^{p+}_{\mu}(T_{p+}(a,t) + T^v_{p+}(a,t)) - \Psi f^{p+}_{\mu}(T_{p-}(a,t) + T^v_{p-}(a,t) + T_{np}(a,t) + T^v_{np}(a,t))\end{aligned} \tag{9.14}$$

now, we also introduce a term $\Delta_N(a,t)$, but this time, it is not aimed at balancing $\dot{N}_o(a,t)$, but to force the total temporal evolution of $N(a,t)$ to follow precisely the function $N_{UN}(a,t)$. That is achieved by defining, at each time step:

$$\Delta_N(a,t) = \dot{N}_{UN}(a,t) - \dot{N}_o(a,t) \tag{9.15}$$

and introducing those $\Delta_N(a,t)$ terms into the system dynamics, so having: $\dot{N}(a,t) = \dot{N}_o(a,t) + \Delta_N(a,t) = \dot{N}_{UN}(a,t)$. Finally, provided that the initial conditions have been properly set: $N(a,t=0) = N_{UN}(a,t=0) \forall a$, this yields the desired behavior for the demographic pyramid; i.e. $N(a,t) = N_{UN}(a,t) \forall (a,t)$.

Again, the $\Delta_N(a,t)$ forcing terms have to be introduced into the different dynamical states within the same age class preserving their proportions, at least in the age groups $a > 0$:

$$\Delta_X(a,t) = \Delta_N(a,t)\frac{X(a,t)}{N(a,t)} \forall a > 0 \tag{9.16}$$

and, under this assumption, the terms $\Delta_N(a,t)$ for $a > 0$, represents the variations of volume of the age group a due to whatever causes foreign to TB infection and individuals aging; which would include all death not caused by TB, and migration, assuming that these factors affect all the dynamics classes regardless their state with respect to TB infection.

The situation is different for the first age class $a = 0$. In the first age group, the introduction of newborn individuals is the main cause of population variation and it is not distributed uniformly among the different dynamical states, but only among S

and L states. For these reasons, and once observed that $\Delta_N(a=0,t) > 0 \forall t$ in all regions under consideration, for simplicity $\Delta_N(a=0,t)$ is directly associated to the number of newborns, as it was previously done in the section devoted to immunization and mother-child disease transmision, in which we describe the way in which these $\Delta_N(a=0,t)$ newborns are distributed among classes S, S^v, L and L^v.

The uncertainty of United nations demographic projections is also reported at [23], which allows us to reconstruct the demographic pyramids at the extreme of the confidence interval $N_{UN}^{low}(a,t)$ and $N_{UN}^{high}(a,t)$. Therefore, its influence on the model forecasts is also evaluable, as we will discuss on the section devoted to uncertainty and sensitivity analysis.

Overall ODEs system

The following system of differential equations describe evolution of the different dynamical states of the model:

$$\begin{aligned}
\dot{S}(a,t) &= -f(a)S(a,t) - \lambda(a,t)S(a,t) - \\
&- (S(a,t) - (1-\delta(a))S(a-1,t))/\tau + \\
&+ \delta(a)(1-m_t m_d(t))(1-f_{new})\Delta_N(a,t) + \\
&+ \delta(a)\Delta_N(a,t)S(a,t)/N(a,t)
\end{aligned} \tag{9.17}$$

$$\begin{aligned}
\dot{L}_s(a,t) &= \gamma L_s^v(a,t) + (1-p(a))\lambda(a,t)(S(a,t)+S^w(a,t)) - \\
&- p(a)q\lambda(a,t)L_s - (\omega_{sp+}(a)+\omega_{sp-}(a)+\omega_{snp}(a))L_s(a,t) - \\
&- (L_s(a,t) - (1-\delta(a))L_s(a-1,t))/\tau + \\
&+ \delta(a)m_t m_d(t)(1-p(a))\Delta_N(a,t) + \delta(a)\Delta_N(a,t)L_s(a,t)/N(a,t)
\end{aligned} \tag{9.18}$$

$$\begin{aligned}
\dot{L}_f(a,t) &= \gamma L_f^v(a,t) + p(a)\lambda(a,t)(S(a,t)+S^w(a,t)) - \\
&- (\omega_{fp+}(a)+\omega_{fp-}(a)+\omega_{fnp}(a))L_f(a,t) + \\
&+ p(a)q\lambda(a,t)(L_s(a,t)+R_{p+N}(a,t)+R_{p-N}(a,t)) + \\
&+ p(a)q\lambda(a,t)(R_{npN}(a,t)+R_{p+S}(a,t)+R_{p-S}(a,t)) + \\
&+ p(a)q\lambda(a,t)(R_{npS}(a,t)+R_{p+D}(a,t)+R_{p-D}(a,t)+R_{npD}(a,t)) - \\
&- (L_f(a,t) - (1-\delta(a))L_f(a-1,t))/\tau + \\
&+ \delta(a)m_t m_d(t)p(a)\Delta_N(a,t) + \delta(a)\Delta_N(a,t)L_f(a,t)/N(a,t)
\end{aligned} \tag{9.19}$$

$$\begin{aligned}
\dot{D}_{p+}(a,t) &= \gamma D_{p+}^v(a,t) + \omega_{fp+}(a)L_f(a,t) + \omega_{sp+}(a)L_s(a,t) - \mu_{p+}D_{p+}(a,t) - \\
&- d(t)D_{p+}(a,t) - \nu D_{p+}(a,t) + r_N R_{p+N}(a,t) + \\
&+ r_S R_{p+S}(a,t) + r_D R_{p+D}(a,t) + \theta D_{p-}(a,t) - \\
&- (D_{p+}(a,t) - (1-\delta(a))D_{p+}(a-1,t))/\tau + \\
&+ \Delta_N(a,t)D_{p+}(a,t)/N(a,t)
\end{aligned} \tag{9.20}$$

$$\begin{aligned}\dot{D}_{p-}(a,t) &= \gamma D^v_{p-}(a,t) + \omega_{fp-}(a)L_f(a,t) + \omega_{sp-}(a)L_s(a,t) - \mu_{p-}D_{p-}(a,t) - \\ &- \eta d(t)D_{p-}(a,t) - \nu D_{p-}(a,t) + r_N R_{p-N}(a,t) + \\ &+ r_S R_{p-S}(a,t) + r_D R_{p-D}(a,t) - \theta D_{p-}(a,t) - \\ &- (D_{p-}(a,t) - (1-\delta(a))D_{p-}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)D_{p-}(a,t)/N(a,t) \end{aligned} \tag{9.21}$$

$$\begin{aligned}\dot{D}_{np}(a,t) &= \gamma D^v_{np}(a,t) + \omega_{fnp}(a)L_f(a,t) + \omega_{snp}(a)L_s(a,t) - \mu_{np}D_{np}(a,t) - \\ &- \eta d(t)D_{np}(a,t) - \nu D_{np}(a,t) + r_N R_{npN}(a,t) + \\ &+ r_S R_{npS}(a,t) + r_D R_{npD}(a,t) - \\ &- (D_{np}(a,t) - (1-\delta(a))D_{np}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)D_{np}(a,t)/N(a,t) \end{aligned} \tag{9.22}$$

$$\begin{aligned}\dot{T}_{p+}(a,t) &= \gamma T^v_{p+}(a,t) + d(t)D_{p+}(a,t) - \Psi T_{p+}(a,t) + \theta T_{p-}(a,t) - \\ &- (T_{p+}(a,t) - (1-\delta(a))T_{p+}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T_{p+}(a,t)/N(a,t) \end{aligned} \tag{9.23}$$

$$\begin{aligned}\dot{T}_{p-}(a,t) &= \gamma T^v_{p-}(a,t) + \eta d(t)D_{p-}(a,t) - \Psi T_{p-}(a,t) - \theta T_{p-}(a,t) - \\ &- (T_{p-}(a,t) - (1-\delta(a))T_{p-}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T_{p-}(a,t)/N(a,t) \end{aligned} \tag{9.24}$$

$$\begin{aligned}\dot{T}_{np}(a,t) &= \gamma T^v_{np}(a,t) + \eta d(t)D_{np}(a,t) - \Psi T_{np}(a,t) - \\ &- (T_{np}(a,t) - (1-\delta(a))T_{np}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T_{np}(a,t)/N(a,t) \end{aligned} \tag{9.25}$$

$$\begin{aligned}\dot{F}(a,t) &= \Psi f_F^{p+}(T_{p+}(a,t) + T^v_{p+}(a,t)) + \\ &+ \Psi f_F^{p-}(T_{p-}(a,t) + T^v_{p-}(a,t) + T_{np}(a,t) + T^v_{np}(a,t)) - \\ &- (F(a,t) - (1-\delta(a))F(a-1,t))/\tau + \Delta_N(a,t)F(a,t)/N(a,t) \end{aligned} \tag{9.26}$$

$$\begin{aligned}\dot{R}_{p+N}(a,t) &= \gamma R^v_{p+N}(a,t) + \nu D_{p+}(a,t) - r_N R_{p+N}(a,t) - \\ &- p(a)q\lambda(a,t)R_{p+N}(a,t) - \\ &- (R_{p+N}(a,t) - (1-\delta(a))R_{p+N}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p+N}(a,t)/N(a,t) \end{aligned} \tag{9.27}$$

$$\begin{aligned}\dot{R}_{p-N}(a,t) &= \gamma R^v_{p-N}(a,t) + \nu D_{p-}(a,t) - r_N R_{p-N}(a,t) - \\ &- p(a)q\lambda(a,t)R_{p-N}(a,t) - \\ &- (R_{p-N}(a,t) - (1-\delta(a))R_{p-N}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p-N}(a,t)/N(a,t) \end{aligned} \tag{9.28}$$

$$\begin{aligned}\dot{R}_{npN}(a,t) &= \gamma R^v_{npN}(a,t) + \nu D_{np}(a,t) - r_N R_{npN}(a,t) - \\ &- p(a)q\lambda(a,t)R_{npN}(a,t) - \\ &- (R_{npN}(a,t) - (1-\delta(a))R_{npN}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{npN}(a,t)/N(a,t) \end{aligned} \tag{9.29}$$

$$\begin{aligned}\dot{R}_{p+S}(a,t) &= \gamma R^v_{p+S}(a,t) + \Psi(1 - f_D^{p+} - f_F^{p+} - f_\mu^{p+})T_{p+}(a,t) - \\ &- r_S R_{p+S}(a,t) - p(a)q\lambda(a,t)R_{p+S}(a,t) - \\ &- (R_{p+S}(a,t) - (1-\delta(a))R_{p+S}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p+S}(a,t)/N(a,t) \end{aligned} \tag{9.30}$$

$$\begin{aligned}\dot{R}_{p-S}(a,t) &= \gamma R^v_{p-S}(a,t) + \Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-})T_{p-}(a,t) - \\ &- r_S R_{p-S}(a,t) - p(a)q\lambda(a,t)R_{p-S}(a,t) - \\ &- (R_{p-S}(a,t) - (1-\delta(a))R_{p-S}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p-S}(a,t)/N(a,t) \end{aligned} \tag{9.31}$$

$$\begin{aligned}\dot{R}_{npS}(a,t) &= \gamma R^v_{npS}(a,t) + \Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-})T_{np}(a,t) - \\ &- r_S R_{npS}(a,t) - p(a)q\lambda(a,t)R_{npS}(a,t) - \\ &- (R_{npS}(a,t) - (1-\delta(a))R_{npS}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{npS}(a,t)/N(a,t) \end{aligned} \tag{9.32}$$

$$\begin{aligned}\dot{R}_{p+D}(a,t) &= \gamma R^v_{p+D}(a,t) + \Psi f_D^{p+} T_{p+}(a,t) - r_D R_{p+D}(a,t) - \\ &- p(a)q\lambda(a,t)R_{p+D}(a,t) - \\ &- (R_{p+D}(a,t) - (1-\delta(a))R_{p+D}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p+D}(a,t)/N(a,t) \end{aligned} \tag{9.33}$$

$$\begin{aligned}\dot{R}_{p-D}(a,t) &= \gamma R^v_{p-D}(a,t) + \Psi f_D^{p-} T_{p-}(a,t) - r_D R_{p-D}(a,t) - \\ &- p(a)q\lambda(a,t)R_{p-D}(a,t) - \\ &- (R_{p-D}(a,t) - (1-\delta(a))R_{p-D}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R_{p-D}(a,t)/N(a,t) \end{aligned} \tag{9.34}$$

$$\begin{aligned}\dot{R}_{npD}(a,t) &= \gamma R^v_{npD}(a,t) + \Psi f_D^{p-} T_{np}(a,t) - r_D R_{npD}(a,t) - \\ &- p(a) q \lambda(a,t) R_{npD}(a,t) - \\ &- (R_{npD}(a,t) - (1-\delta(a)) R_{npD}(a-1,t))/\tau + \\ &+ \Delta_N(a,t) R_{npD}(a,t)/N(a,t) \end{aligned} \tag{9.35}$$

$$\begin{aligned}\dot{S}^w(a,t) &= \gamma S^v(a,t) - \lambda(a,t) S^w(a,t) + \Delta_N(a,t) S^w(a,t)/N(a,t) \\ &- (S^w(a,t) - (1-\delta(a)) S^w(a-1,t))/\tau \end{aligned} \tag{9.36}$$

$$\begin{aligned}\dot{S}^v(a,t) &= f(a) S(a,t) - \gamma S^v(a,t) - \epsilon \lambda(a,t) S^v(a,t) - \\ &- (S^v(a,t) - (1-\delta(a)) S^v(a-1,t))/\tau + \\ &+ \delta(a)(1 - m_t m_d(t)) f_{new} \Delta_N(a,t) + \delta(a) \Delta_N(a,t) S^v(a,t)/N(a,t) \end{aligned} \tag{9.37}$$

$$\begin{aligned}\dot{L}^v_s(a,t) &= -\gamma L^v_s(a,t) + (1-p(a)) \epsilon \lambda(a,t) S^v(a,t) - \\ &- (\omega_{sp+}(a) + \omega_{sp-}(a) + \omega_{snp}(a)) L^v_s(a,t) - p(a) q \epsilon \lambda(a,t) L^v_s(a,t) - \\ &- (L^v_s(a,t) - (1-\delta(a)) L^v_s(a-1,t))/\tau + \Delta_N(a,t) L^v_s(a,t)/N(a,t) \end{aligned} \tag{9.38}$$

$$\begin{aligned}\dot{L}^v_f(a,t) &= -\gamma L^v_f(a,t) + p(a) \epsilon(a) \lambda(a,t) S^v(a,t) - (\omega_{fp+}(a) + \omega_{fp-}(a) + \omega_{fnp}(a)) L^v_f(a,t) + \\ &+ p(a) q \epsilon \lambda(a,t) (L^v_s(a,t) + R^v_{p+N}(a,t) + R^v_{p-N}(a,t) + R^v_{npN}(a,t)) + \\ &+ p(a) q \epsilon \lambda(a,t) (R^v_{p+S}(a,t) + R^v_{p-S}(a,t) + R^v_{npS}(a,t)) + \\ &+ p(a) q \epsilon \lambda(a,t) (R^v_{p+D}(a,t) + R^v_{p-D}(a,t) + R^v_{npD}(a,t)) - \\ &- (L^v_f(a,t) - (1-\delta(a)) L^v_f(a-1,t))/\tau + \Delta_N(a,t) L^v_f(a,t)/N(a,t) \end{aligned} \tag{9.39}$$

$$\begin{aligned}\dot{D}^v_{p+}(a,t) &= -\gamma D^v_{p+}(a,t) + \omega_{fp+}(a) L^v_f(a,t) + \omega_{sp+}(a) L^v_s(a,t) - \mu_{p+} D^v_{p+}(a,t) - \\ &- d(t) D^v_{p+}(a,t) - \nu D^v_{p+}(a,t) + r_N R^v_{p+N}(a,t) + r_S R^v_{p+S}(a,t) + r_D R^v_{p+D}(a,t) + \\ &+ \theta D^v_{p-}(a,t) - (D^v_{p+}(a,t) - (1-\delta(a)) D^v_{p+}(a-1,t))/\tau + \\ &+ \Delta_N(a,t) D^v_{p+}(a,t)/N(a,t) \end{aligned} \tag{9.40}$$

$$\begin{aligned}\dot{D}^v_{p-}(a,t) &= -\gamma D^v_{p-}(a,t) + \omega_{fp-}(a) L^v_f(a,t) + \omega_{sp-}(a) L^v_s(a,t) - \mu_{p-} D^v_{p-}(a,t) - \\ &- \eta d(t) D^v_{p-}(a,t) - \nu D^v_{p-}(a,t) + r_N R^v_{p-N}(a,t) + r_S R^v_{p-S}(a,t) + r_D R^v_{p-D}(a,t) - \\ &- \theta D^v_{p-}(a,t) - (D^v_{p-}(a,t) - (1-\delta(a)) D^v_{p-}(a-1,t))/\tau + \\ &+ \Delta_N(a,t) D^v_{p-}(a,t)/N(a,t) \end{aligned} \tag{9.41}$$

$$\begin{aligned}\dot{D}^v_{np}(a,t) &= -\gamma D^v_{np}(a,t) + \omega_{fnp}(a)L^v_f(a,t) + \omega_{snp}(a)L^v_s(a,t) - \mu_{np}D^v_{np}(a,t) - \\ &- \eta d(t)D^v_{np}(a,t) - \nu D^v_{np}(a,t) + r_N R^v_{npN}(a,t) + r_S R^v_{npS}(a,t) + \\ &+ r_D R^v_{npD}(a,t) - (D^v_{np}(a,t) - (1-\delta(a))D^v_{np}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)D^v_{np}(a,t)/N(a,t) \end{aligned} \tag{9.42}$$

$$\begin{aligned}\dot{T}^v_{p+}(a,t) &= -\gamma T^v_{p+}(a,t) + d(t)D^v_{p+}(a,t) - \Psi T^v_{p+}(a,t) + \theta T^v_{p-}(a,t) - \\ &- (T^v_{p+}(a,t) - (1-\delta(a))T^v_{p+}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T^v_{p+}(a,t)/N(a,t) \end{aligned} \tag{9.43}$$

$$\begin{aligned}\dot{T}^v_{p-}(a,t) &= -\gamma T^v_{p-}(a,t) + \eta d(t)D^v_{p-}(a,t) - \Psi T^v_{p-}(a,t) - \theta T^v_{p-}(a,t) - \\ &- (T^v_{p-}(a,t) - (1-\delta(a))T^v_{p-}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T^v_{p-}(a,t)/N(a,t) \end{aligned} \tag{9.44}$$

$$\begin{aligned}\dot{T}^v_{np}(a,t) &= -\gamma T^v_{np}(a,t) + \eta d(t)D^v_{np}(a,t) - \Psi T^v_{np}(a,t) - \\ &- (T^v_{np}(a,t) - (1-\delta(a))T^v_{np}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)T^v_{np}(a,t)/N(a,t) \end{aligned} \tag{9.45}$$

$$\begin{aligned}\dot{R}^v_{p+N}(a,t) &= -\gamma R^v_{p+N}(a,t) + \nu D^v_{p+}(a,t) - r_N R^v_{p+N}(a,t) - \\ &- p(a)q\epsilon\lambda(a,t)R^v_{p+N}(a,t) - \\ &- (R^v_{p+N}(a,t) - (1-\delta(a))R^v_{p+N}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p+N}(a,t)/N(a,t) \end{aligned} \tag{9.46}$$

$$\begin{aligned}\dot{R}^v_{p-N}(a,t) &= -\gamma R^v_{p-N}(a,t) + \nu D^v_{p-}(a,t) - r_N R^v_{p-N}(a,t) - \\ &- p(a)q\epsilon\lambda(a,t)R^v_{p-N}(a,t) - \\ &- (R^v_{p-N}(a,t) - (1-\delta(a))R^v_{p-N}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p-N}(a,t)/N(a,t) \end{aligned} \tag{9.47}$$

$$\begin{aligned}\dot{R}^v_{npN}(a,t) &= -\gamma R^v_{npN}(a,t) + \nu D^v_{np}(a,t) - r_N R^v_{npN}(a,t) - \\ &- p(a)q\epsilon\lambda(a,t)R^v_{npN}(a,t) - \\ &- (R^v_{npN}(a,t) - (1-\delta(a))R^v_{npN}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{npN}(a,t)/N(a,t) \end{aligned} \tag{9.48}$$

$$\begin{aligned}\dot{R}^v_{p+S}(a,t) &= -\gamma R^v_{p+S}(a,t) + \Psi(1 - f_D^{p+} - f_F^{p+} - f_\mu^{p+})T^v_{p+}(a,t) - r_S R^v_{p+S}(a,t) - \\ &- p(a)q\epsilon\lambda(a,t)R^v_{p+S}(a,t) - (R^v_{p+S}(a,t) - (1-\delta(a))R^v_{p+S}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p+S}(a,t)/N(a,t) \end{aligned} \tag{9.49}$$

$$\begin{aligned}\dot{R}^v_{p-S}(a,t) &= -\gamma R^v_{p-S}(a,t) + \Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-})T^v_{p-}(a,t) - \\ &- r_S R^v_{p-S}(a,t) - p(a)q\epsilon\lambda(a,t)R^v_{p-S}(a,t) - \\ &- (R^v_{p-S}(a,t) - (1-\delta(a))R^v_{p-S}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p-S}(a,t)/N(a,t) \end{aligned} \tag{9.50}$$

$$\begin{aligned}\dot{R}^v_{npS}(a,t) &= -\gamma R^v_{npS}(a,t) + \Psi(1 - f_D^{p-} - f_F^{p-} - f_\mu^{p-})T^v_{np}(a,t) - \\ &- r_S R^v_{npS}(a,t) - p(a)q\epsilon\lambda(a,t)R^v_{npS}(a,t) - \\ &- (R^v_{npS}(a,t) - (1-\delta(a))R^v_{npS}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{npS}(a,t)/N(a,t) \end{aligned} \tag{9.51}$$

$$\begin{aligned}\dot{R}^v_{p+D}(a,t) &= -\gamma R^v_{p+D}(a,t) + \Psi f_D^{p+} T^v_{p+}(a,t) - r_D R^v_{p+D}(a,t) - p(a)q\epsilon\lambda(a,t)R^v_{p+D}(a,t) - \\ &- (R^v_{p+D}(a,t) - (1-\delta(a))R^v_{p+D}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p+D}(a,t)/N(a,t) \end{aligned} \tag{9.52}$$

$$\begin{aligned}\dot{R}^v_{p-D}(a,t) &= -\gamma R^v_{p-D}(a,t) + \Psi f_D^{p-} T^v_{p-}(a,t) - r_D R^v_{p-D}(a,t) - p(a)q\epsilon\lambda(a,t)R^v_{p-D}(a,t) - \\ &- p(a)q\epsilon\lambda(a,t)R^v_{p+D}(a,t) - (R^v_{p-D}(a,t) - (1-\delta(a))R^v_{p-D}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{p-D}(a,t)/N(a,t) \end{aligned} \tag{9.53}$$

$$\begin{aligned}\dot{R}^v_{npD}(a,t) &= -\gamma R^v_{npD}(a,t) + \Psi f_D^{p-} T^v_{np}(a,t) - r_D R^v_{npD}(a,t) - p(a)q\epsilon\lambda(a,t)R^d_{npD}(a,t) - \\ &- (R^v_{npD}(a,t) - (1-\delta(a))R^v_{npD}(a-1,t))/\tau + \\ &+ \Delta_N(a,t)R^v_{npD}(a,t)/N(a,t) \end{aligned} \tag{9.54}$$

sdvsdcdscd

where $\delta(a)$ stands for the Kronecker delta function, and the four quantities that depend on time are the force of infection $\lambda(a,t)$, the proportion of mothers with TB $m_d(t)$, the diagnosis rate $d(t)$ and the correction terms $\Delta_N(a,t)$ standing for any demographic variation in the population due to causes foreign to TB and aging. The force of infection is constructed like follows:

$$\lambda(a,t) = \beta_o e^{\alpha_\beta t} \sum_{a'} \xi(a,a') \Upsilon(a',t) \tag{9.55}$$

where $\Upsilon(a',t)$ stands for the weighted sum of infectious individuals:

$$\begin{aligned} \Upsilon(a',t) = D_{p+}(a',t) + F(a',t) + \phi_{p-} D_{p-}(a',t) + \phi_D R_{p+D}(a',t) + \phi_{p-}\phi_D R_{p-D} + \\ + D^v_{p+}(a',t) + \phi_{p-} D^v_{p-}(a',t) + \phi_D R^v_{p+D}(a',t) + \phi_{p-}\phi_D R^v_{p-D}(a',t) \end{aligned} \tag{9.56}$$

Additionally, the proportion of mothers who are responsible of mother-child disease transmission reads as:

$$\begin{aligned} m_d(t) = \frac{\sum_{a=3}^{a=7} D_{p+}(a,t) + D_{p-}(a,t) + D_{np}(a,t) + R_{p+D}(a,t) + R_{p-D}(a,t)}{\sum_{a=3}^{a=7} N(a,t)} + \\ + \frac{\sum_{a=3}^{a=7} R_{npD}(a,t) + F(a,t) + D^v_{p+}(a,t) + D^v_{p-}(a,t)}{\sum_{a=3}^{a=7} N(a,t)} + \\ + \frac{\sum_{a=3}^{a=7} D^v_{np}(a,t) + R^v_{p+D}(a,t) + R^v_{p-D}(a,t) + R^v_{npD}(a,t)}{\sum_{a=3}^{a=7} N(a,t)} \end{aligned} \tag{9.57}$$

Instead, we have, for the diagnosis rate:

$$d(t) = d_o e^{(\alpha_d t)} \tag{9.58}$$

And finally, the demographic correction terms $\Delta_N(a,t)$ obeys the following expression:

$$\begin{aligned} \Delta_N(a,t) = \dot{N}_{UN}(a,t) - [((1-\delta(a))N(a-1,t) - N(a,t))/\Delta_t - \\ - \mu_{p+}(D_{p+}(a,t) + F(a,t) + D^v_{p+}(a,t)) - \\ - \mu_{p-}(D_{p-}(a,t) + D^v_{p-}(a,t)) - \mu_{np}(D_{np}(a,t) + D^v_{np}(a,t)) - \\ - \Psi \hat{f}_\mu (T_{p+}(a,t) + T^v_{p+}(a,t) + T_{p-}(a,t) + T^v_{p-}(a,t) + T_{np}(a,t) + T^v_{np}(a,t))] \end{aligned} \tag{9.59}$$

It is interesting to define two additional variables, fully dependent on the dynamical state of the system, such as the accumulated number of TB incident cases in each age group, from the beginning of the period under analysis $I(a,t)$, and the accumulated number of TB deaths equally defined $M(a,t)$. Their respective temporal evolution reads as follows:

$$\begin{aligned}\dot{I}(a,t) &= d(t)(D_{p+}(a,t) + D^v_{p+}(a,t) + \\ &+ \eta(D_{p-}(a,t) + D^v_{p-}(a,t) + D_{np}(a,t) + D^v_{np}(a,t))\end{aligned} \tag{9.60}$$

$$\begin{aligned}\dot{M}(a,t) &= \mu_{p+}(D_{p+}(a,t) + F(a,t) + D^v_{p+}(a,t)) + \mu_{p-}(D_{p-}(a,t) + D^v_{p-}(a,t)) + \\ &+ \mu_{np}(D_{np}(a,t) + D^v_{np}(a,t)) + \Psi f^{p+}_{\mu}(T_{p+}(a,t) + T^v_{p+}(a,t)) + \\ &+ \Psi f^{p-}_{\mu}(T_{p-}(a,t) + T^v_{p-}(a,t) + T_{np}(a,t) + T^v_{np}(a,t))\end{aligned} \tag{9.61}$$

Because of that, from these variables, once summed over all age groups, we explicitly get the incidence rate as the number of new cases per year $i(t)$, and the annual mortality rate as the total number of TB deaths per year $m(t)$, both normalized by 100.000 individuals:

$$i(t) = \frac{100.000 \cdot \sum_a (I(a,t+1) - I(a,t))}{(\mathcal{N}(t+1) + \mathcal{N}(t))/2} \tag{9.62}$$

$$m(t) = \frac{100.000 \cdot \sum_a (M(a,t+1) - M(a,t))}{(\mathcal{N}(t+1) + \mathcal{N}(t))/2} \tag{9.63}$$

The sums of $I(a,t)$ and $M(a,t)$ over all ages at the end of the period under study provide the total number of cases and deaths due to the disease during the whole period.

Initial conditions setup

Once we have detailed the forces driving the time evolution of our state variables, it remains pendent answer how do we set the initial conditions $X(a, t = 0)$ for each possible state X. This problem is traditionally solved [422, 22] just by considering that, at the beginning of the period analyzed, the system is at the stationary that is reached after blocking the temporal evolution of the time dependent parameters to their values at the beginning of the period: $d(t = 0) = d_o$ and $\beta(t = 0) = \beta_o$, as well as the demographic boundary conditions $\vec{N}(a,t) = \vec{N}(a,0)$, where $\vec{N}(a,t)$ represents the vector of the different populations at each age group. We denote those stationary levels as $X^*(a, d_o, \beta_o, \vec{N}(a,0))$, so we have $\dot{X}^*(a, d_o, \beta_o, \vec{N}(a,0)) = 0$ (provided that the time "freezing" of time-dependent parameters and the demography is acting), and this approach consists in using them to set up the initial conditions of the system: $X(a,0) = X^*(a, d_o, \beta_o, \vec{N}(a,0))$ for any dynamic state X. From this point on, it is worth noticing that no variation in the dynamic state of the system would be achieved if the parameters d and β as well as the demographic pyramid remains constant; because of, in this point, the dynamical system has reached a stationary state. When this kind of approach is used, instead, the only way of modifying this stationary state is to

trigger both the temporal evolution of $d(t)$ and $\beta(t)$ and/or the temporal evolution of the demographic constraints $\vec{N}(a,t)$, which is precisely what it is done so as to start the dynamical description during the period under analysis.

In this work we also abandon this initial stationarity assumption and we do not impose that the system must lye at the stationary in $t = 0$. Instead of that, we calculate the stationary values of all states $X^*(a, d_o, \beta_o, \vec{N}(a,0))$, and we set up an initial state that can correspond either to higher or lower levels of disease prevalence. So as to map these possible variations on TB burden from the stationary vector of states $\vec{X}^*$, we distinguish the unexposed state $S(a,t)$, from the rest of the states in the non immune branch joined by individuals that, at least, have been infected with the bacillus once. Additionally, we consider that, at the immune branch $X^v(a, t = 0) = 0$, (also $S^w(a, t = 0) = 0$), as any vaccination campaign is introduced for some time $t > 0$. Additionally, we define a parameter $\zeta \in [-1, 1]$. In case $\zeta < 0$, it means that the initial conditions correspond to a state with lower TB burden than the stationary, which we build as follows:

$$X(a, t = 0) = (1 + \zeta)X^*(a, d_o, \beta_o, \vec{N}(a,0)) \qquad \forall (X \neq S) \tag{9.64}$$

$$S(a, t = 0) = S^*(a, d_o, \beta_o, \vec{N}(a,0)) \left(1 - \zeta \sum_{X \neq S} X^*(a, d_o, \beta_o, \vec{N}(a,0))\right) \tag{9.65}$$

Instead, if $\zeta > 0$, the initial conditions are displaced up to higher burden levels from the stationary:

$$S(a, t = 0) = S^*(a, d_o, \beta_o, \vec{N}(a,0))(1 - \zeta) \tag{9.66}$$

$$X(a, t = 0) = X^*(a, d_o, \beta_o, \vec{N}(a,0)) \left(1 + \frac{\zeta S^*(a, d_o, \beta_o, \vec{N}(a,0))}{\sum_{X \neq S} X^*(a, d_o, \beta_o, \vec{N}(a,0))}\right) \tag{9.67}$$

9.3.2 Fitting procedure

Fitting scheme

Our model is based in the determination of the initial distance of the system from the stationary ζ, the diagnosis rate $d(t)$ and the scaled infectivity $\beta(t)$, for each region that are adequate to make the model reproduce certain disease burden measurements on a certain time window which we take as an input. The last two time dependent parameters are supposed to vary exponentially: $d(t) = d_o e^{(\alpha_d t)}$ and $\beta(t) = \beta_o e^{(\alpha_\beta t)}$, which means that the total amount to parameters to fit is five: $(\zeta, d_o, \beta_o, \alpha_d, \alpha_\beta)$

The goal of the fitting procedure is to minimize the overall error H of the model outcome with respect to the input burden measurements, calculated as follows:

$$H = \sum_{t=t_o}^{t_F} \left(\frac{i(t) - \bar{i}(t)}{\bar{\sigma}_i(t)} \right)^2 + \left(\frac{m(t) - \bar{m}(t)}{\bar{\sigma}_m(t)} \right)^2 \tag{9.68}$$

where the fitting window spans from t_o to t_F (in years); and $\bar{i}(t)$ and $\bar{m}(t)$ stand for the annual incidence and mortality rates, calculated by grouping the national estimations available at the the WHO database for TB research [24], corresponding to the different countries within each region. These measurements of TB incidence and mortality have their correspondent confidence intervals $(\bar{i}_{low}(t), \bar{i}_{high}(t))$ and $(\bar{m}_{low}(t), \bar{m}_{high}(t))$, which are not necessarily symmetrical with respect to the central values $\bar{i}(t)$ and $\bar{m}(t)$. Using these confidence intervals, and taking into consideration their asymmetry, the correspondent terms $\bar{\sigma}_i(t)$ $\bar{\sigma}_m(t)$ are constructed like follows:

$$\bar{\sigma}_i(t) = \begin{cases} \bar{i}(t) - \bar{i}_{low}(t) & \text{if } i(t) \leq \bar{i}(t) \\ \bar{i}_{high}(t) - \bar{i}(t) & \text{if } i(t) > \bar{i}(t) \end{cases} \tag{9.69}$$

$$\bar{\sigma}_m(t) = \begin{cases} \bar{m}(t) - \bar{m}_{low}(t) & \text{if } m(t) \leq \bar{m}(t) \\ \bar{m}_{high}(t) - \bar{m}(t) & \text{if } m(t) > \bar{m}(t) \end{cases} \tag{9.70}$$

The conceptual scheme for the fitting of these parameters (see figure 9.4) essentially consists in an iterative evaluation of the model across the parameter space $(\zeta, d_o, \beta_o, \alpha_d, \alpha_\beta)$, which is navigated according to a certain "routing" that eventually guarantees the localization of a certain parameter set that yields an error H which is small enough. In our case, we have used a Levemberg-Marquardt algorithm [413], implemented, as for the rest of the model, in programming language C.

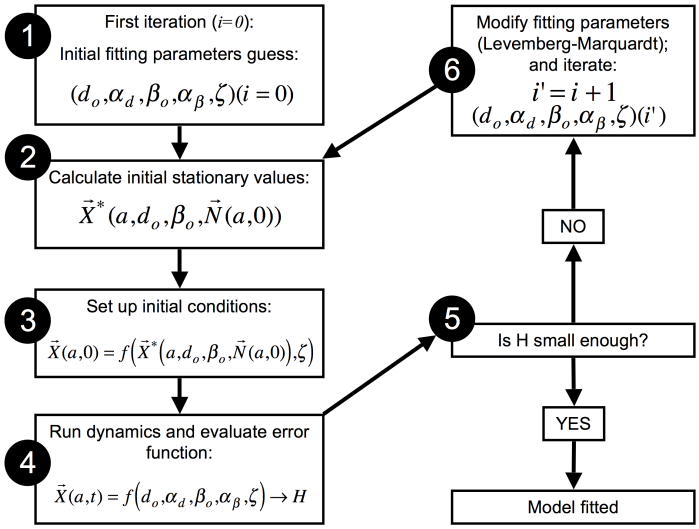


FIGURE 9.4: Schematic representation of the fitting procedure.

The temporal window chosen for the fitting spans from 2001 to 2011, for which burden data is available in [24]. The fitting window could be arbitrarily extended backwards in time -even beyond the limits of the registers of the database [24], if the appropriate data sources or estimations were used-, but the exercise would lose sense as the window grew and grew. The reason is that the values of the parameters fitted reflect the socio-economic situation of the populations under study, as well as the capacity of the health systems to contain the epidemics and their paces of improvement; all these factors quantified during the time window of the fitting. For this reason, the eventual extension of the time window to the past is not a convenient strategy, because, as we go back in the past, all these socio-economic factors diverge from what they are in the current times and ultimately, to intend that their values were representative of what it is expected to happen in the next decades is less and less reasonable. In this sense, the choose of a ten years time window is a convention that takes into account the compromise between a window short enough to represent a period stable enough from a socio-economic point of view in most countries, and a period large enough to offer a reasonable agreement between number of data and number of tunable parameters.

Disease burden estimations

The reproduction by our model of the temporal evolution of the variables $\bar{i}(t)$ and $\bar{m}(t)$ represent the main goal of the fitting procedure. However, a problem regarding the association of TB disease and HIV infection and the way in which these magnitudes are registered at the World Health Organization database for TB [24]. Indeed, the cases of TB+HIV confection are currently accounted as TB cases when speaking about incidence in the database, but, when these patients dead, the event is not accounted as a TB related death, as, by convention, these deaths are attributed to HIV and so, WHO avoids registering the same death twice across its databases.

Taking that into account, so as to calculate $\bar{m}(t)$, we are constricted to perform the sum $\bar{m}(t) = \bar{m}_{exc.TB+HIV}(t) + \bar{m}_{TB+HIV}(t)$. Indeed the contribution of the cases of TB excluding coinfection, $\bar{m}_{exc.TB+HIV}(t)$ is explicitly available at [24], while the estimation of $\bar{m}_{TB+HIV}(t)$ is more concerned. Indeed, the estimation of the relative mortality rates among HIV+ positive individuals by TB, compared to HIV uninfected individuals is not trivial. In a recent work, Au-Yeung et al. [414] estimate the quotient $x(t) = \bar{m}_{TB+HIV}(t)/\bar{m}_{exc.TB+HIV}(t)$ for the different WHO-regions, also used in this chapter, in the period 2006-2008. Taking 2007 as reference, we can complement this information with the data about HIV+TB confection occurrence among incident TB, $TB + HIV(t)$, available at [24], to reach the following final estimation of $\bar{m}(t)$:

$$\bar{m}(t) = \bar{m}(t)_{exc.TB+HIV} \left(1 + x(2007)\frac{HIV(t)}{HIV(2007)}\right) \tag{9.71}$$

Leaving apart the well known fact that HIV status has a critical influence on individuals epidemic risk related to TB at many stages of its natural history, our aim is not to describe these dynamical differences or to evaluate the differential risks due to HIV infection of a certain segment of the populations under study. Instead of that, we just

provide an "average" description of TB burden, as we do not explicitly differentiate, as a first approximation, the dynamics of HIV+ and HIV- individuals. However, this does not make a reasonable approach to consider the incidence of new active cases of TB associated to TB+HIV coinfection and, simultaneously, to neglect the deaths occurred as a result of that coinfection, and by this reason we are constricted to introduce this additional contribution to $\bar{m}(t)$ due to HIV-TB coinfection.

In figure 9.5, we represent, for each region, the contribution to $\bar{m}(t)$ of deaths due to TB alone and to TB+HIV coinfection. As it is can be seen from there, TB+HIV coinfection supposes a dramatically high fraction of all deaths related to TB in AFRH region, which is nonetheless the region with highest levels of HIV prevalence in the world. Errorbars in $\bar{m}(t)$ has been propagated from equation 9.71 from the uncertainties of $\bar{m}(t)_{exc.TB+HIV}$, $x(2007)$ and the quotient $HIV(t)/HIV(2007)$

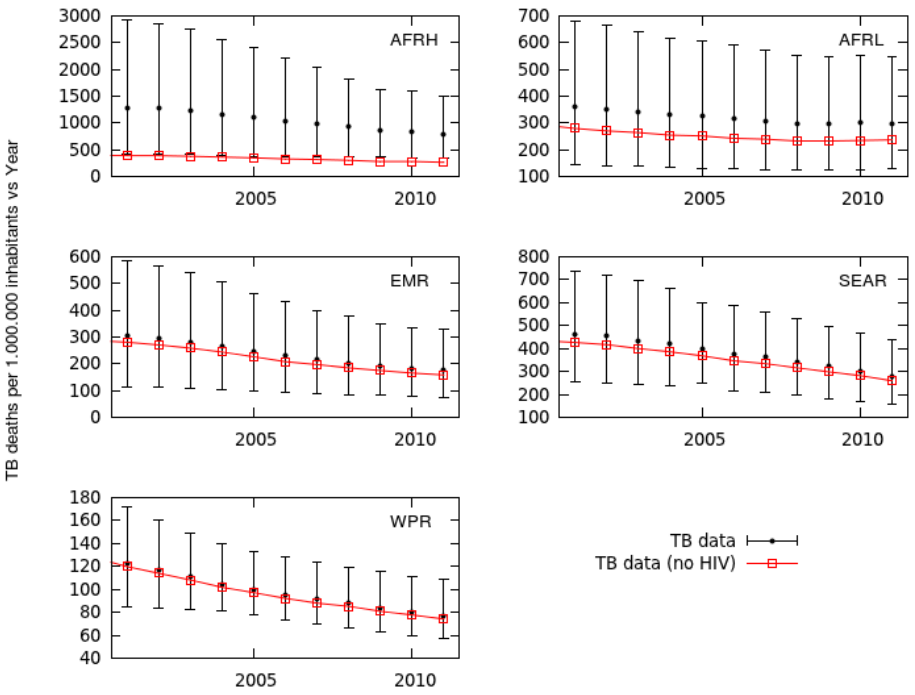


FIGURE 9.5: Red: $\bar{m}_{exc-TB+HIV}(t)$ Mortality rates excluding TB plus HIV confection. Blue: total mortality rates $\bar{m}(t)$

9.3.3 Model states and parameters summary

In this section we list all the dynamical states and parameters used in the model, along with their values, definitions, confidence intervals and bibliographical sources:

Dynamic states

In table 9.2 all dynamic states of the model are detailed. It is worth emphasizing that X^v states are described as "immune" states, which means that, after being vaccinated, individuals in this second branch have a partial immunity to infection that translates into a reduced probability of getting infected or re-infected, regardless they were previously infected or not. In contrast, states labeled as non-immune contains individuals that not have been vaccinated ever and also those that, despite having received the vaccine, have lost its immunogenic effect.

State	Definition
$S(a,t)$	Susceptible (not previously exposed to infection) unvaccinated individuals
$S^v(a,t)$	Susceptible (not previously exposed to infection) vaccinated individuals retaining vaccine's (partial) immunity
$S^w(a,t)$	Susceptible (not previously exposed to infection) vaccinated individuals whose immunity has waned
$L_s(a,t)$	Non-immune Infected individuals (slow latency)
$L_f(a,t)$	Non-immune Infected individuals who will develop primo-infection (fast latency)
$L_s^v(a,t)$	Immune Infected individuals (slow latency)
$L_f^v(a,t)$	Immune Infected individuals who will develop primo-infection (fast latency)
$D_{ps+}(a,t)$	Non-immune untreated sick individuals: Smear positive pulmonar disease
$D_{ps-}(a,t)$	Non-immune untreated sick individuals: Smear negative pulmonar disease
$D_{np}(a,t)$	Non-immune untreated sick individuals: non pulmonar disease
$D_{ps+}^v(a,t)$	Immune untreated sick individuals: Smear positive pulmonar disease
$D_{ps-}^v(a,t)$	Immune untreated sick individuals: Smear negative pulmonar disease
$D_{np}^v(a,t)$	Immune untreated sick individuals: non pulmonar disease
$T_{ps+}(a,t)$	Immune sick individuals under treatment: Smear positive pulmonar disease
$T_{ps-}(a,t)$	Non-immune sick individuals under treatment: Smear negative pulmonar disease
$T_{np}(a,t)$	Non-immune sick individuals under treatment: non pulmonar disease
$T_{ps+}^v(a,t)$	Immune sick individuals under treatment: Smear positive pulmonar disease
$T_{ps-}^v(a,t)$	Immune sick individuals under treatment: Smear negative pulmonar disease
$T_{np}^v(a,t)$	Immune sick individuals under treatment: non pulmonar disease
$F(a,t)$	Patients who faultily finished their treatment.
$R_{p+S}(a,t)$	Non-immune patients of smear positive pulmonary TB who successfully finished their treatment.
$R_{p+D}(a,t)$	Non-immune patients of smear positive pulmonary TB who defaulted their treatment by two consecutive months or more.
$R_{p+N}(a,t)$	Non-immune patients of smear positive pulmonary TB that naturally –i.e. without treatment– recovered from the disease.
$R_{p+S}^v(a,t)$	Immune patients of smear positive pulmonary TB who successfully finished their treatment.
$R_{p+D}^v(a,t)$	Immune patients of smear positive pulmonary TB who defaulted their treatment by two consecutive months or more.

State	Definition
$R^v_{p+N}(a,t)$	Immune patients of smear positive pulmonary TB that naturally –i.e. without treatment– recovered from the disease.
$R_{p-S}(a,t)$	Non-immune patients of smear negative pulmonary TB who successfully finished their treatment.
$R_{p-D}(a,t)$	Non-immune patients of smear negative pulmonary TB who defaulted their treatment by two consecutive months or more.
$R_{p-N}(a,t)$	Non-immune patients of smear negative pulmonary TB that naturally –i.e. without treatment– recovered from the disease.
$R^v_{p-S}(a,t)$	Immune patients of smear negative pulmonary TB who successfully finished their treatment.
$R^v_{p-D}(a,t)$	Immune patients of smear negative pulmonary TB who defaulted their treatment by two consecutive months or more.
$R^v_{p-N}(a,t)$	Immune patients of smear negative pulmonary TB that naturally –i.e. without treatment– recovered from the disease.
$R_{npS}(a,t)$	Non-immune patients of non pulmonary TB who successfully finished their treatment.
$R_{npD}(a,t)$	Non-immune patients of non pulmonary TB who defaulted their treatment by two consecutive months or more.
$R_{npN}(a,t)$	Non-immune patients of non pulmonary TB that naturally –i.e. without treatment– recovered from the disease.
$R^v_{npS}(a,t)$	Immune patients of non pulmonary TB who successfully finished their treatment.
$R^v_{npD}(a,t)$	Immune patients of non pulmonary TB who defaulted their treatment by two consecutive months or more.
$R^v_{npN}(a,t)$	Immune patients of non pulmonary TB who naturally –i.e. without treatment– recovered from the disease.
$N(a,t)$	Total number of individuals in age group a
$I(a,t)$	Accumulated number of detected TB cases in age group a from the beginning of the period.
$M(a,t)$	Accumulated number of TB deaths in age group a from the beginning of the period.
$i(t)$	Tuberculosis incidence rate per 100000 individuals.
$m(t)$	Tuberculosis mortality rate per 100000 individuals.

TABLE 9.2: List of all the dynamic variables contemplated in our model.

Global parameters

Regional parameters

Region dependent parameters are the following

- η: Coefficient of modification of the diagnosis rate modification for smear positive and extrapulmunoray TB types with respect to smear possitive [22]

Meaning	Parameter	Value	Confidence interval	Reference
Primo-infection fraction	p(a)	0.050 (children)	(0.043-0.058)	[22]
		0.150 (adults)	(0.128-0.173)	[22]
Fast progression rate to smear + TB (y^{-1})	$\omega_{fp+}(a)$	0.090 (children)	(0.071-0.109)	[22]
		0.450 (adults)	(0.355-0.545)	[22]
Fast progression rate to smear - TB (y^{-1})	$\omega_{fp-}(a)$	0.585 (children)	(0.461-0.709)	[22]
		0.360 (adults)	(0.284-0.436)	[22]
Fast progression rate to non-pulmonary TB (y^{-1})	$\omega_{fnp}(a)$	0.225 (children)	(0.177-0.273)	[22]
		0.090 (adults)	(0.071-0.109)	[22]
Slow progression rate to smear + TB (y^{-1})	$\omega_{sp+}(a)$	$7.50 \cdot 10^{-5}$ (children)	$(5.91 - 9.09) \cdot 10^{-5}$	[22]
		$3.75 \cdot 10^{-4}$ (adults)	$(2.95 - 4.55) \cdot 10^{-4}$	[22]
Slow progression rate to smear - TB (y^{-1})	$\omega_{sp-}(a)$	$4.88 \cdot 10^{-4}$ (children)	$(3.84 - 5.91) \cdot 10^{-4}$	[22]
		$3.00 \cdot 10^{-4}$ (adults)	$(2.36 - 3.64) \cdot 10^{-4}$	[22]
Slow progression rate to non-pulmonary TB (y^{-1})	$\omega_{snp}(a)$	$1.88 \cdot 10^{-4}$ (children)	$(1.48 - 2.27) \cdot 10^{-4}$	[22]
		$7.50 \cdot 10^{-5}$ (adults)	$(5.91 - 9.09) \cdot 10^{-5}$	[22]
Mortality rate by pulmonary smear + TB (y^{-1})	μ_{p+}	0.250	(0.213-0.288)	[22]
Mortality rate by pulmonary smear - TB (y^{-1})	μ_{p-}	0.100	(0.085-0.115)	[22]
Mortality rate by non-pulmonary TB (y^{-1})	μ_{np}	0.100	(0.085-0.115)	[22]
Re-infection vs. first infection relative risk	q	0.650	(0.553-748)	[22]
Treatment completion rate (y^{-1})	Ψ	2.00	(1.70-2.30)	[22]
Smear progression rate (y^{-1})	θ	0.015	(0.007-0.020)	[422]
Relapse rate after treatment success (y^{-1})	r_S	$9.392 \cdot 10^{-4}$	$(6.364 - 12.45) \cdot 10^{-4}$	[96], this Thesis
Relapse rate after treatment default (y^{-1})	r_D	$3.774 \cdot 10^{-3}$	$(1.354 - 8.620) \cdot 10^{-3}$	[96], [423],this Thesis
Relapse rate for naturally recovered individuals (y^{-1})	r_N	0.030	(0.020-0.040)	[422]
Natural recovery rate (y^{-1})	ν	0.100	(0.085-0.115)	[422]
Infectivity reduction of D_{p-} with respect to D_{p+}	ϕ_{p-}	0.250	(0.213-0.288)	[22]
Infectivity reduction of R_{p+D} with respect to D_{p+}	ϕ_D	0.500	(0.250-0.750)	[422]
Proportion of mothers infecting their newborns	m_t	0.150	(0.100-0.200)	[424]

TABLE 9.3: List of global parameters whose values are considered not to depend on the geographic area being modeled

Region	η	f_D^{p+}	f_F^{p+}	f_μ^{p+}
AFRH	0.843 (0.717-0.970)	0.081(0.045-0.118)	0.016 (0.006-0.026)	0.063(0.023-0.085)
AFRL	0.510 (0.434-0.587)	0.099 (0.044-0.154)	0.024 (0.013-0.035)	0.061 (0.048-0.074)
EMR	1.178 (1.001-1.354)	0.067 (0.038-0.095)	0.016 (0.013-0.018)	0.036 (0.033-0.039)
SEAR	0.797 (0.677-0.916)	0.061 (0.040-0.077)	0.019 (0.015-0.021)	0.042 (0.032-0.049)
WPR	0.641 (0.545-0.737)	0.013 (0.008-0.018)	0.009 (0.006-0.011)	0.017 (0.014-0.020)

Region	f_D^{p-}	f_F^{p-}	f_μ^{p-}
AFRH	0.079 (0.042-0.116)	0.003 (0.001-0.006)	0.087 (0.054-0.120)
AFRL	0.076 (0.044-0.108)	0.004 (0.001-0.007)	0.096 (0.055-0.137)
EMR	0.076 (0.046-0.107)	0.004 (0.001-0.007)	0.033 (0.024-0.042)
SEAR	0.062 (0.047-0.075)	0.005 (0.002-0.009)	0.034 (0.032-0.035)
WPR	0.017 (0.008-0.026)	0.002 (0.002-0.003)	0.014 (0.011-0.017)

TABLE 9.4: Regional parameters in each region

- $(f_D^{p+}, f_F^{p+}, f_\mu^{p+})$: fraction of default, failure and death outcomes for smear positive pulmonary TB [24]
- $(f_D^{p-}, f_F^{p-}, f_\mu^{p-})$: fraction of default, failure and death outcomes for smear negative pulmonary and non pulmonary TB [24]

and their values, for each of the regions analyzed are listed in table 9.3.3:

Other inputs for the model, that are not formally model parameters but that also has to be determined for the model to work are the following:

- $N_{UN}(a,t)$: Demographic pyramid projected for each region by the UN population division [23]
- $\bar{i}(t)$, $\bar{m}(t)$: incidence and mortality rates, estimated by the WHO [24]

Fitted parameters

Once fixed all the mentioned parameters, we have proceeded to fit the diagnosis rate $d(t) = d_o e^{\alpha_d t}$ and the base infectivity $\beta(t) = \beta_o e^{\alpha_\beta t}$ that reproduces more accurately the burden time series of incidence and mortality estimated by the WHO. The results, in each of the regions analysed are listed in table 9.5

TABLE 9.5: Fitted parameters of the model.

Region	$d_0\,(y^{-1})$	α_d	$\beta_0\,(y^{-1})$	α_β	ς
AFRH	0.281 (0.037-1.007)	0.082 (0.034-0.101)	5.330 (3.481-10.056)	-0.009 (-0.030–0.005)	0.136 (-0.032-0.425)
AFRL	0.965 (0.354-3.085)	0.043 (0.039-0.055)	18.972 (11.371-41.478)	-0.012 (-0.016–0.007)	0.051 (0.018-0.104)
EMR	0.379 (0.126-1.328)	0.099 (0.078-0.131)	4.896 (3.215-9.915)	0.013 (0.005-0.024)	0.130 (0.056-0.208)
SEAR	0.679 (0.342-1.322)	0.073 (0.069-0.087)	1.962 (1.340-3.042)	0.003 (-0.004-0.013)	0.053 (0.006-0.127)
WPR	3.034 (0.981-3.560)	0.038 (0.032-0.039)	4.245 (3.215-5.418)	0.001 (-0.006-0.009)	0.132 (0.092-0.190)

Strategy	$f\,(a)\ \ a \in [0, 13]$	f_{neo}	ϵ
Mass vaccination	$1\,y^{-1}\ \forall a$	0	$0.2\ \forall a$
Children-focused campaign	$1\,y^{-1}\ a = 0$ $0\ a \neq 0$	0	$0.2\ \forall a$
Adolescent-focused campaign	$1\,y^{-1}\ a = 3$ $0\ a \neq 3$	0	$0.2\ \forall a$
Newborn-focused campaign	$0\ \forall a$	1	$0.2\ \forall a$
Newborn+Teenager mixed campaign	$1\,y^{-1}\ a = 3$ $0\ a \neq 3$	1	0.2 if $a < 3$ $\bar{\epsilon}\,(t_{vac}, a)$ if $a \geq 3$

TABLE 9.6: Vaccine description under different immunization strategies

Vaccine descriptors

We have applied different vaccination strategies in this chapter but all of them have the following features in common:

- The year of beggining of the vaccination campaign: 2015
- The immunity waning rate: $\gamma\,(a) = 0.015\,y^{-1} \forall a$

The difference between the different strategies will be given by the dependence of the immunity acquisition rate $f\,(a)$ with the age group, –introduced to describe the age at which individuals are immunized– and the fraction of newborns f_{neo} that are vaccinated. In table 9.4.4 summarizes the different vaccines:

A further description is needed in the case of a combined strategy adolescent+newborn, in which two vaccines are foreseen –a BCG substitutive vaccine applied on newborn and a booster vaccine on adolescents–; as we wish to take into account the cooperation between them. We consider that both vaccines act with the same observed efficacy coefficient $\epsilon = 0.2$, as in the other strategies considered but we take the assumption that individuals receiving both vaccines will gain an enhanced immunity described by an observed efficacy coefficient $\epsilon = 0.2^2 = 0.04$.

Individuals within the first three groups –children less than 15 years old–, will only benefit from the BCG substitutive vaccine of observed efficacy equal to $\epsilon = 0.2$. Instead, for the other age groups, we will eventually have doubly immunized individuals, if enough time from the vaccination campaign start has passed so as the people who

have received the newborn vaccine gets older and receive the adolescent vaccine. The function $\bar{\epsilon}\,(t_{vac}, a)$ gives us the dependence of the effective efficacy coefficient with the age group, a, and the time since the beggining of the campaign, t_{vac}:

$$\bar{\epsilon}\,(t_{vac}, a) = \begin{cases} 0.2 & \text{if } t_{vac} < \Delta_t a \\ 0.2\left(1 - \frac{t_{vac} - \Delta_t a}{\Delta_t}\right) + 0.04\frac{t_{vac} - \Delta_t a}{\Delta_t} & \text{if } \Delta_t a < t_{vac} < \Delta_t\,(a+1) \\ 0.04 & \text{if } t_{vac} > \Delta_t\,(a+1) \end{cases}$$

Contact patterns

Perhaps, one of the most pervasive simplifying hypothesis in the mathematical modelling of TB spreading is the homogenous mixing of the populations under study. In this chapter, we abandon this hypothesis by defining an heterogeneous contact matrix between different age groups denoted as $\xi(a, a')$. The matrix elements $\xi(a, a') \in [0, 1]$ represent the fraction of all possible contacts between individuals within age groups a and a' that actually take place. For example, if we have $N(a, t)$ and $N(a', t)$ in each group at a certain moment, then we count as many as $\xi(a, a')N(a, t)N(a', t)$ between these two groups of individuals.

For the computation of the contact matrix used in our model, we take advantage from the data of the survey performed in the so-called Polymod proyect [12]. In this survey people is asked how many contacts they have had in a certain time window, and the ages of the people who have maintained these contacts with them. Let us denote as $P(a, a')$ the elements of this matrix, which represents the average number of contacts that an individual in the age group a has, per unit time, with people in the age group a'. These matrices $P(a, a')$ are provided in [12] for different countries. In the following, we explain how do we get our contact matrix $\xi(a, a')$ from these Polymod data.

First of all, it is important to note that, for any country, the matrix $P(a, a')$ is not, –in general–, symmetrical, due to the different number of people in each age group. The number of total contacts between two age groups a and a' can be computed in two different ways: if we rely on the reports of class a we would write it as the product $P(a, a')N(a, t)$, where $N(a, t)$ is the number of individuals in age group a at moment t. Instead, if we rely our estimation of the number of contacts on the reports by class a' individuals, we would obtain $P(a', a)N(a', t)$. Obviously, if the survey has enough quality, one should find:

$$P(a, a')N(a, t) \simeq P(a', a)N(a', t) \Rightarrow \frac{P(a, a')}{N(a', t)} \simeq \frac{P(a', a)}{N(a, t)} \tag{9.72}$$

So, if t_s is the moment in which surveys took place, we can construct the following matrix, with the same units of $\xi(a, a')$ (i.e. fraction of total contacts between age groups a and a' per unit time):

$$A(a, a') = \frac{P(a, a')}{N(a', t_s)} \tag{9.73}$$

which is not perfectly symmetrical. Indeed, what we do is consider the expected value of the fraction of total contacts as its symmetric part, denoted as $B(a,a')$:

$$B(a,a') = B(a',a) = \frac{A(a,a') + A(a',a)}{2} \tag{9.74}$$

And use the asymmetric part, (which is the measure of the incoherence between the reports of age groups a and a' when referring the number of contacts between them) as an estimation of the error:

$$\delta B(a,a') = \delta B(a',a) = \frac{1}{\sqrt{2}}\sqrt{\frac{\left(A(a,a') - B(a,a')\right)^2 + \left(A(a',a) - B(a,a')\right)^2}{2}} \tag{9.75}$$

Polymod project collected and rendered publicly available [12] the matrices $P(a,a')$ for eight different european countries, as resulted from the surveys performed between 2005 and 2006. The mixing patterns are highly coherent among the eight countries analysed, which allows us to build the following average, which we assimilate to $\xi(a,a')$:

$$\xi(a,a') = \frac{\sum_c \left[B(a,a')\right]_c \left[N(a,t_s)\right]_c \left[N(a',t_s)\right]_c}{\sum_c \left[N(a,t_s)\right]_c \left[N(a',t_s)\right]_c} \tag{9.76}$$

where c is an index that indicates the country. Finally, we consider $B(a,a')$ as the only source of error:

$$\begin{aligned}\delta\xi(a,a') &= \sqrt{\sum_c \left(\frac{\partial\xi(a,a')}{\partial\left[B(a,a')\right]_c}\left[\delta B(a,a')\right]_c\right)^2} = \\ &= \frac{\sqrt{\sum_c \left(\left[\delta B(a,a')\right]_c \left[N(a,t_s)\right]_c \left[N(a',t_s)\right]_c\right)^2}}{\sum_c \left[N(a,t_s)\right]_c \left[N(a',t_s)\right]_c}\end{aligned} \tag{9.77}$$

So, our hypothesis is that we can use the matrix $\xi(a,a')$ (see figure 9.3.3) to undertake a first approximation to the description of the mixing patterns heterogeneity in any part of the world. In relation to this, it is worth noticing different aspects of our approximation. In the one hand, by scaling the original matrices $P(a,a')$ towards the final matrix $\xi(a,a')$ using the demographic pyramids of the countries surveyed $N(a,t)$ at $t = t_s$ (available at [23]), we eliminate the influence of the demographic structure of these countries on the mixing matrix, and so, we come up with an object that is useful to describe the age dependent mixing patterns regardless its specific demographical details. In conclusion, the fact that other regions have different demographic structures different from Europe impose no limitation to the generalisation of $\xi(a,a')$ out from Europe. In the other hand, the overall intensity of the mixing terms is modulated by the common factor $\beta(t)$, which is fitted independently for each region: this is of utmost relevance, given that the average strength and frequency of social interactions is, undoubtedly, highly dependent on cultural and socio-economic factors.

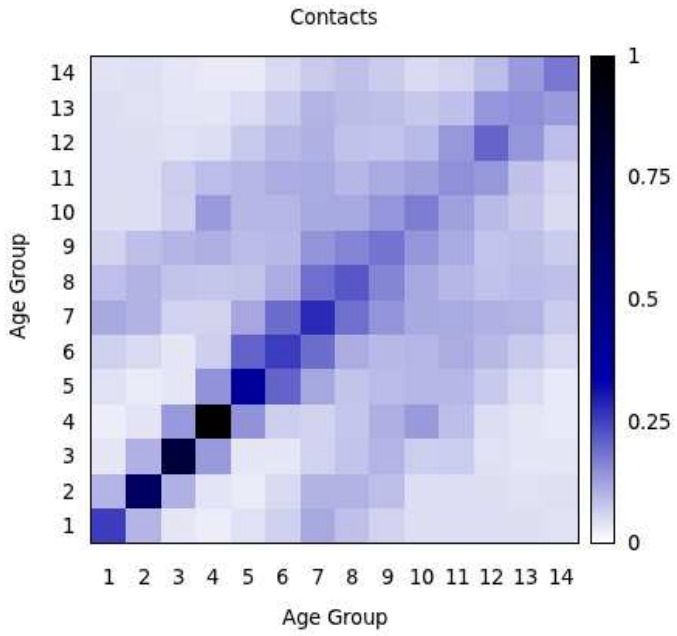


FIGURE 9.6: Normalized contact matrix $\xi(a, a')/\xi(a, a')_{max}$

In any case, even if we account for eventual variations on the overall contact strength by fitting the parameter $\beta(t)$, it is also obviously true that the structure of the specific mixing patterns in regions foreign to Europe could present differential variations among the different fields of the matrices with respect to the european case calculated as $\xi(a, a')$. Our hypothesis is that, however, the relevance of these differential variations might not be very relevant, as the main features of the contact matrixes are stable from a country to another and obey to very general characteristics of our society that can be very reasonably extrapolated to any part of the world. These are 1. a strong trend of people to be in contact with people of their same age 2. the presence of more frequent and intense contacts during the childhood and 3. the appearance of inter-generational contacts stablished between parents and children (or grandparents and grandchildren).

In this sense, and recalling again the fact that the average strength of the contact is modulated independently in each region via $\beta(t)$, our usage of $\xi(a, a')$ in all regions as the age dependent contact matrix, is a first approximation to the description of heterogeneous contact patterns and its influence of model outcomes, which is, as we shown in the main text, of quantitative relevance.

9.3.4 Regions

The definition of the regions used in this study have been taken from [22]. We have analyzed the five regions of highest TB burden levels worldwide. These are the countries that form each of the regions analysed:

- AFRH: Africa, high HIV prevalence: Botswana, Burundi, Cameroon, Central African Republic, Congo, Cote d'Ivoire, Democratic Republic of the Congo, Ethiopia, Gabon, Kenya, Lesotho, Malawi, Mozambique, Namibia, Nigeria, Rwanda, South Africa, Swaziland, Uganda, United Republic of Tanzania, Zambia, Zimbabwe.

- AFRL: Africa, low HIV prevalence: Algeria, Angola, Benin, Burkina Faso, Cape Verde, Chad, Comoros, Equatorial Guinea, Eritrea, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Madagascar, Mali, Mauritania, Mauritius, Niger, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Togo.

- EMR: Eastern Mediterranean Region: Bahrain, Djibouti, Egypt, Iraq, Islamic Republic of Iran, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Pakistan, Qatar, Saudi Arabia, Somalia, Sudan, Syrian Arab Republic, Tunisia, United Arab Emirates, Occupied Palestinian Territory, Yemen

- SEAR: South-Southeast Asia Region: Bangladesh, Bhutan, India, Democratric People's Republic of Korea, Indonesia, Maldives, Myanmar, Nepal, Sri Lanka, Thailand, Timor-Leste

- WPR Western Pacific Region: Brunei Darussalam, Cambodia, China, Fiji, Lao People's Democratic Republic, Malaysia, Micronesia, Mongolia, New Caledonia, Papua New Guinea, Philippines, Polynesia, Republic of Korea, Solomon Islands, Vanuatu, Viet Nam.

For each region, besides the set of regional parameters, –adapted from [22]–, the demographic pyramid $N(a,t)_{UN}$ has to be constructed by summing up the national data available at the database of the population division of UN [23]. Additionally, the series of measured incidence and mortality rates $\bar{i}(t)$ and $\bar{m}(t)$ has also to be built up, in this case, by summing the national data from the TB database curated by the World Health Organization [24], weighted by the countries populations (and corrected for accounting TB+HIV related deaths). For some regions, however, there exist some issues with the databases, mostly regarding small territories for which not all the data are available as well as countries that have experienced process of division or emancipation:

- Some small countries do not appear in the databases of population by age group in [23]. These are Seychelles, in AFRL; and Anguilla, Antigua and Barbuda, Bermuda, British Virgin Islands, Cayman Islands, Dominica, Montserrat, Saint Kitts and Nevis and Turks and Caicos Islands in AMR. For these countries, –which nonetheless represent around 0.1% of the total population of their regions–,

we have supposed that their total populations –which are, instead, available at [23]– are distributed among the different age groups following the same proportions than the rest of the countries belonging to their regions AFRL and AMR, respectively.

- Some countries are denoted in a different way in the demographic database of population by age of [23] and in the TB database [24]. That is the case of the island territories of Polynesia and Micronesia, that appear under these names at the population by age files in [23] but, either in the files of total population series in [23] or in the WHO TB database, [24], they appear broke down by the countries that belong to each one: American Samoa, Cook Islands, French Polynesia, Niue, Samoa, Tokelau, Tonga, Tuvalu and Wallis and Futuna Islands; which are the Polynesian Islands and Guam, Kiribati, Marshall Islands, Federated States of Micronesia, Nauru, Northern Mariana Islands and Palau, which are the Micronesian ones.

- Some countries have experienced processes of division or independence that –at least in the case of [24]– are reflected in the databases. These are the cases of the independence of Timor-Leste in 2010 (previously it belonged to Indonesia) and the division of the Netherlands Antilles in 2010 (which gives rise to Bonaire, Saint Eustatius and Saba; Sint Maarten –Dutch part– and Curacao).

9.4 Results

9.4.1 Model fitting and generic vaccination impact evaluations

If we focus on the South East Asia Region (SEAR)[24], the figure 9.4.1, panels A-B show the projection of the incidence and death rates associated with the disease up to year 2050. As we can see, the model captures the behavior of the incidence and mortality curves during the fitted window, and foresees the decline of the disease up to circa 2025, moment in which incidence and mortality levels remarkably stop their decay.

The dynamical reason underlying that behavior shift is the evolution of the stationary levels, represented in red in figure 9.4.1, as a function of the time dependent parameters and the demographic pyramid at each time step. As a result of the fitting procedure (see supplementary information), in South east Asia, both time dependent parameters are increasing functions of time. This result, in the case of $\beta(t)$, contributes to enhance disease burden as time goes by, while, in the case of $d(t)$ contributes to reduce it. In addition, the marked process of demographic aging which is expected to take place in the region during the whole XXI century (see supplementary information) has an influence on disease burden levels that is not trivial at all. On the one hand, as the proportion of older individuals grow in the population, the probability of them to have been exposed to infection also increases, which would contribute to enhance disease burden levels. However, the structure and strength of social contacts

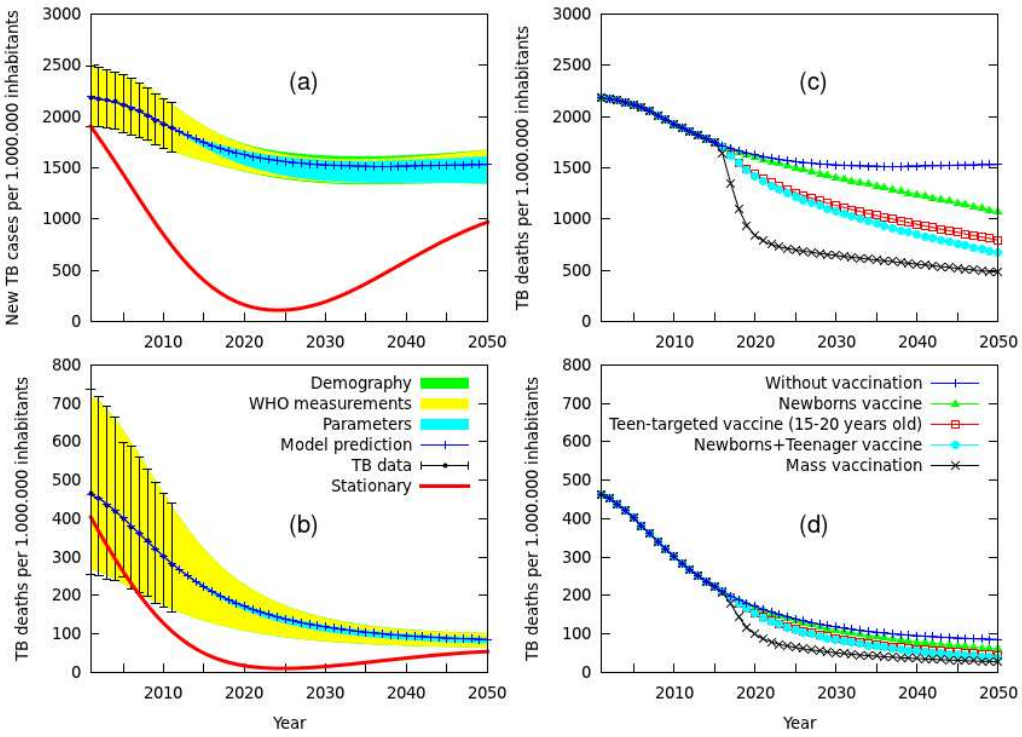


FIGURE 9.7: Model fitting and forecasts in South East Asia region. Panels A,B: Incidence and mortality rates, respectively: blue: model forecasts. Black: WHO estimations during the fitting window (2001-2011). Red: stationary levels at each time step. Error bars in model forecasts are due to the propagated uncertainties of model parameters (yellow) WHO burden and treatment outcomes estimations themselves (red) and demographic data (green). Panels C,D: vaccine impacts foreseen for different vaccination strategies. The vaccines are applied either on newborn individuals, adolescents (15-20 years old), or all the population. Simulations corresponding the vaccination strategy focused simultaneously on newborns and adolescents are designed to describe cooperative effects between both vaccines.

of an individual changes with age, which will also modify her attitudes to transmit the disease to others. A complex interplay of these factors is responsible, in this case, of this worrying behavior according to which the burden levels associated to the stationary attractors of the dynamics show an initial decreasing up to an inversion point that prevents the region to reach the final conditions for disease eradication; provided that the time dependent parameters maintain their time variation rates during the time window analyzed. Another relevant conclusion that is derived from the figure is that stationary values vary in time much faster than burden levels themselves, which show a very strong inertia. This implies that the location of the dynamic state of the system on the stationary may be a more unlikely event than expected, which gives us an additional argument so as to abandon the classical hypothesis of initial stationarity.

In the same figure, panels C and D represent the variation of the same magnitudes that would be achieved after the introduction of a vaccine able to reduce the probability

of infection of immunized individuals up to a 20 percent of the base case (80% efficacy) applied according to vaccination strategies focused on different age groups. Although the long term levels of disease burden are similar among all the strategies considered, a much faster fall in either incidence or mortality rates is achieved after a mass vaccination campaign on all age groups, rather than any alternative strategy, as expected. The combined vaccination campaign on newborns and adolescents is designed to simulate the implementation of two cooperative vaccines, one substitutive vaccine to be applied on newborns instead of BCG (newborn-focused vaccine, or NFV in what follows) and a booster vaccine to be administrated on adolescents regardless their primary vaccine was BCG or its novel substitute (adolescents-focused vaccine, AFV). In figure 9.4.1 the accumulated impact of each vaccination campaign is represented as a function of the vaccine efficacy.

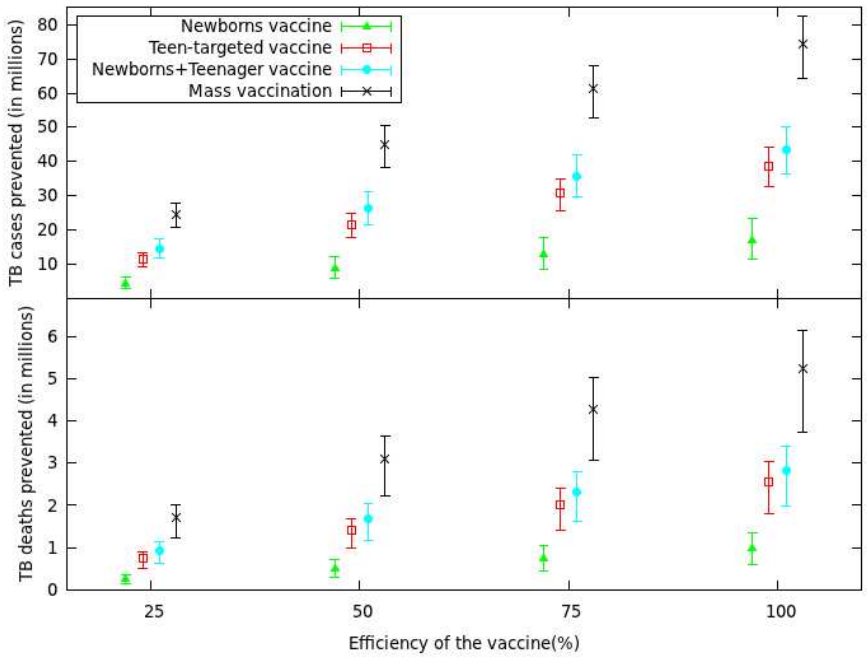


FIGURE 9.8: Impact evaluation of different immunization campaigns in South east Asia as a function of the efficacy of the vaccine, in terms of infectiousness reduction.

Finally, in addition to south east asia we have extended our analysis to the other four regions of greater TB burden in absolute terms: AFRL (Africa, low HIV), AFRH (Africa, high HIV), EMR (East Mediterranean region) and WPR (West pacific region)[22]. A summary of the results is annotated in table 9.4.1.

. Measure	SEAR	AFRH	AFRL	EMR	WPR
Number of TB cases (in millions)	165 (149-179)	82 (68-93)	27 (24-29)	46 (37-54)	56 (45-66)
Number of TB deaths (in millions)	17 (11-24)	15 (9-32)	4 (2-6)	4 (2-6)	4 (3-5)
Incidence rate (2001)	2191 (1900-2524)	3876 (2578-5928)	1606 (1269-1985)	1198 (970-1464)	1286 (1066-1538)
Mortality rate (2001)	462 (268-719)	1243 (421-2763)	355 (143-657)	293 (118-520)	121 (90-166)
Incidence rate (2030)	1527 (1349-1621)	1513 (1103-1982)	1641 (1451-1738)	1068 (815-1248)	613 (470-733)
Mortality rate (2030)	118 (80-142)	199 (153-335)	228 (112-349)	67 (44-78)	31 (20-47)
Vacc. impact (NFV) (in millions)	14 (9-19)	12 (3-19)	6 (5-8)	8 (4-11)	2 (1-3)
Vacc. impact (AFV) (in millions)	32 (27-37)	19 (6-32)	11 (9-12)	15 (9-19)	5 (4-6)
Vacc. impact (NFV+AFV) (in millions)	37 (30-42)	23 (8-37)	12 (11-14)	17 (10-22)	6 (4-7)
Vacc. impact (mass) (in millions)	64 (55-71)	30 (11-46)	16 (14-18)	25 (16-32)	15 (12-18)
Vacc. doses (AFV) (in millions)	963 (803-1126)	615 (544-685)	224 (200-249)	434 (375-493)	600 (486-718)
Vacc. doses (mass) (in millions)	2292 (1916-2707)	1370 (1167-1577)	523 (457-594)	1043 (900-1197)	1785 (1536-2056)
Vacc. performance (AFV) (in thousands)	34 (29-37)	32 (10-51)	42 (38-44)	34 (21-43)	9 (6-11)
Vacc. performance (mass) (in thousands)	28 (24-31)	19 (9-28)	31 (28-33)	24 (16-30)	9 (6-10)

TABLE 9.7: Model forecasts on different regions. Number of TB cases (TB deaths): total number of TB cases (deaths) from 2001 to 2050. Incidence rate: number of new TB cases per 1.000.000 individuals. Mortality rate: number of TB deaths per 1.000.000 individuals. Vaccine impact: number of TB cases prevented from 2015 to 2050. Vaccine doses: number of vaccines needed from 2015 to 2050. Vaccine performance: number of cases prevented from 2015 to 2050 per 1.000.000 vaccines applied. Newb.: vaccination campaign on newborns. Teen.: vaccination campaign on adolescents (15-20 years old). Newb.+adol.: combined vaccination campaign on newborn and adolescents. Mass: mass vaccination campaign on all age groups.(See supplementary information).

9.4.2 Influence of novel hypothesis: demographic evolution

As we have annotated before, one of the driving forces influencing the temporal trends of disease burden is the evolution of the demographic structure of the population. In order to address the quantitative influence of that factor on disease burden rates, we compare the results from our model to what it is obtained by looking at the original population pyramid in 2001 and taking it as constant during the dynamic evolution of the system, neglecting its time evolution, as it is made in some other former works [422].

As we can see in table 9.4.2, the variations both in number of cases and deaths under no intervention, are minimum regardless the structure of the population is heterogeneous or not. This is logical, as these forecasts are the extrapolation of the incidence and mortality curves that have been fitted between 2001 and 2011, and so, regardless the contact patterns, the fitting procedure forces them to converge. The situation is very different when we enter at evaluating the impact of different age-specific vaccination strategies, and the bias in model forecasts introducing after neglecting the social structure of the contact patterns can reach up to the 84%. This can be also easily understood, as the degree of connectedness of each age group strongly differs if heterogeneous contact patterns are considered, and so does their infectious potential. The errors in impact evaluation of mass vaccination campaign are, again, smaller,

Measure	Demographic evolution neglected	Demographic evolution considered	Relative error
Average incidence rate over all ages (2001-2050)	1725 (1568-1830)	1686 (1532-1823)	2.3% (-1.2-5.2)
Average mortality rate over all ages (2001-2050)	189 (120-264)	188 (118-265)	0.7% (-1.5-2.7)
Average incidence rate for children <5 y. old (2001-2050)	673 (570-759)	622 (532-711)	8.2% (3.4-12.0)
Average mortality rate for children <5 y. old (2001-2050)	52 (33-78)	51 (31-78)	2.4% (-2.3-6.8)
Fraction of cases prevented with a newborn vaccine in 2015	22% (20-24)	12% (8-15)	85% (44-155)
Fraction of deaths prevented with a newborn vaccine in 2015	16% (14-18)	9% (6-12)	84% (41-134)
Fraction of cases prevented with a AFV in 2015	40% (36-44)	28% (25-31)	44% (36-52)
Fraction of deaths prevented with a AFV vaccine in 2015	34% (30-38)	24% (21-27)	44% (36-51)
Fraction of cases prevented with a NFV+AFV vaccine in 2015	47% (43-50)	32% (29-35)	49% (38-61)
Fraction of deaths prevented with a NFV+AFV vaccine in 2015	39% (35-43)	26% (23-30)	47% (37-59)
Fraction of cases prevented with a mass vaccine in 2015	65% (60-69)	56% (51-61)	16% (12-19)
Fraction of deaths prevented with a mass vaccine in 2015	58% (52-63)	51% (45-56)	14% (11-17)

TABLE 9.8: Effects of demography coupling of model outcomes

as the fact that all age groups are simultaneously immunized hinders the aforementioned effect. Taken together, these results make evident that the impact assessment of any age-specific intervention needs an accurate description of the time variation of the demographic pyramids to avoid the introduction of systematic biases.

9.4.3 Influence of novel hypothesis: contact heterogeneity and impact evaluations of age focused vaccines.

As we have already said, different factors can limit, or even make unfeasible a massive administration of an eventual new anti-tubercle vaccine on individuals of any age. Foreseeing an eventual scenario in which a new vaccine has to be applied on a specific age group, it is obviously necessary to know, as reliably as possible, the way the impact of a drug of this kind depends on the age of the target population. In order to tackle this forecast task, it is clear that the selected model must make a description of the age structure of the population as precise as possible. This requires not only an explicit consideration of the evolution of the demographic pyramid as it has been explained, but also reflecting the heterogeneity of the contact patterns between individuals of different ages with the highest accuracy [12, 425]. For this reason, it is a mandatory exercise to test the way in which the heterogeneity of contacts, as described in this chapter, influences the impact of age focused vaccination campaign.

In figure 9.4.3, we represent the impact of a eventual vaccine applied in 2015 in South East Asia, applied over an specific age group, in terms of number of cases and deaths prevented up to 2050. Black dots represent the model forecasts when, as it is done in previous models, the contact patterns are considered homogeneous. Blue lines, instead, in this figure stand for the case in which the description contemplates heterogeneity in the mixing patterns.

As we can see, a proper description of the heterogeneity in the contact patterns among age groups modify substantially the relative impact of vaccines applied over different age groups. More specifically, while impacts associated to vaccination campaigns focused on children of 0-5 years old are slightly lower than predicted by homogeneous models, the impact of the same vaccine, when applied to adolescents is 19% higher than predicted by models with homogeneous mixing. This makes that the difference of impact between both vaccination strategies is a 71% higher (16.5 million individuals) when the model to estimate it contemplated contact heterogeneities than when it does not (9.7 million individuals).

9.4.4 Model forecasts for different regions

Base scenario

In figure 9.4.4, we represent the projected incidence and mortality rates for the five regions in the world with highest TB burden levels.

The red, solid curves represent the stationary values $i^*(t) = i^*(d(t), \beta(t), \{N(a,t)\})$ and $m^*(t) = m^*(d(t), \beta(t), \{N(a,t)\})$, calculated on each time point by letting the

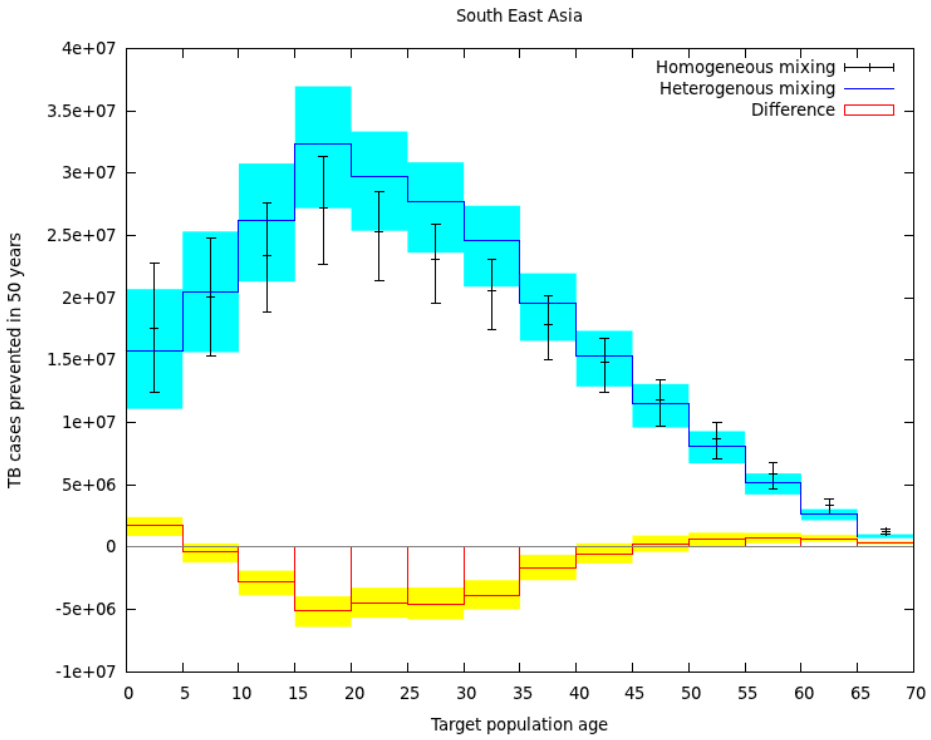


FIGURE 9.9: Impact of different vaccination campaigns focused on individuals of different ages in terms of number of cases prevented. Black: homogeneous mixing model. Blue: heterogeneous mixing model. Red: difference.

dynamics to relax up to the stationary when the parameters and the demographic pyramids are frozen on the values of each time point. As we can see, the temporal evolution of the stationary values –that illustrates the shift due to parameters and demography variations of the attractor of the dynamics– is surprisingly faster than the dynamics itself, with relevant implications and divergent behaviors in the different regions under study; which noticeably shows that the classical hypothesis of initial stationarity in hardly justifiable. The figure evidences that the high inertia of the system will force health authorities worldwide to maintain high standards of disease control and containment even decades after the burden levels will fall beneath really low limits; if eventual future relapses of the global epidemiological situation wants to be avoided.

Alternative scenarios

The reliability of our temporal forecasts presents, however, an intrinsic limitation relayed to the extrapolation of present time variation trends of the fitted parameters $\beta(t)$ and $d(t)$ to the future. Within the scenario showed up to this point, we assume that the annual rates of variation of both parameters α_β and α_d remain constant during

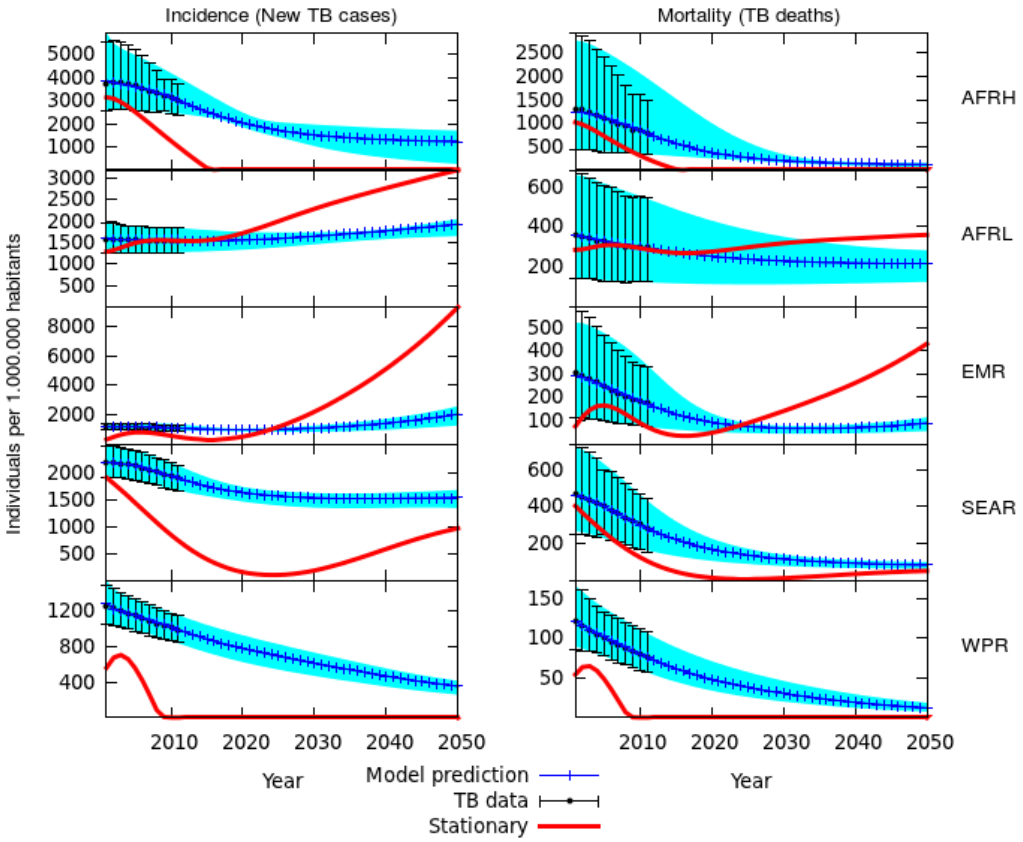


FIGURE 9.10: Fitted forecast curves for incidence and mortality rates in Africa, low HIV (A), Africa, High HIV (B), Euro-mediterranean region (C), South East Asia (D), West pacific region (E)

the whole period of the projection. However, there are two relevant caveats regarding this point. On the one hand, future social, economical, political and scientific changes might modify the outlook in the following decades in a way that is, by definition, unpredictable. On the other hand, at the long term, the maintenance of constant variation rates for these parameters, will make them explode; and more specifically, parameter values –mostly $d(t)$– start taking rather unrealistic values even a few decades in the future. Taking into account these questions, in figure 9.4.4 we sketch how incidence and mortality rates would be affected by a modification in the annual rates of variation of $\beta(t)$ and $d(t)$. by deriving two alternative scenarios: a first scenario according to which the time evolution of time-dependent parameters stops from 2030 on and a third, extreme scenario according to which, since 2030, time evolution of $\beta(t)$ and $d(t)$ reverts its trend up to recover, in 2050, the initial values of 2001.

In order to interpret these forecasts, it is convenient to highlight that, for most

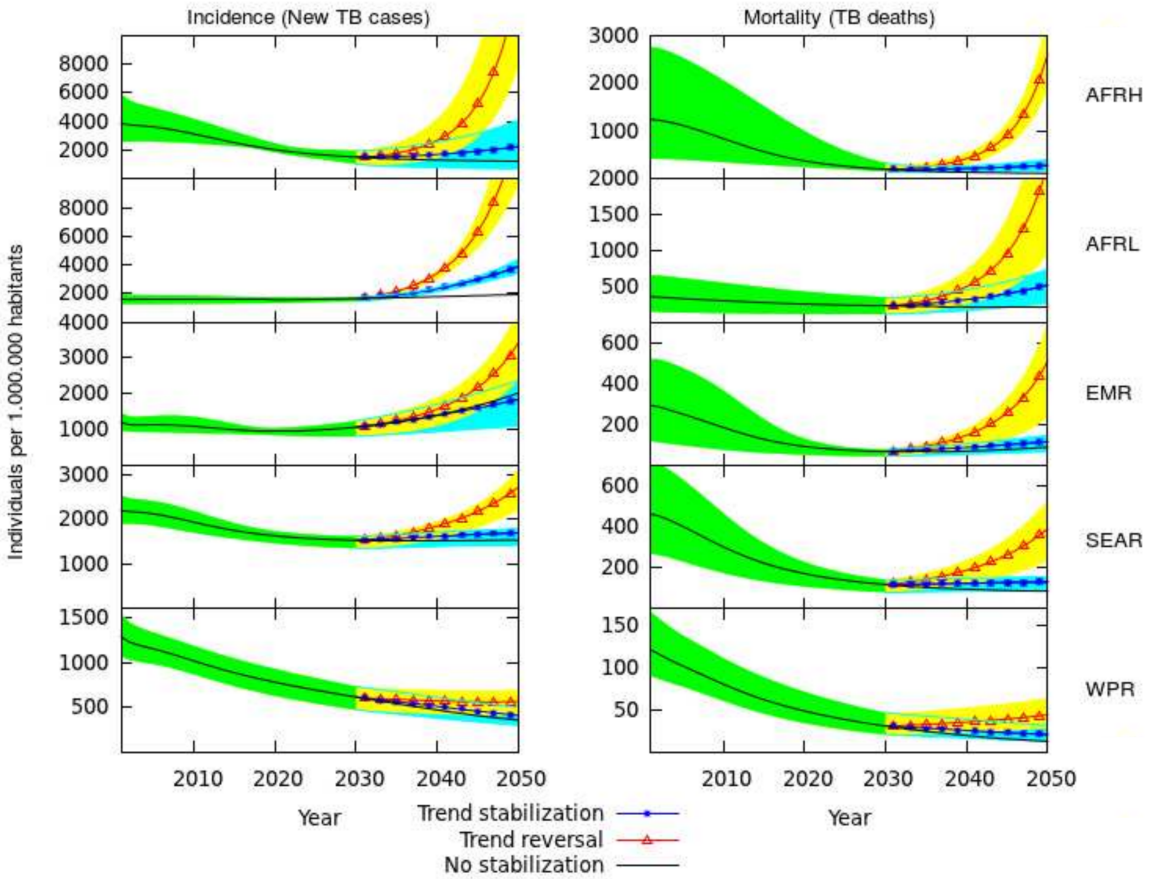


FIGURE 9.11: Alternative forecast scenarios for incidence and mortality rates.

regions, the fitting outcomes correspond to a increase in detection rates as well as to a remarkably lower variation of infectivity ($|\alpha_\beta| < |\alpha_d|$), that can either be positive or negative. In that sense, stopping the time evolution of the parameters or reverting it, has the effect of enhancing disease spreading, as the more important effect is that due to stabilize $d(t)$ (or revert its temporal trend).This implies that the two alternative scenarios are, in general, more pessimistic than the base case; with the exception of the trend stabilization of the parameter trends in the east Mediterranean region, in which, the effect of stabilizing β –a increasing function with time– contributes to contain disease burden up to the point of counteracting the effect of stopping time decay of the detection rate $d(t)$. As it can be seen in 9.4.4, the decay in TB burden is tightly associated, in terms of our model, to the time evolution of these parameters, which, from an epidemiological point of view, means that the decay in TB burden observed during the last decades around the world has to be interpreted as an achievement as fragile as precious. The stabilization of detection rates and infectivities, from 2030 on,

disturbingly yields to relevant disease relapses in at least three regions, in addition to the relapse in TB burden which is observed in the east mediterranean region and low HIV Africa, also in the base case. In addition, when the time variations accumulated up to 2030 are reverted from that point on, the relapse of the burden levels raises the burden levels even beyond their initial values.

In conclusion, if the final goal of public health authorities is that of maintaining the disease burden decay in the following decades so as to walk the way to disease eradication, public health efforts must be maintained in the future days; or otherwise, all the achievements reached to this date could even disappear.

Vaccine impacts forecasts

In figure 9.4.4, the incidence and mortality rates projected in the different regions are represented; after the simulation of different vaccination strategies applied in 2015 in each case.

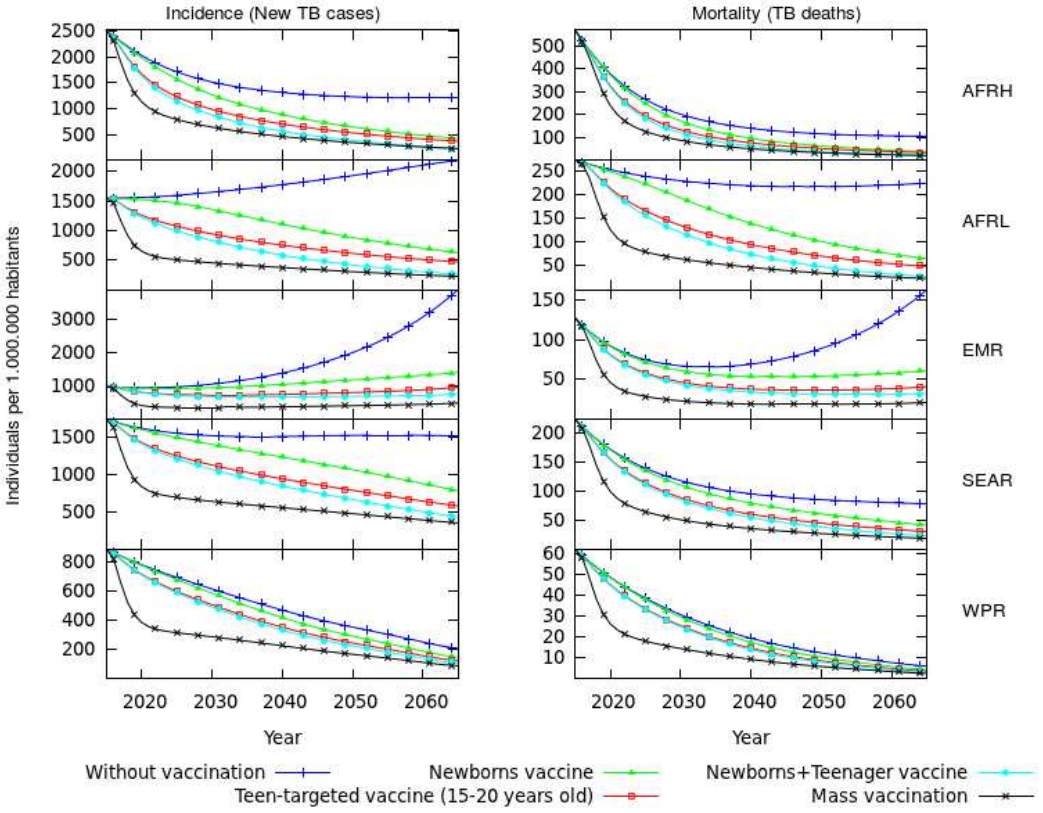


FIGURE 9.12: Incidence and mortality rates after vaccination, for different vaccination strategies in each region.

9.4.5 Model uncertainty and sensitivity analysis

The model parameters –both regional and global–, as well as the demographic data and the TB burden estimations carry intrinsic uncertainties whose influence on the model fitting and forecast has to be evaluated. In order to accomplish that task, we have performed an exhaustive uncertainty and sensitivity analysis that allows us to evaluate a confidence interval for our model forecasts as well as the part of this uncertainty that is originated by each of the model inputs.

Uncertainty sources analysis

In our model, we consider four main different types of inputs, which constitute uncertainty sources that are mutually independent:

- Global parameters (and regional parameter η): each of the estimation of these parameters is based on independent bibliographic sources. This makes a total amount of 21 parameters, each of them treated as an independent uncertainty source $u_i, i \in [1, 21]$.
- Burden and treatment outcomes estimations provided by WHO: based upon a number of case notifications and treatment outcomes of finite cohorts surveilled in each country, the World Health Organization provides estimations for incidence and mortality rates $\bar{i}(t)$ and $\bar{m}(t)$ and for the treatment outcome fractions $(f_D^{p+}, f_F^{p+}, f_\mu^{p+})$, and $(f_D^{p-}, f_F^{p-}, f_\mu^{p-})$. For the purpose of our model, we have grouped these WHO-curated measurements and considered them as a single uncertainty source, labeled as u_{22}
- Demographic pyramids $N(a, t)$: which are also considered as a single uncertainty source, labeled as u_{23}.
- Contact matrix $\Upsilon(a, a')$; the last single uncertainty source u_{24}

By proceeding this way, we count with 24 uncertainty sources $u_i, i \in [1, 24]$ considered independent, whose contributions to the uncertainty of a certain model measurement x can be evaluated by defining the variables $d_i^{low}(x) = x(u_1, ..., u_i^{low}, ..., u_{24}) - x(u_1, ..., u_i, ..., u_{24})$ and $d_i^{up}(x) = x(u_1, ..., u_i^{up}, ..., u_{24}) - x(u_1, ..., u_i, ..., u_{24})$, which represent the differences of the measurement x (i.e. an incidence or mortality rate, either temporal or accumulated, for example), evaluated after the only variation of the uncertainty source u_i, which takes the values of the lower and upper limits of its confidence interval u_i^{low} and u_i^{up}, respectively.

At this point it worths mentioning two relevant aspects. On the one hand, it has to be noted that each uncertainty source is not a single parameter, but a group of them of related origin, whose confidence intervals plausibly carry strong correlations. For example, some of the parameters are age-dependent, WHO estimations comprise several measurements of different nature (treatment outcomes fractions, mortalities and incidence rates), and either demographic pyramids or contact matrixes are multidimensional objects too. This means that, when evaluating, for example $x(u_1, ..., u_i^{low}, ..., u_{24})$,

all the "components" of u_i (i.e. the value of a parameter on all ages, or all the entries of the contact matrix, for example) are simultaneously evaluated at the bottom of their confidence intervals.

On the other hand, the evaluation of $x(u_1, ..., u_i^{low}, ..., u_{24})$ or $x(u_1, ..., u_i^{high}, ..., u_{24})$ –for example– implies the repetition of the fitting procedure after variating u_i so as precisely to address the influence of the error associated to the input represented by u_i on the reliability of the whole fitting and forecasting procedure of our model. This implies that the sign of the factors $d_i^{up}(x)$ is not trivial for any uncertainty source, and even can coincide with the sign of $d_i^{low}(x)$, due to that, after the shifts in u_i the fitting is iterated and will produce results of x which may be equally greater or lower than the central value.

Taking that into consideration; we can evaluate the uncertainty in measure x that is due to the uncertainty of the parameters reported in the bibliography. In order to do so, we have to sum up the contributions to the error of the first uncertainty source –the bibliographic parameters–, by performing the following partial sums:

$$\Delta_x^{param,up} = \sqrt{\sum_1^{21} \delta\left(\frac{|d_i^{low}(x)|}{d_i^{low}(x)} - 1\right) d_i^{low}(x)^2 + \delta\left(\frac{|d_i^{up}(x)|}{d_i^{up}(x)} - 1\right) d_i^{up}(x)^2} \tag{9.78}$$

$$\Delta_x^{param,low} = \sqrt{\sum_1^{21} \delta\left(\frac{|d_i^{low}(x)|}{d_i^{low}(x)} + 1\right) d_i^{low}(x)^2 + \delta\left(\frac{|d_i^{up}(x)|}{d_i^{up}(x)} + 1\right) d_i^{up}(x)^2} \tag{9.79}$$

and, in a similar way, we can isolate the contribution to the model uncertainty when determining x of the other uncertainty sources as follows:

$$\Delta_x^{WHO,up} = \sqrt{\delta\left(\frac{|d_{22}^{low}(x)|}{d_{22}^{low}(x)} - 1\right) d_{22}^{low}(x)^2 + \delta\left(\frac{|d_{22}^{up}(x)|}{d_{22}^{up}(x)} - 1\right) d_{22}^{up}(x)^2} \tag{9.80}$$

$$\Delta_x^{WHO,low} = \sqrt{\delta\left(\frac{|d_{22}^{low}(x)|}{d_{22}^{low}(x)} + 1\right) d_{22}^{low}(x)^2 + \delta\left(\frac{|d_{22}^{up}(x)|}{d_{22}^{up}(x)} + 1\right) d_{22}^{up}(x)^2} \tag{9.81}$$

$$\Delta_x^{demo,up} = \sqrt{\delta\left(\frac{|d_{23}^{low}(x)|}{d_{23}^{low}(x)} - 1\right) d_{23}^{low}(x)^2 + \delta\left(\frac{|d_{23}^{up}(x)|}{d_{23}^{up}(x)} - 1\right) d_{23}^{up}(x)^2} \tag{9.82}$$

$$\Delta_x^{demo,low} = \sqrt{\delta\left(\frac{|d_{23}^{low}(x)|}{d_{23}^{low}(x)} + 1\right) d_{23}^{low}(x)^2 + \delta\left(\frac{|d_{23}^{up}(x)|}{d_{23}^{up}(x)} + 1\right) d_{23}^{up}(x)^2} \tag{9.83}$$

$$\Delta_x^{contacts,up} = \sqrt{\delta\left(\frac{|d_{24}^{low}(x)|}{d_{24}^{low}(x)} - 1\right) d_{24}^{low}(x)^2 + \delta\left(\frac{|d_{24}^{up}(x)|}{d_{24}^{up}(x)} - 1\right) d_{24}^{up}(x)^2} \tag{9.84}$$

$$\Delta_x^{contacts,low} = \sqrt{\delta\left(\frac{|d_{24}^{low}(x)|}{d_{24}^{low}(x)} + 1\right) d_{24}^{low}(x)^2 + \delta\left(\frac{|d_{24}^{up}(x)|}{d_{24}^{up}(x)} + 1\right) d_{24}^{up}(x)^2} \tag{9.85}$$

Along this chapter (see fig. 9.4.4), turquoise areas represent the error bars in incidence ($\Delta_i^{low}, \Delta_i^{high}$) and mortality projection curves ($\Delta_m^{low}, \Delta_m^{high}$), which are calculated by summing up all the contributions:

$$\Delta_x^{low} = \sqrt{(\Delta_x^{param,low})^2 + (\Delta_x^{WHO,low})^2 + (\Delta_x^{demo,low})^2 + (\Delta_x^{contacts,low})^2} \tag{9.86}$$

$$\Delta_x^{up} = \sqrt{(\Delta_x^{param,up})^2 + (\Delta_x^{WHO,up})^2 + (\Delta_x^{demo,up})^2 + (\Delta_x^{contacts,up})^2} \tag{9.87}$$

In figure 9.4.5, we represent the total amount of error due to each of the three main uncertainty sources -bibliographic parameters, WHO estimations and demography– to the total:

- Bibliographic parameters contribution: red area: $(\Delta_x^{param,low}, \Delta_x^{param,up})$
- WHO contribution: yellow area: $\left(\left(\sqrt{(\Delta_x^{param,low})^2 + (\Delta_x^{WHO,low})^2} - \sqrt{(\Delta_x^{param,low})^2}\right), \left(\sqrt{(\Delta_x^{param,up})^2 + (\Delta_x^{WHO,up})^2} - \sqrt{(\Delta_x^{param,up})^2}\right)\right)$
- Demography contribution: blue area: $\left(\left(\Delta_x^{low} - \sqrt{(\Delta_x^{param,low})^2 + (\Delta_x^{WHO,low})^2}\right), \left(\Delta_x^{up} - \sqrt{(\Delta_x^{param,up})^2 + (\Delta_x^{WHO,up})^2}\right)\right)$

The contribution to the total error bars due to the uncertainty in the contact matrix fields is negligible when compared to the other three.

Model (empiric) sensitivity

Additionally to distinguishing the amount of uncertainty derived from each type of input, we have to isolate the model sensitivity to each of the parameters. More specifically, we estimate the sensitivities ς_x of model measurement x to the uncertainty source u_i as the following percentages:

$$\varsigma_x^{up}(u_i) = 100 \cdot \frac{d_i^{up}(x)}{x(\{u_i\})} \tag{9.88}$$

$$\varsigma_x^{low}(u_i) = 100 \cdot \frac{d_i^{low}(x)}{x(\{u_i\})} \tag{9.89}$$

where $x(\{u_i\})$ represents the value of measurement x when all the inputs u_i are evaluated at their central, expected values. In figure 9.4.5, we see the sensitivities of the total incidence $(\varsigma_I^{low}(u_i), \varsigma_I^{up}(u_i)$ and mortality $(\varsigma_M^{low}(u_i), \varsigma_M^{up}(u_i)$ with respect to all uncertainty sources, in all the regions studied.

As we see, the uncertainty of WHO estimations of disease burden and treatment outcomes constitute the main single source of uncertainty for both incidence and mortality projections, which is a logical result, provided that the uncertainties in the burden levels has a direct correlate on the burden projections by the model.

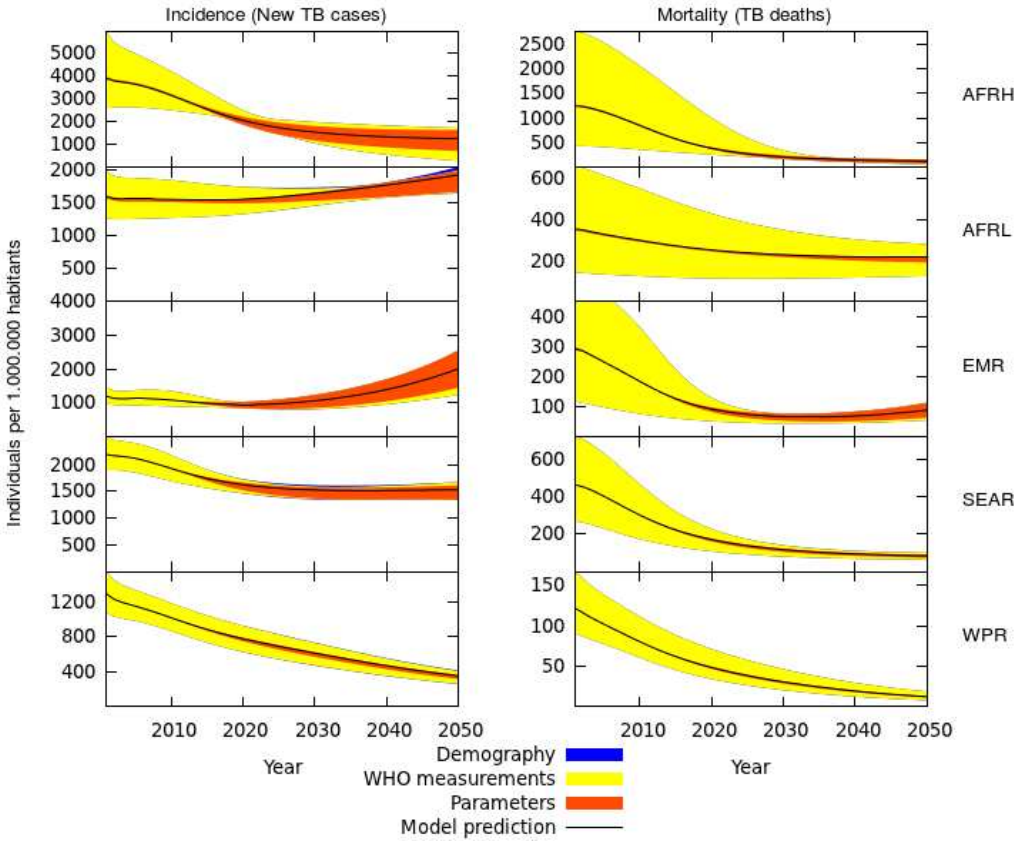


FIGURE 9.13: Model projections and uncertainty sources.

Intrinsic sensitivity

The confidence intervals of the different inputs of the model, which have been used in the last section to evaluate model sensitivities, are of disparate relative width. In that sense, the sensitivity calculated there is not an intrinsic sensitivity of the model to each input, but an empirical sensitivity influenced by the degree of accuracy of each model input estimation.

In order to isolate that factor and evaluate an intrinsic sensitivity of our model to the variation of each input, we have repeated the sensitivity analysis by substituting the disparate confidence intervals of each model input by intervals of equivalent width–in relative terms– equal to the 15% of each uncertainty source. The results are represented in figure 9.4.5

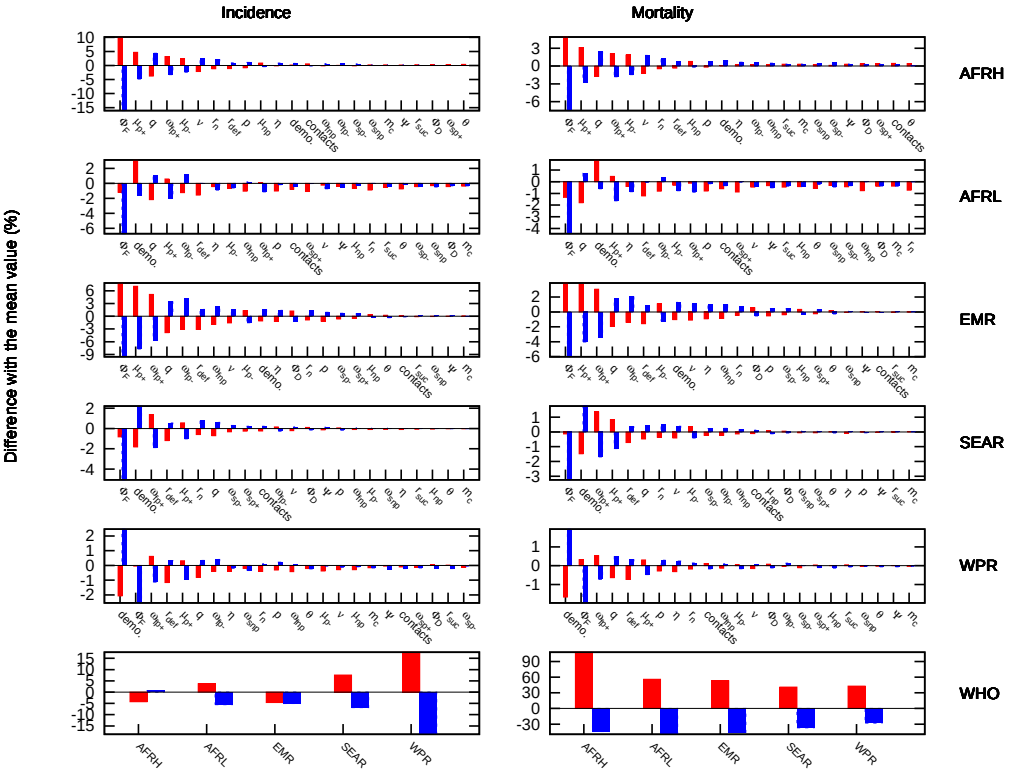


FIGURE 9.14: Model sensitivity analysis. In panels labeled as AFRH, AFRL,EMR, SEAR and WPR we represent the model sensitivity to the variation of each model input according to their confidence intervals. Bottom panels, represented by WHO contains the sensitivity of the model, within each region, to the uncertainty of the estimations provided by the WHO at [24].

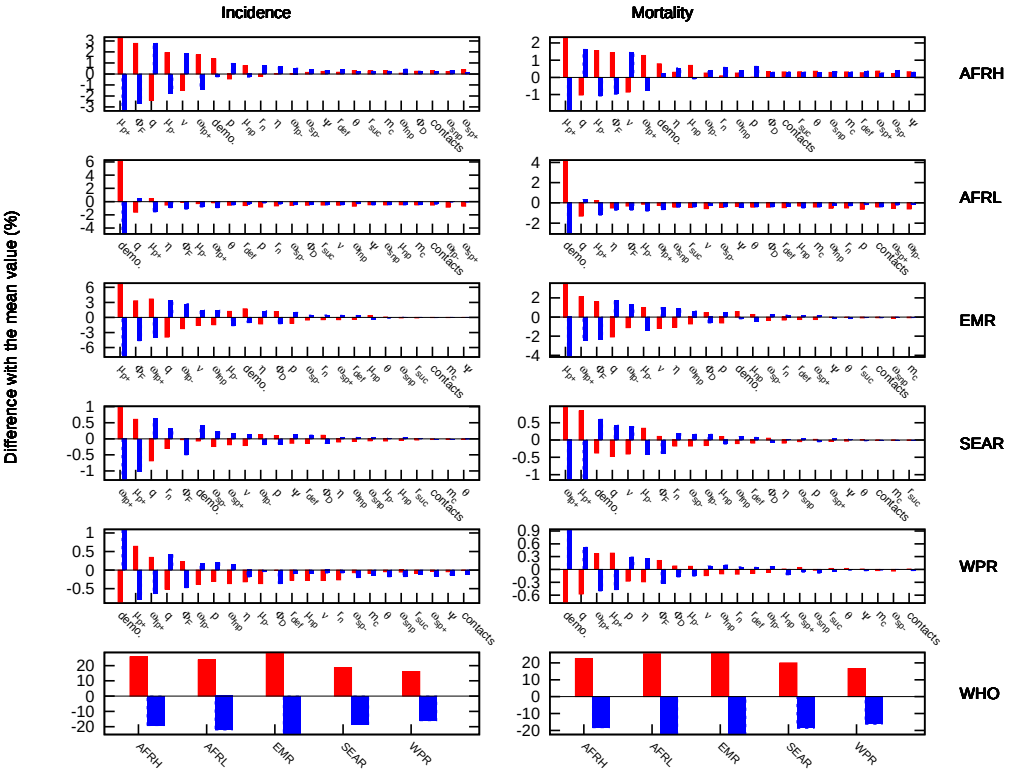


FIGURE 9.15: Intrinsic model sensitivity analysis. The variations in model measurements are calculating after equivalent variations equal to a 15% of each model input.

9.5 Conclussions

In this chapter we introduce a new model for the description of TB spreading that gets rid of certain oversimplifications responsible of relevant biases on model outcomes. As it has been previously discussed, either the explicit consideration of the temporal evolution of the demographic structure and the incorporation of realistic, heterogeneous mixing patters among different age groups in the model, relevantly modify model forecasts, which make obvious the need to improve the classical models of TB spreading towards a more faithful description of the disease dynamics and its dependence on the age structure of the populations.

In this sense, the incorporation of the mentioned heterogeneous patterns of interaction spotted in former works [12], remarkably influences the comparison of different age-specific vaccination campaigns impacts.

In fact, the impact of a vaccine applied to adolescents is the highest, and it is a 19% higher than predicted by homogeneous models. This is due to the co-ocurrence of two factors. First, it is a well known fact that children are less prune to develop the infectious forms of the disease than adults [22]. Being adolescents the first adult-like age group, they are the first at having high probabilities of developing infectious forms of pulmonary TB; which make them –also in homogeneous models– an optimal target for immunization strategies. In addition, once contacts heterogeneity is considered, adolescents are one of the most strongly connected classes, and most of their contacts are established with adult individuals (prune to develop infectious TB), which enhances the impact of adolescent focused vaccination campaigns (AFV) with respect to previous models without contact structure.

On the contrary, impacts of age focused vaccines, when applied on younger and younger individuals, decay faster as a consequence of the contact structure. The reason is again related to the assortative structure of the contact matrix, according to which children tend to interact among them rather than with adults. This implies a prophylactic effect on children, because, as they oddly develop infectious forms of TB, to concentrate their social interactions among them maintains them protected from TB infection; which finally makes the impact of any children focused vaccine to be lower than expected by homogeneous models [422, 22].

Taken together, these effects causes differences between AFV and NFV that are 71% higher than expected, which makes evident the relevance of discarding over-simplifying hypothesis in TB spreading modeling, always considering the data available which models must always confront to. However, further model refinements have to be implemented before vaccine impact assessments provided by our model could be considered relevant from a quantitative point of view. Remarkably, for an age-focused vaccine to be evaluated by the model, this must offer a faithful description of the disease burden distribution among the different age groups, prior to vaccination, something that has been neither explored in this chapter nor in previous literature of TB spreading modeling at a global scale [22, 422]. Fortunately, WHO TB database [24] contains the information needed to accomplish that goal, and that will be precisely our main objective for the following chapter.

Chapter 10
Age-focused vaccine impact estimations

10.1 Introduction

As we commented in the previous chapter, in the current portfolio of novel vaccines under development we can find candidates designed to substitute current BCG on newborns and booster vaccines whose intended application on previously BCG immunized people may allow their administration to older individuals [428, 429, 430, 186]. Additionally, prime vaccines aimed at substituting BCG could be also applied on people previously vaccinated with BCG [421], which makes the debate about what segments of the populations constitute optimal targets for age-focused immunization campaigns a global question that will must be faced by all the successful candidates for novel anti TB vaccines. The implications of vaccinating one age group or other may differ, as individuals' susceptibility to disease depends on age, as well as their social behavior -e.g. the number and nature of contacts [12]- does.

Two age groups appear concurrently as most promising targets for such new immunization campaigns: newborns and adolescents (i.e. 15-20 years old), with arguments in favor of both of them. On the one hand, adolescents and young adults concentrate more incidence of active TB than any other group, mainly because they are the first population segments for whom highly infectious pulmonary tuberculosis is the prevalent form of the disease. Also, these individuals maintain a larger number of contacts, making them the main spreaders of the disease. Therefore, to accomplish an immediate immunization of them would derive in faster, greater vaccine impact in terms of number of TB cases and casualties prevented.

But, on the other hand, a vaccination strategy focused exclusively in adolescents would leave unprotected a significant segment of the population -i.e. children-, creating a reservoir for the disease that is unreachable for the vaccine, which could compromise the goal of eradicating TB by 2050. A vaccination campaign focused on newborns would immunize, at the long term, all the age groups in the population, provided that its effects are persistent enough. Besides, recent epidemiological studies conducted in Cape Town (South Africa) indicate that in such a high-burden setting, incidence of TB in children between 0 and 4 years old may be almost as high as that of adolescents [159], which would highlight the relevance of immunizing people from their infancy.

The context of this disjunctive takes place after the publication of the results of the Phase IIb clinical trials of the novel booster vaccine MVA85A [213], which are the first trials of this kind –aimed at determining not just vaccine safety but also efficacy– conducted since the implementation of BCG, one century ago. These trials, conducted in Worcester (South Africa) over almost 3000 BCG-vaccinated infants (6 months old), concluded that the protective effects of the novel vaccine were very reduced (a 17.3%

of protection against disease and no protection against infection). These disappointing results have made the entire community of TB research enter in a state of alarm, and, as a consequence, rethink about the need of establish more stringent criteria that, for the sake of resources saving, were able to identify failed candidates before they enter in the costly phases II and IIb of the production pipeline. This may include an exclusive focalization in vaccines designed to be applied on specific age groups: ideally newborns or adolescents.

The question is complex and delicate, as there are powerful arguments supporting both types of vaccination campaigns. In the following sections we refine and use the model developed in chapter 9 to quantitatively explore these arguments and discuss their implications.

10.2 Methods

All the vaccine impact estimations conducted in this manuscript up to this point have been obtained through the model developed in the previous chapter, which offers a faithful picture of overall TB spreading dynamics. However, the model still presents an essential limitation that hinders its usage for quantitatively meaningful comparisons among age-focused vaccines. This limitation has to do rather than with the model itself, with the design of the fitting procedure. As portrayed in chapter 9, the fitting algorithm is designed to make the model reproduce the total incidence and mortality rates over time, and so, the procedure offers no means to control how the disease burden is distributed among the different age segments of the population. In figure 10.1 we see how the TB burden distribution among age groups offered by the model, as presented in chapter 9 for the case of high South East Asia region (SEAR) is remarkably deficient, mostly for what regards some age groups (e.g. people between 15 and 24 y.o.).

This problem becomes crucial when trying to evaluate age-focused vaccine impacts on a quantitative basis, as the initial distribution of TB incident and prevalent cases among the different age groups will definitely influence the impact that a vaccine applied on each group is able to offer. Therefore, if an age-detailed reproduction of TB burden levels is desired, some of the parameters should be age-specifically fitted. This supposes to go beyond the classical fitting scheme proposed by Dye and colleagues [422, 22], and, somehow, to question the universal validity of the parameter values accepted in these works and also used by us in chapter 9. In this line, two relevant modifications to the parameter set globally used in the last chapter are proposed here:

On the one hand, in chapter 9, we calculated the rates of disease progression from latency (fast or slow) to active disease (extrapulmonary or pulmonary, smear positive or negative) as the product of two magnitudes reported in [22] as global parameters, non dependent on the geographical setting: 1) the fast and slow crude transmission rates from latency, regardless the class of disease that infected individuals will develop (that is 0.9 y^{-1} from fast latency, c.i. $0.765 - 1.035$ and $7.5 \cdot 10^{-4}$ y^{-1} from slow latency, c.i. $(6.38 \cdot 10^{-4} - 8.63 \cdot 10^{-4})$ and 2) the fractions of progressing individuals joining each type of TB class. However, for the case of the probabilities of developing

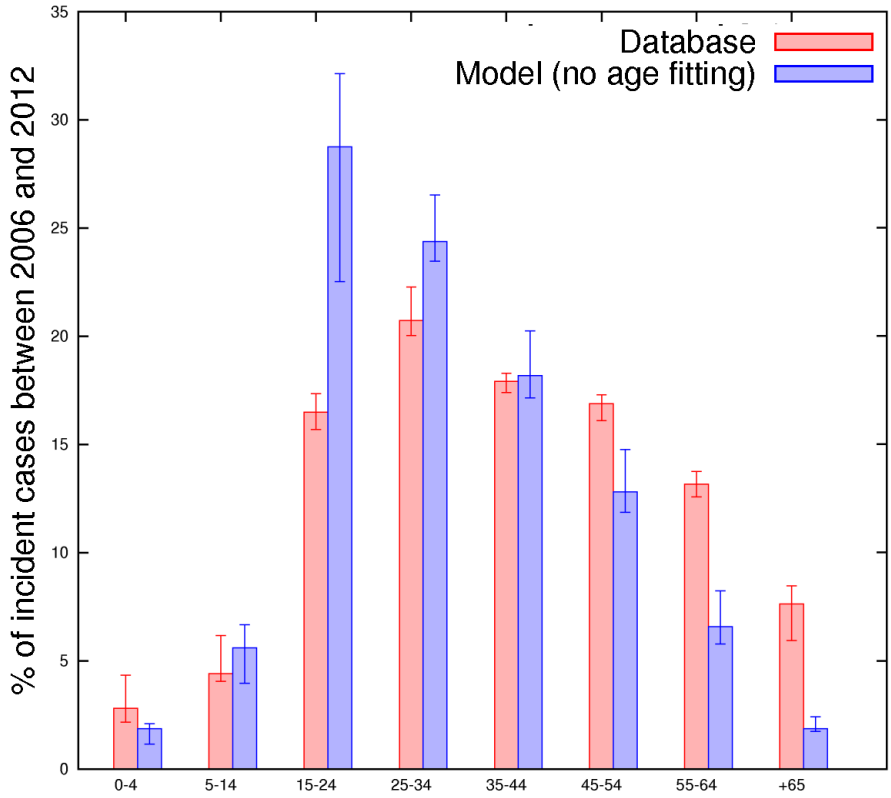


FIGURE 10.1: TB burden distribution among the different age-groups for the model following the fitting procedure presented in chapter 9 and according to data collected in [24] in SEAR region. Although the overall trend is qualitatively similar between the model and the database, relevant divergences arise, mostly for the age segment between 15 and 24 y.o. Y axis indicates total number of TB cases between 2006 and 2012, only period in which age-structured data in [24] are globally available significantly enough.

each type of TB, there is no justification to adopt a general value for every region as these fractions tend to vary in different geographic basins and, more important, the information needed for calculating these fractions in each country is available in the WHO database for TB surveilling [24].

Admittedly, this sophistication, even if constitute a more plausible estimation of the progression rates as a function of age, does not solve yet the main limitation of our model, i.e. its inability to offer an adequate picture of TB burden distribution among age groups, as the new progression rates are also an *a priori* input. However, further inspection of parameter sources allowed us to identify primo-infection probabilities as a strongly age-dependent parameter for whose values in the different age segments a remarkable low consensus can be tracked in the literature. For example, we find that for children younger than 5 y.o., reported values are highly divergent, between 5% [392] and 24% (and even 43% for children younger than one year old) [393]. This points to a strong dependence of this parameter on the particular geographical conditions in which

it is measured, as these studies were conducted in United Kingdom and South Africa, respectively.

This issue motivated us to discard global bibliographical sources of $p(a)$ and reconstruct a fitting procedure based on fitting primo-infection probabilities in each region and age group with the aim of reproducing age-distributed TB burden levels. What we aim to reproduce is, more explicitly, total active TB cases between 2006 and 2012, the only burden metrics for which enough information distributed by age-segment is available at [24].

10.2.1 Determination of the primo-infection probabilities

The primo-infection probabilities are fitted to make the model reproduce the proportion of active cases in the different age groups between 2006 and 2012. It is not immediate to obtain such proportions, as some difficulties arise. The database is structured as a matrix where the rows represent the countries and years, and the columns represent age group, sex and type of disease. The value of every cell gives us the number of notifications that fulfill the characteristics given by the row and column where it belongs -i.e. country, year, sex, age and type of disease-. Unfortunately, many cells of the database are empty, especially before 2000. To deal with this issue we had to discard many data. The following criterion is taken: we only accept rows that have at least one notification in every box. The reason behind this rule is that there are many countries that failed to identify some type of disease (sometimes just in some age groups), leaving many empty cells in their data, specially, as said, before 2000. Proceeding this way we avoid the introduction of bias due to technical limitations in the reports of some countries. The data that we found reliable under this criterion are bounded between the years 2006 and 2012, and many countries -some of them with huge populations as India- are completely discarded.

In addition, the age groups used by the database do not correspond exactly to the age groups used in the spreading model. The database uses an age window of 10 years, except for the first group that is 0 to 4 years old; and in the spreading model an age window of 5 years is implemented -that concur with the age groups used by the UN population database [23]-. To compare the results of the model with the data obtained we have to merge some age groups, as it is shown in table 10.1:

Finally, it is important to note that the last group has no upper limit of age in the database, but it is constrained in a 5 year window in the spreading model, which forces us to leave it out of the fitting process (otherwise, the primoinfection probabilities at this age-group would present strong biases). Anyway, burden levels in individuals older than 65 years constitute a reduced fraction of the total and so, this only supposes a residual error in the age distribution of TB burden.

We define the proportions of cases of the age group α and the year y, as:

$$f_{\alpha,y} = \frac{\sum\limits_{c} I_{\alpha,c,y}}{\sum\limits_{c}\sum\limits_{\alpha'=0}^{6} I_{\alpha',c,y}} = \frac{I'_{\alpha,y}}{\sum\limits_{\alpha'=0}^{6} I'_{\alpha',y}} \tag{10.1}$$

Model structure		Database structure	
Age	Index: a	Age	Index: α
0-4	0	0-4	0
5-9	1	5-14	1
10-14	2		
15-19	3	15-24	2
20-24	4		
25-29	5	25-34	3
30-34	6		
35-39	7	35-44	4
40-44	8		
45-49	9	45-54	5
50-54	10		
55-59	11	55-64	6
60-64	12		
65-70	13	+65	7

TABLE 10.1: Age groups reparametrization.

where $I_{\alpha,c,y}$ is the number of cases in the age group α, in the country c and the year y. To obtain the total fractions we perform a weighted average:

$$f_\alpha = \frac{\sum_y f_{\alpha,y} \left(\sum_{\alpha'} \sum_c N_{\alpha',c,y} \right)}{\sum_{\alpha'} \sum_c \sum_y N_{\alpha',c,y}} = \frac{\sum_y f_{\alpha,y} \left(\sum_{\alpha'} N'_{\alpha',y} \right)}{\sum_{\alpha'} \sum_y N'_{\alpha',y}} \tag{10.2}$$

where $N_{\alpha,c,y}$ is the population of the group age α, in the country c and the year y. The notation I' or N' is referring to quantities aggregated to all countries in the region of interest.

The fitting algorithm is the following:

1. We choose initial values of p_α

2. With these values we used the spreading model -after a proper fitting of the parameters, as explained in chapter 9- to obtain certain outcomes of incidence.

3. We used the incidence obtained in each age group to construct the proportion of cases given by the model: f_α^{model}

4. We compare the fractions provided by our model to those coming from the database f_α, by calculating the ratio $\frac{f_\alpha}{f_\alpha^{model}}$:

   (a) If it is constrained between 0.99 and 1.01 for all the age groups -except the last one (which we are constrained to neglect)-, we stop.

Age	p	Confidence interval
0-4	0.039	(0.030-0.060)
5-14	0.014	(0.013-0.019)
15-24	0.051	(0.049-0.054)
25-34	0.073	(0.070-0.078)
35-44	0.081	(0.079-0.083)
45-54	0.125	(0.119-0.128)
55-64	0.177	(0.169-0.184)
+65	0.177	(0.169-0.184)

TABLE 10.2: Primo-infection probabilities fitted for South East Asia Region.

(b) In any other case, we construct new p_α -for $\alpha \in [0, 6]$- and go to step 2:

$$p_\alpha^{new} = p_\alpha \frac{f_\alpha}{f_\alpha^{model}} \tag{10.3}$$

For $p_{\alpha=7}$ we choose the value of the previous age group, $\alpha = 6$.

Finally, we have to define the confidence interval for these primo-infection probabilities. First we define the confidence interval for the cases $I'_{\alpha,y}$ assuming a multinomial distribution:

$$\Delta I'_{\alpha,y} = \sqrt{I'_{\alpha,y}\left(1 - f_{\alpha,y}\right)} \tag{10.4}$$

and we propagate these errors in the calculation of $f_{\alpha,y}$, taken them as independent sources of error. After that, for every age group, we took the minimum and the maximum value of $f_{\alpha,y} \pm \Delta f_{\alpha,y}$ as the confidence interval of f_α.

Then we obtained the upper and lower limit of p_α -called p_α^{up} and p_α^{low}- with the following formula:

$$p_\alpha^{up} = p_\alpha \frac{f_\alpha + \Delta f_\alpha^+}{f_\alpha^{model}} \tag{10.5}$$

$$p_\alpha^{low} = p_\alpha \frac{f_\alpha - \Delta f_\alpha^-}{f_\alpha^{model}} \tag{10.6}$$

By completing this new fitting algorithm, we obtained the following primoinfection probabilities in SEAR, which we show in table 10.2:

10.2.2 Determination of the progression rates

As the database [24] reports the division of the notified cases according to the type of disease (pulmonary smear positive, negative and extra-pulmonary), we can also obtain the probability of progression to each kind of disease for each country and region. Then,

we multiplicate these probabilities per the net transition rates from latency to active disease regardless the type of disease, which are

As a first approximation it would seem that these probabilities are just the proportion of cases of each type; but as we have different diagnosis rates we have to rescale these numbers of cases with the parameter η introduced in chapter 9:

$$p_{p+}(\alpha,t) = \frac{i_{p+}(\alpha,t)}{i_{p+}(\alpha,t) + i_{p-}(\alpha,t) + i_{np}(\alpha,t)} \tag{10.7}$$

$$p_{p-}(\alpha,t) = \frac{i_{p-}(\alpha,t)}{i_{p+}(\alpha,t) + i_{p-}(\alpha,t) + i_{np}(\alpha,t)} \tag{10.8}$$

$$p_{np}(\alpha,t) = \frac{i_{np}(\alpha,t)}{i_{p+}(\alpha,t) + i_{p-}(\alpha,t) + i_{np}(\alpha,t)} \tag{10.9}$$

where:

$$i_{p+}(\alpha,t) = I_{p+}(\alpha,t) \tag{10.10}$$

$$i_{p-}(\alpha,t) = \frac{I_{p-}(\alpha,t)}{\eta} \tag{10.11}$$

$$i_{np}(\alpha,t) = \frac{I_{np}(\alpha,t)}{\eta} \tag{10.12}$$

As we did in the previous section, we obtain the uncertainty of $I_x(\alpha,t)$ by assuming a multinomial distribution but in this case the categories are the types of disease instead of the age groups.

$$I^{up}_{new,x}(\alpha,t) = I_{new,x}(\alpha,t) + \sqrt{\left(1 - \frac{I_{new,x}(\alpha,t)}{\sum\limits_x I_{new,x}(\alpha,t)}\right) I_{new,x}(\alpha,t)} \tag{10.13}$$

$$I^{low}_{new,x}(\alpha,t) = I_{new,x}(\alpha,t) - \sqrt{\left(1 - \frac{I_{new,x}(\alpha,t)}{\sum\limits_x I_{new,x}(\alpha,t)}\right) I_{new,x}(\alpha,t)} \tag{10.14}$$

Then we obtain the errors of $p_{p+}(\alpha,t)$,$p_{p-}(\alpha,t)$ and $p_{np}(\alpha,t)$, by propagating 4 independent sources of error, $I_{p+}(\alpha,t)$, $I_{p-}(\alpha,t)$, $I_{np}(\alpha,t)$ and η. Also, the outcomes for the age groups from 25 years old, are merged.

We obtain the following results:

The use of these parameters of tables 10.2 and 10.3 allows us to obtain more reliable age-detailed results, which is imperative in this kind of work. As a result, we recover a more accurate distribution for the TB burden among the different age groups, as we see in figure 10.2. Differences between model and database in figure 10.2 arise from truncating our model at 70 years old.

	Age	SP	SN	EP
SEAR	0-4	0.024 (0.010-0.053)	0.930 (0.912-0.947)	0.046 (0.021-0.060)
	5-14	0.157 (0.126-0.205)	0.654 (0.585-0.787)	0.188 (0.083-0.253)
	15-24	0.658 (0.595-0.714)	0.239 (0.191-0.335)	0.103 (0.038-0.146)
	25+	0.685 (0.618-0.731)	0.249 (0.201-0.341)	0.065 (0.016-0.092)

TABLE 10.3: Probabilities of developing each type of TB when progressing from latency in South East Asia region.

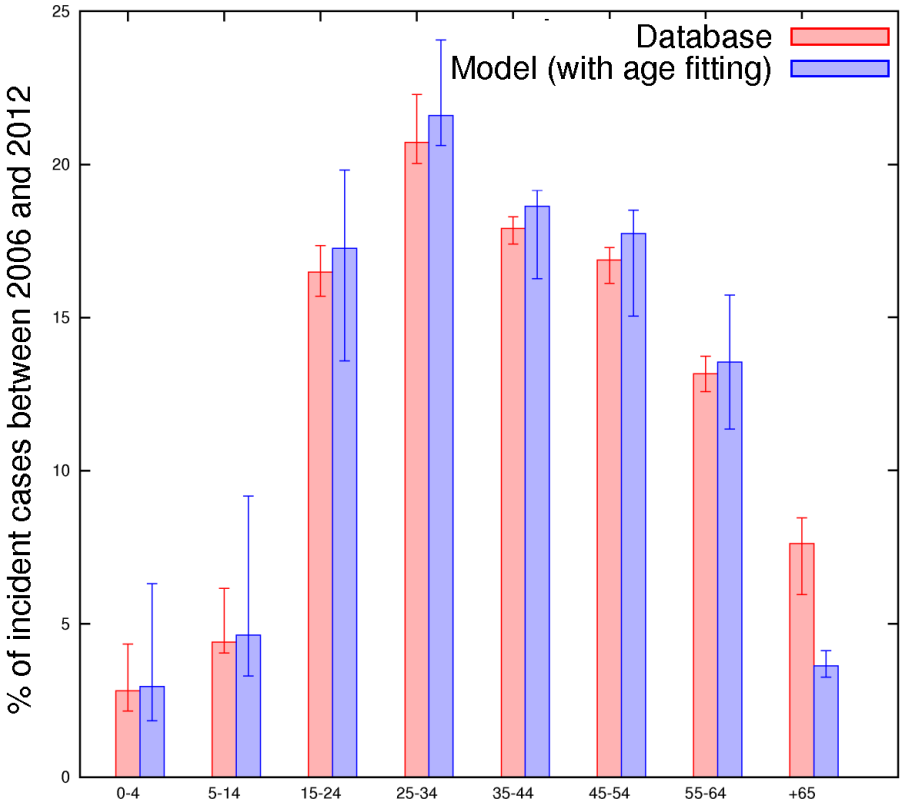


FIGURE 10.2: TB burden distribution among the different age-groups for the model presented here and according to data collected in [24] in SEAR, once progression rates and primo-infection probabilities have been re-calculated. An adequate fit of the age-structured data in [24] is recovered. For individuals older than 65 y.o., lower accuracy is obtained, as for this age group no fit for $p(a)$ is feasible because of model truncation at 70 y.o. What is fitted is the distribution of TB durned among the different classes younger than 65 y.o.

10.3 Vaccine impacts evaluations of vaccines of equal observed efficacy

In this section we will explore the different vaccine impacts that two vaccines with the same observed efficacy (i.e. equal $\epsilon = 0.3$) are able to offer when applied over the two population groups that are considered the most promising candidates to constitute optimal targets in an age-structured vaccination campaign: newborns and adolescents (15-20 y.o.)

10.3.1 Global and age distributed impacts of AFVs and NFVs

As we mentioned before, the common assumption that adolescents and young adults might suppose an optimal group for age-focused antituberculosis vaccination campaigns obeys to the fact that, typically, highest incidence and mortality rates of tuberculosis are concentrated on these age groups, as it is the case of SEAR between 2006 and 2012 (figure 10.2).

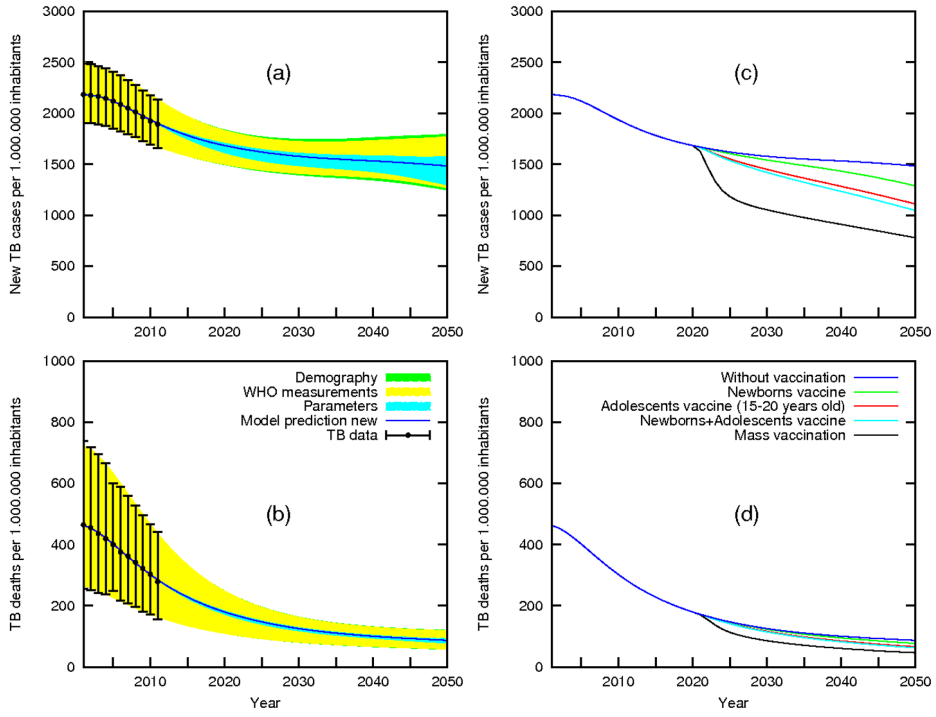


FIGURE 10.3: Model outcomes in SEAR. (a): incidence rate foreseen by the fitted model. (b) TB-related mortality rate foreseen by the fitted model.(c) Incidence rate after different vaccination strategies. (d) Mortality rate after different vaccinations strategies. Vaccine efficacies are 70% protection against infection ($\epsilon = 0.3$). Multiplicative protection is predicted for doubly vaccinated individuals (NFV+AFV)

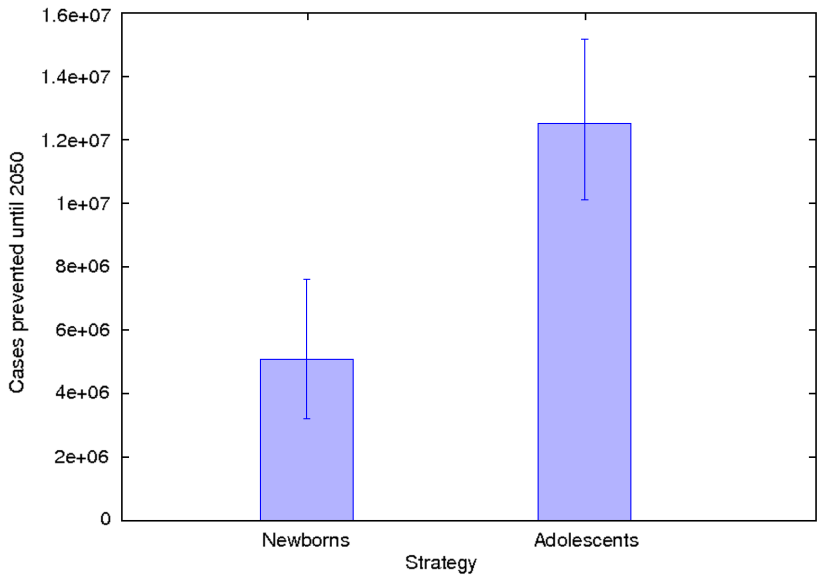


FIGURE 10.4: Impact of new tuberculosis vaccines of 70% efficacy ($\epsilon = 0.3$) applied on newborns and adolescents in 2023, measured as number of cases prevented up to 2050 in SEAR.

As we can see in figures 10.3 and 10.4, our model confirms this commonly admitted situation: the impact of an adolescents-focused vaccine (AFV) is higher than the impact of a newborn focused vaccine (NFV), when measured as the number of active TB cases prevented up to 2050 in both cases. However, as we see in figures 10.4 and 10.6, that does not mean that the impact of a NTV was negligible; neither when measured over the total population nor when it is detailed by ages. For a more general picture of how vaccine impacts depend on age (not only on newborns or adolescents), and, more remarkably, how model forecasts are affected by the re-parametrization proposed in this chapter, in figure 10.5, we represent the number of TB cases prevented up to 2050 by a vaccine introduced in 2020:

Indeed, from figure 10.6, we see that NFV's may also generate measurable effects on adult age groups, and on the other hand, leaving children unvaccinated reduces the impact of AFVs on the first age groups to a marginal side effect due to general reduction of transmission in the population. This situation could be specially relevant in certain high-burden settings, since, according to recent epidemiological studies conducted in Cape Town (South Africa) [159], incidence of TB in children between 0 and 4 years old can be almost as high as that of young adults in certain areas: in such a setting, taking into account the results of figure 10.6 in comparison to the burden distribution by ages represented in figure 10.2, the burden reduction in the youngest age groups would represent a more important fraction than found here, which would derive in reducing impact differences between AFVs and NFVs. Furthermore, from the

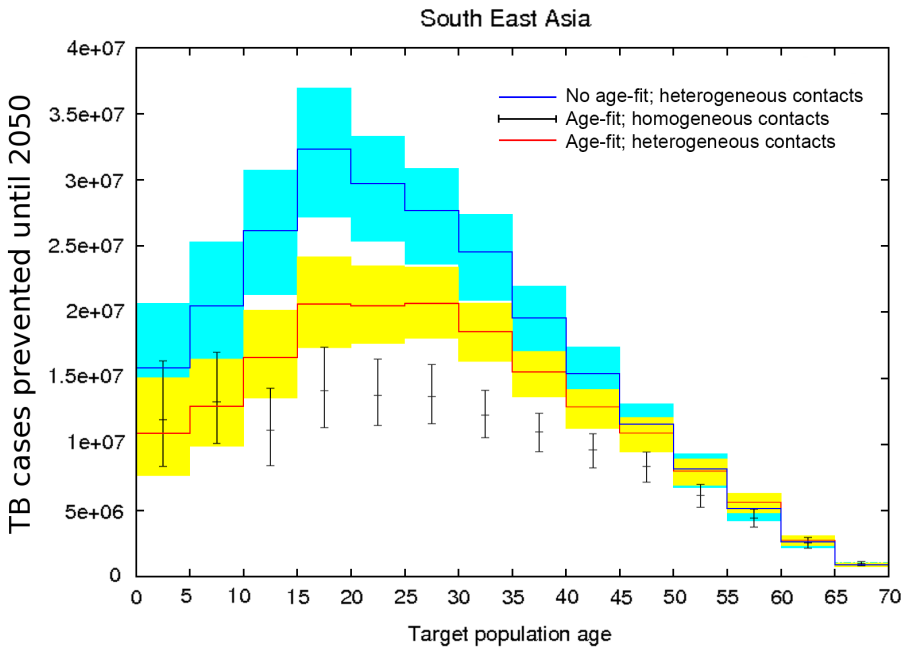


FIGURE 10.5: Vaccine impacts depending on age: both heterogeneous contact patterns and a proper description of the TB burden distributed by ages exert relevant influences on age focused vaccine impacts. For the sake of comparison with figure 9.4.3, vaccines are now introduced in 2015. (Blue series here and in 9.4.3 are the same.

lower panel of figure 10.6, a long-term problem of AFVs becomes evident: by leaving unvaccinated children youngest than 15 years old; a permanent reservoir of mycobacteria is guaranteed, compromising the final goal of total disease eradication within the next decades.

10.3.2 Time horizon for impact evaluations

In the precedent lines, we have concluded that, in general, AFVs get greater impacts than NFVs paying the prize of sacrificing the youngest age segments of the population. However, in case vaccine-derived immunity is persistent enough, part of the outperformance of AFV is only a matter of the time scale at which it is evaluated, as it only appears as a consequence of that, by quickly immunizing adolescents, adult segments of population -who are responsible of the bulk of disease transmission- get immune sooner. In figures 10.4 and 10.6, impacts are evaluated as number of active TB cases and infections prevented up to 2050; however, if this time horizon is expanded, the difference in the performance of both vaccination strategies is reduced, and finally, reversed (figure 10.7, upper panel).

Looking to the lower panel in figure 10.7, the reason of that situation becomes clearer. AFV produces a fast initial fall in disease incidence, but, later, the decline of

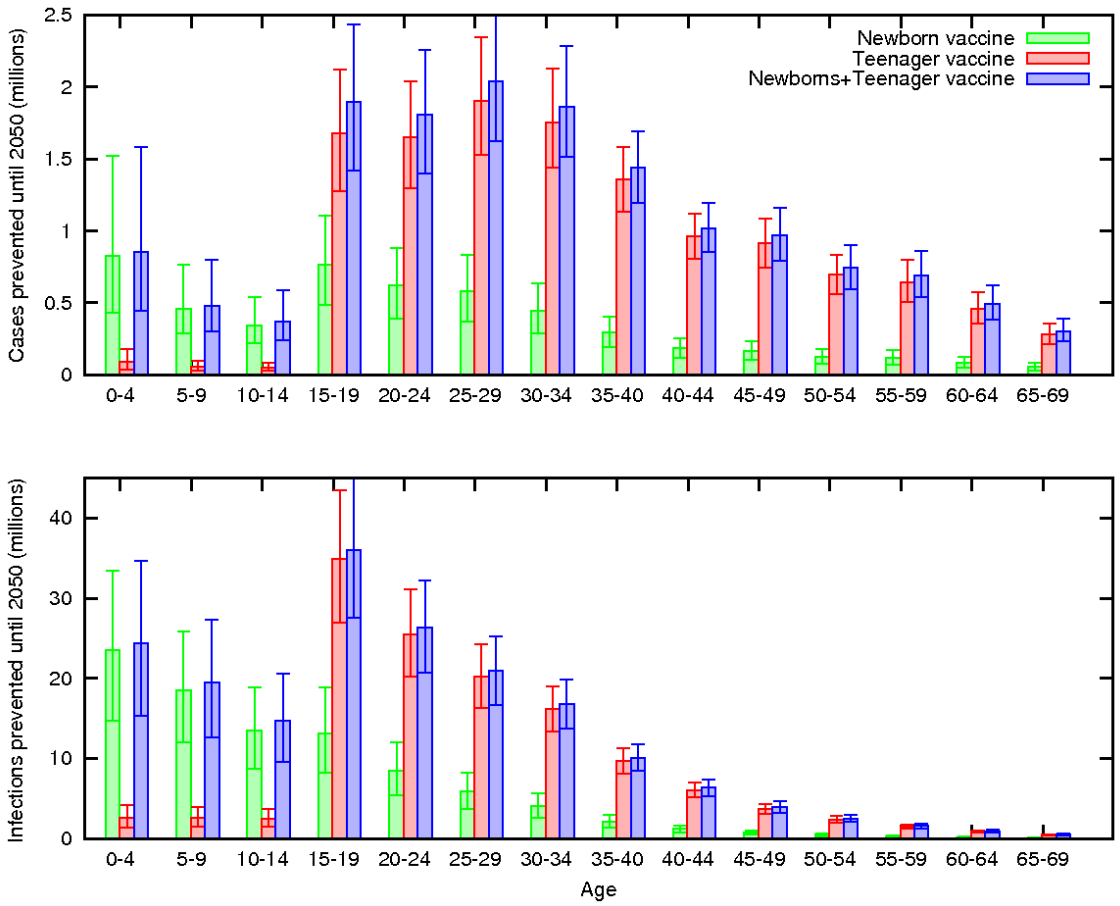


FIGURE 10.6: Number of incident cases (up) and infections (low) prevented in each age group, for the different vaccination strategies considered. (Vaccine efficacy: 70% protection against infection: $\epsilon = 0.3$).

incidence associated to the NFV is more sustained, until the point in which incidence rate after NFV is lower than that of the AFV; which is due to the fact that, as decades gone by, the immunization of the whole population is only available after a NFV based strategy.

Summarizing, each type of vaccine provides complementary advantages: to immunize adolescents provide fast impacts, while NFVs, instead, left covered the age groups unprotected by AFVs; which would guarantee a better long term performance. Therefore, the only way of reaching a fast decline, yet sustained in the long term, is to implement a combined strategy targeting both age groups, as shown in figures 10.6 and 10.7.

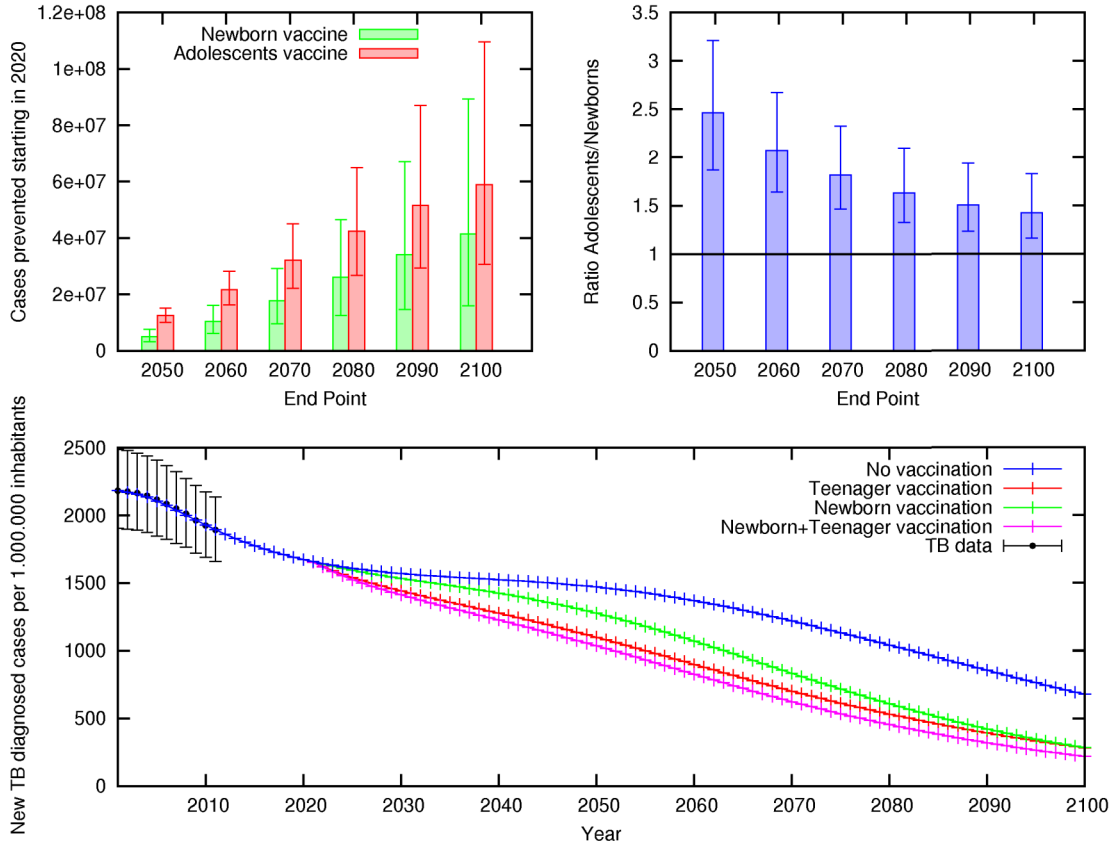


FIGURE 10.7: Effects on impact evaluations of the end point of the evaluation timespan. (a): impacts measured as number of cases prevented up to different moments for adolescents and newborn focused vaccines. (b) Impact ratio adolescents/newborns. (c) temporal projections for different vaccination strategies extended up to 2100. (Vaccine efficacy: 70% protection against infection: $\epsilon = 0.3$).

10.4 Discussion

As we have seen before, even when AFVs provide a faster impact, there exist reasons - in terms of vaccine impact- advising the deepening in the development of vaccines designed to be applied on newborns.

Although less performant at the short term, vaccination campaigns focused on newborn individuals offer beneficial effects which other vaccination strategies focused on older age groups, like adolescents, lack. In this sense, although AFVs are expected to have faster impacts than NFVs, the difference tends to be mitigated as impact estimations are evaluated on larger timespans. This situation emerges as a consequence of that, at the long term, while NFV campaigns assure fully coverage to all age groups, AFV leave unprotected the first age groups, which results in lower long-term impacts. Leaving children unprotected from infection, in addition to the problems related to

pediatric disease, guarantees the survival of an infection reservoir, which ultimately compromises our chances to finally eradicate TB, at least within the present century.

The results here presented strongly suggest that the optimal strategy to follow would ideally be that of contemporary combining the two vaccination strategies on newborns and adolescents, as the only way of simultaneously providing a fast impact, yet sustained in the long term on all age segments of the populations.

Nonetheless, impact evaluations performed in this chapter correspond to different vaccines of identical levels of observed efficacy (i.e. identical ϵ, in terms of our spreading model), that are essentially persistent in time ($\gamma = 0.015$, which corresponds to a period of immunity of 67 years after vaccination). However, different biological factors affect differently the performance of a vaccine when it is intended to be applied on different age groups, so modifying its observed efficacy and its persistence in time. In the following chapter we will explore these effects and we will quantify their affections on vaccine impact forecasts.

Chapter 11

Age-dependent effects on vaccine efficacy: masking, blocking and efficacy waning

11.1 Introduction

In the last section, we argued how vaccine impact evaluations, when two vaccines having identical observed efficacy (i.e. offering the same levels of protection against infection) are compared, depend on the precise age on the people being immunized. The analysis indicated that AFVs are expected to provide higher impacts than NFVs, but certain advantages which are privative of the last type of vaccination strategy would ideally advise the adoption of a combined immunization campaign over both age groups.

In the comparison between the performance of AFVs and NFVs made in the last section, there persists, however, a fundamental issue missed, i.e. the existence of intrinsic effects that modify the efficacy of a vaccine as a function of the age of the individuals being immunized. Additionally to the immunity waning of vaccine protection as time after vaccination passes by, there exist other effects that may reduce the observed efficacy of vaccines when applied on older age groups, which appear as a consequence of prior exposure of individuals to non-tubercle mycobacteria (or other sources of mycobacterial antigens) that are common in the environment, specially in tropical and/or high TB-prevalent areas. That is the case of the blocking and masking

effects, identified as likely major causes for the observed variation in BCG efficacy, as discussed in the introduction of this thesis.

According to these hypothetical effects [139], human immune response of individuals previously exposed to environmental mycobacteria interfere with the vaccine, impairing its protective performance in different ways. On the one hand, according to masking hypothesis, environmental sensitization confers a significant protection against TB that can not be mensurably improved by the application of a vaccine. On the other hand, the blocking hypothesis postulates that prior exposition to mycobacteria may trigger an immune response capable to block correct vaccine "take", impairing its performance. As a consequence, the older the target populations are, the stronger their sensitization to environmental mycobacteria is, and so, the stronger the vaccine efficacy's loss, no matter the precise mechanism driving it. Similarly, these effects are more intense in lower latitudes, as environmental mycobacteria are more abundant in these areas. All these trends, as previously discussed, have been identified for the case of BCG [194, 195], which reinforces the hypothesis of these effects being the main responsible for the enormous variations among different efficacy measurements –that span from 0 to 80% of protection– for the old *M.bovis* derived vaccine.

Therefore, given that these issues have been identified for BCG, there wouldn't be surprising to find them in the novel vaccines currently under development. If that was the actual case, as resulting from blocking and masking, an eventual vaccine of given efficacy ϵ, measured in newborns, would have a different efficacy ϵ' if applied on a different age group, typically associated to lower levels of protection (in our model, ϵ stands for the infectivity reduction due to the vaccine, which means that the lower is ϵ the greater is vaccine's protection, and so, we'll typically have, in this context $\epsilon < \epsilon'$). The existence of these age-dependent effects would make of limited relevance the strict comparison of impacts between equally-efficient vaccines, as, for their cause, two identical vaccines will have different efficacies when applied on different age groups (the older the target population, the lower the vaccine efficacy), or, in other words, for an AFV to present the same efficacy offered by an NFV, the vaccine applied on the older age group would have to be intrinsically better so as to compensate the efficacy loss due to prior sensitization to environmental mycobacteria.

In this chapter we aim at studying these effects, and quantifying the affection that they could exert, in terms of vaccine impact, on an hypothetical novel vaccine of a given efficacy level. In order to do so, we start by studying the case of efficacy variation in BCG, so as to estimate the most likely levels of masking and blocking that are compatible with current observations [195]. Then, using these results, we will study the behavior of different novel vaccines, both primes aimed at substituting BCG and BCG-booters.

Trials	Efficacy measured in Salvador	Efficacy measured in Manaus
Newborn vaccination	40% (22-54%)	36% (11-56%)
First-dose at school age	34% (7-53%)	8% (-39 to 40%)
Second-dose at school age	19% (3-33%)	1% (-27 to 23%)

TABLE 11.1: Results of the trial performed by Barreto et al. [195]

11.2 Masking and blocking quantification in BCG

11.2.1 A mathematical model for blocking and masking

To demonstrate the existence and importance of the masking and blocking effects, in a recently published work [195], Barreto and colleagues have conducted a set of clinical trials, specifically designed to measure BCG efficacy when applied on populations on different ages in two cities of a big country -Manaus and Salvador, in Brazil- of different latitude. In this section, we use the data of these clinical trials to extract the levels of masking and blocking in the populations of the trials, based on a simple mathematical model; and we discuss the implications for other BCG clinical trials.

As we say, the trials conducted by Barreto et al. were designed to explore the causes of BCG efficacy variation. First they choose a population of non-infected schoolchildren between 7 and 14 years old. In this population we can find individuals that have been vaccinated as newborns (they present a BCG scar), and individuals that have not been vaccinated yet. Then a BCG vaccine is applied in half -approx.- of each one of these two subgroups. So, there are 4 cohorts in total: non vaccinated, newborn vaccinated, first-dose at school age and second-dose at school age -after a newborn vaccine-. Then, the follow-up period of these cohorts is 9 years. By comparing the different ratio of diseased individuals in these cohorts, 3 different trials are performed: a newborn vaccination trial -comparing the newborn vaccinated with the non-vaccinated-, a first-dose school age trial -comparing the first-dose at school age vaccinated with the non-vaccinated-, and a revaccination at school age trial -comparing the second-dose at school age vaccinated with the newborn vaccinated-. As we say, this study is conducted simultaneously in two different cities of Brazil: Salvador and Manaus -which is expected to have a higher level of environmental mycobacteria-. The results of these trials are shown in table 11.1

From these data, we build a simple mathematical model to describe how masking and blocking effects could, hypothetically, explain these variations in BCG efficacy. The dependence of the outcomes of these trials with the levels of masking and blocking is described, according to our model, by a series of splittings applied in each cohort. For example, in what regards individual infection status, people within a cohort might face different fates: 1)not being exposed to *MTB* during the follow-up period. 2)Being exposed but protected from infection due to prior sensitization to environmental mycobacteria. 3)being exposed and protected by a vaccine or 4)being exposed and infected (because of vaccine inefficacy, intrinsic or due to blocking). Then, people within

each of these subgroups splits into two groups, as they might progress to active disease or not.

Entering into the details, the first splitting is common in every cohort: the N individuals is divided in a fraction x of exposed individuals -tentatively infected- and a fraction $1-x$ of non-exposed individuals. In the non vaccinated cohort, a fraction m of the exposed individuals is protected by the masking effect.

For a vaccinated cohort, a fraction e of individuals is protected by the vaccine, where e is the base efficacy of the vaccine. However, this efficacy will not be the same in every vaccinated cohort due to the immunity waning rate of BCG. In the case of a newborn-vaccine, an average of 15 years have past during the follow-up period since the administration of the vaccine, and in the case of school vaccination this time lapse is 4.5 years. So, we will distinguish two different efficacies: $e(15)$ and $e(4.5)$.

When we applied a revaccination we have an additional splitting for not protected individuals in e' and $1-e'$. This revaccination efficacy is calculated to recover the levels of protection of the previous vaccine -we assume that the revaccinating process works as a reset of the vaccine-. Essentially, we impose that the fraction of people protected by the combination of the newborn vaccination (of efficacy $e(15)$ at the middle of the follow up) and the revaccination, of efficacy e' is the same that the fraction protected in a first dose, school age vaccination (of efficacy $e(4.5)$). So, we have that $e'(1-e(15))+e(15)=e(4.5)$ which yields $e'=\frac{e(4.5)-e(15)}{1-e(15)}$.

Notwithstanding that, for a vaccine applied on school age -first dose or revaccination- we will have blocking, described through another splitting in fractions b and $1-b$. This will not happen in newborn vaccine, as newborns have not been exposed to mycobacteria in the moment of the vaccination.

Finally, an additional splitting is added in the infected individuals of each cohort: a fraction r will progress to disease, a fraction $1-r$ will not. The endpoint of the trials is measurement of disease, not infection, and that is why this last division is conceptually needed: however, it will not affect the results of the trials, as it is common to every one of them. This means that we consider that both vaccines and protections due to prior sensitization (masking) have the only effect of reducing the infection risk at the level of single individuals, and not to reduce the progression rates from latency to disease. Though this is the simplest case to analyze, vaccines and prior sensitization might also affect the progression to disease. The problem is that to be capable to discriminate between effects over infection or effects over progression to disease, data from trials with several endpoints would be needed (i.e. trials registering final fractions of sick and latently infected individuals, like done in Worcester for the novel vaccine MVA85A [213]), which unfortunately is not the case: the only trials that have been specifically designed to look for masking and blocking effects are –up to our knowledge– the trials of Barreto et al., and they only consider the endpoint of disease (i.e. they only explore fractions of active disease after the follow-up period).

A schematic, yet comprehensive representation of the splitting model is shown in figure 11.1.

The masking effect, parametrized through $m\in[0,1]$ will be, in principle, an increasing function of the age of the individual -in the case of a trial, a function of the

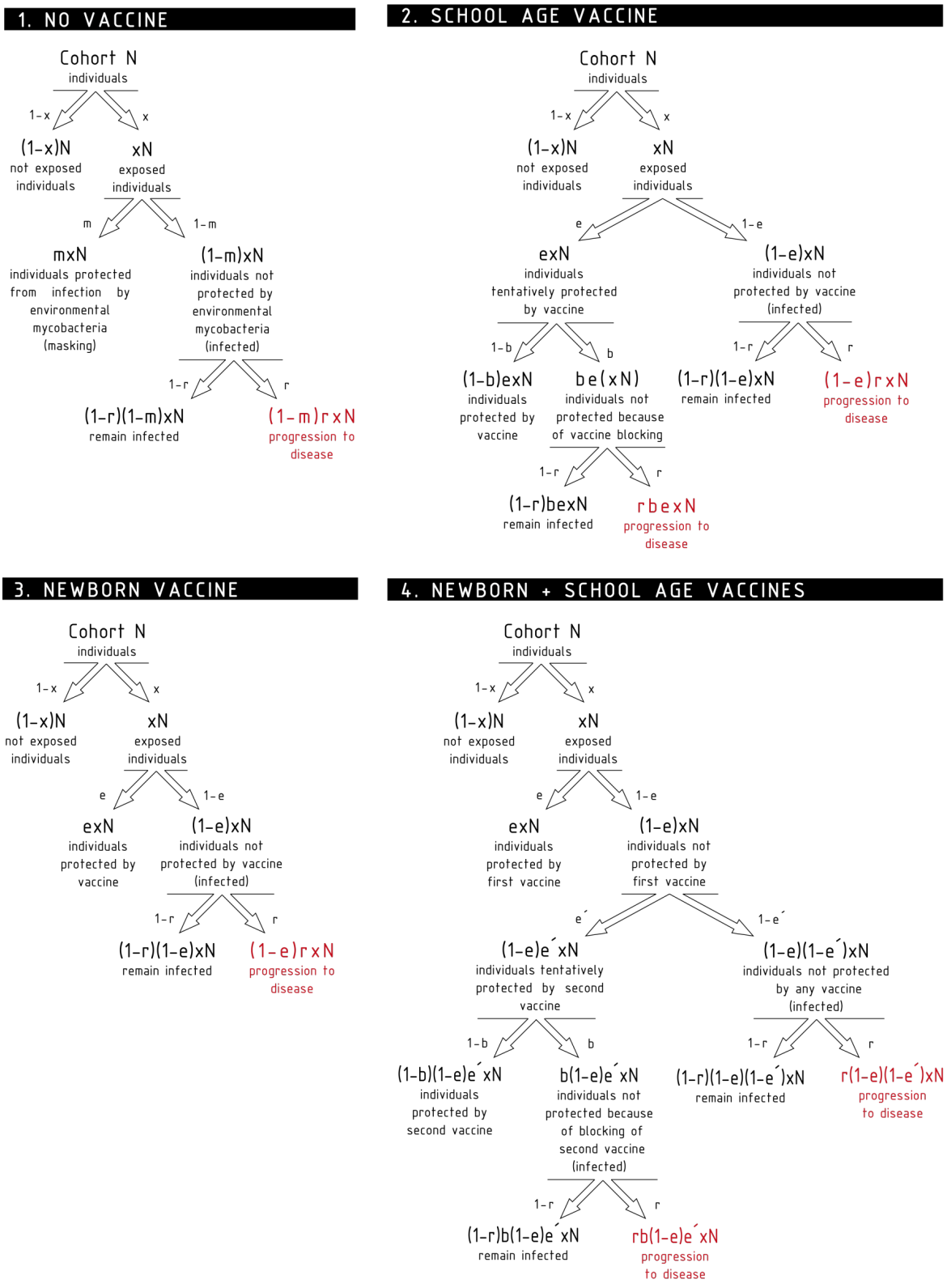


FIGURE 11.1: Schematic representation of the model of the masking and blocking effects in the different cohorts under study

average age of the individuals during the follow up period: 15 years old for our case [195]-. For the blocking effect -described by $b \in [0,1]$–, we have an slightly different situation, as it is an increasing function of the age at the moment of the vaccination: 10.5 years in these trials as a raw average of the limiting ages of the vaccinated children (7 and 14 y.o.[195]). Both, masking and blocking, should be different for each city -higher in Manaus than in Salvador- as the exposition level to mycobacteria differs due to the lower latitude of Manaus.

The proportions of people that finish the trial being sick in each cohort (D_1:non vaccinated, D_2:newborn vaccination, D_3: first dose at school age, D_4: revaccination at school age) are given by:

$$D_1 = (1 - m(15))rx \tag{11.1}$$

$$D_2 = (1 - e(15))rx \tag{11.2}$$

$$D_3 = (1 - e(4.5))rx + rb(10.5)e(4.5)x \tag{11.3}$$

$$D_4 = rb(10.5)(1 - e(15))e'x + r(1 - e(15))(1 - e')x \tag{11.4}$$

The efficacy of a clinical trial is determined by $\bar{e} = 1 - \frac{D_v}{D_c}$ where D_v and D_c are the proprortions of sick individuals the vaccinated group and the control group. Using our branching model, it is straightforward to derive the following expressions for the efficacies measured by the trials of Barreto et al. as a function of vaccine efficacies, blocking and masking effects.

$$\bar{e}_{1,M} = 1 - \frac{D_2}{D_1} = \frac{e(15) - m_M(15)}{1 - m_M(15)} \tag{11.5}$$

$$\bar{e}_{2,M} = 1 - \frac{D_3}{D_1} = \frac{e(4.5)(1 - b_M(10.5)) - m_M(15)}{1 - m_M(15)} \tag{11.6}$$

$$\bar{e}_{3,M} = 1 - \frac{D_4}{D_2} = \frac{e(4.5) - e(15)}{1 - e(15)}(1 - b_M(10.5)) \tag{11.7}$$

$$\bar{e}_{1,S} = 1 - \frac{D_2}{D_1} = \frac{e(15) - m_S(15)}{1 - m_S(15)} \tag{11.8}$$

$$\bar{e}_{2,S} = 1 - \frac{D_3}{D_1} = \frac{e(4.5)(1 - b_S(10.5)) - m_L(15)}{1 - m_S(15)} \tag{11.9}$$

$$\bar{e}_{3,S} = 1 - \frac{D_4}{D_2} = \frac{e(4.5) - e(15)}{1 - e(15)}(1 - b_S(10.5)) \tag{11.10}$$

where numeric subscripts indicate newborns vaccine (1), first dose at school age (2) and second dose at school age (3), respectively. In turn M stands for Manaus and S for Salvador. Thus, we have a system of 6 equation with 6 variables ($e(15)$,

$e(4.5)$, $m_M(15)$, $m_S(15)$, $b_M(10.5)$ and $b_S(10.5)$). As all these parameters are defined as fractions of individuals, they will be ranged between 0 and 1.

11.2.2 Masking and blocking quantification

Once we have obtained the system of equations that depends on different parameters of masking, blocking and base efficacy, the next step is to fit these parameters to obtain the best possible agreement with the data of the trials.

So as to do that task, let us consider a certain set of parameters $e(15)$, $e(4.5)$, $m_M(15)$, $m_S(15)$, $b_M(10.5)$ and $b_S(10.5)$ that produces a corresponding set of efficacies as predicted by our model $\bar{e}_{i,j}^{model}$. For every individual trial, we perform an elementary hypothesis test according to which the probability for $\bar{e}_{i,j}^{model}$ being compatible to the corresponding Barreto's observation $\bar{e}_{i,j}^{data}$ is assimilable to:

$$p_{i,j} = 1 - 2\frac{1}{\sqrt{2\pi}\sigma_{i,j}^{data}} \int_0^{|Z_{i,j}|} e^{-Z'^2_{i,j}} dZ' \tag{11.11}$$

where $Z_{i,j}$ corresponds to the Z-score associated to the model prediction of trial type i in city j:

$$Z_{i,j} = \frac{\bar{e}_{i,j}^{model} - \bar{e}_{i,j}^{data}}{\sigma_{i,j}^{data}} \tag{11.12}$$

and $\sigma_{i,j}^{data}$ corresponds to the typical deviation associated to the gaussian, here assimilated to one half the radius of the confidence interval reported by Barreto et al. As these intervals are not symmetric, we have:

$$\sigma_{i,j}^{data} = \begin{cases} \frac{\bar{e}_{i,j}^{data} - \bar{e}_{i,j}^{data,low}}{2} & \text{if } Z_{i,j} \leq 0 \\ \frac{\bar{e}_{i,j}^{data,high} - \bar{e}_{i,j}^{data}}{2} & \text{if } Z_{i,j} > 0 \end{cases} \tag{11.13}$$

where $[\bar{e}_{i,j}^{data,low}, \bar{e}_{i,j}^{data,high}]$ is the confidence interval reported by Barreto et al. for each trial. So, it is trivial that p_i will be maximum -equal to 1- if the exact data of the trial is reproduced, and it will decrease with the difference between the real observation and the value given by a set of parameters according to our model. From these measures we construct a global error function $\mathcal{Z}$ as follows:

$$\mathcal{Z} = 1 - \prod_i \prod_j p_{i,j} \tag{11.14}$$

Which can be interpreted as the probability of at least one trial reporting values incompatible to the model. We search the set of parameters that gives us the minimum possible value of $\mathcal{Z}$ -for the moment, using a Levemberg-Marquadt algorithm [413]-, and we call it $\vec{p}_0$. Thse values are the following:

$$\vec{p}_0 = (e(4.5), e(15), m_M(15), m_S(15), b_M(10.5), b_S(10.5)) = (0.60, 0.40, 0.01, 0.07, 0.88, 0.41) \tag{11.15}$$

which seems to indicate that blocking is the driving effect for the observed variations in BCG efficacies.

However, LM algorithm only guarantees that $\vec{p}_0$ represents a local minimum for $\mathcal{Z}$. Therefore, if we are in a situation when there are several local minima of $\mathcal{Z}$, the solution given in the previous section could not be considered as a solid answer and we are forced to look after every possible minimum in the landscape of $\mathcal{Z}$

Considering the asymmetric character of the confidence intervals by Barreto et al. makes $\mathcal{Z}$ be strictly non-differentiable. For this reason, although an analytic addressing of all the eventual minima of $\mathcal{Z}$ is feasible, such an operation should be repeated 2^6 times, i.e. one for each of the regions of the parameter space in which the piece-wise definition of $\sigma_{i,j}$ generates a different version of $\mathcal{Z}$.

Instead of doing that we consider reasonable to follow a numeric procedure based on mapping the error landscape on a network. To explore the total space of parameters we construct a directed network, in which nodes are points of the parameters' space and edges represent proximity between two nodes. The algorithm used to construct the network is the following (see figure 11.2, panel A):

- We divided the total space of parameters in a grid, and we find the minimum of $\mathcal{Z}$ inside every different box. Each one of these points will be a node. We define these nodes by iteratively dividing the hyper-cube in other smaller ones, a fixed number of times, retaining each time, the cube associated to the smallest $\mathcal{Z}$ found at each step.

- We represent a edge between nodes if the difference between them for every one of the six parameters is less than the side of a box.

- The edges will be directed, from the highest to the lowest value of $\mathcal{Z}$

Following this algorithm, two nodes corresponding to consecutive boxes are not necessarily linked. If they are attracted by different basins,they might be attracted in different directions, so the distance between them could be bigger than the side of a box, and therefore, they would not be linked. This construction could drive to different communities -or even disconnected components- for every minimum of $\mathcal{Z}$ (panel B in figure 11.2. In addition, if the resolution of the grid is enough, nodes with $k_{out} = 0$ can be associated to local minima and, ultimately, an overall outlook of the energy landscape can be achieved. The complex network for our model is represented in figure 11.2, panel C. Fortunately, the network presents only one minimum; which corresponds to a unique solution for the set of parameters, that is precisely the one found using the LA algorithm, which corresponds to a value $\mathcal{Z} = 0.27$.

11.2.3 Confidence intervals

Once we have seen that there is only one minimum in our parameter space ($\vec{p}_0$), it is time to study the local surroundings of the minimum found. In a local region near

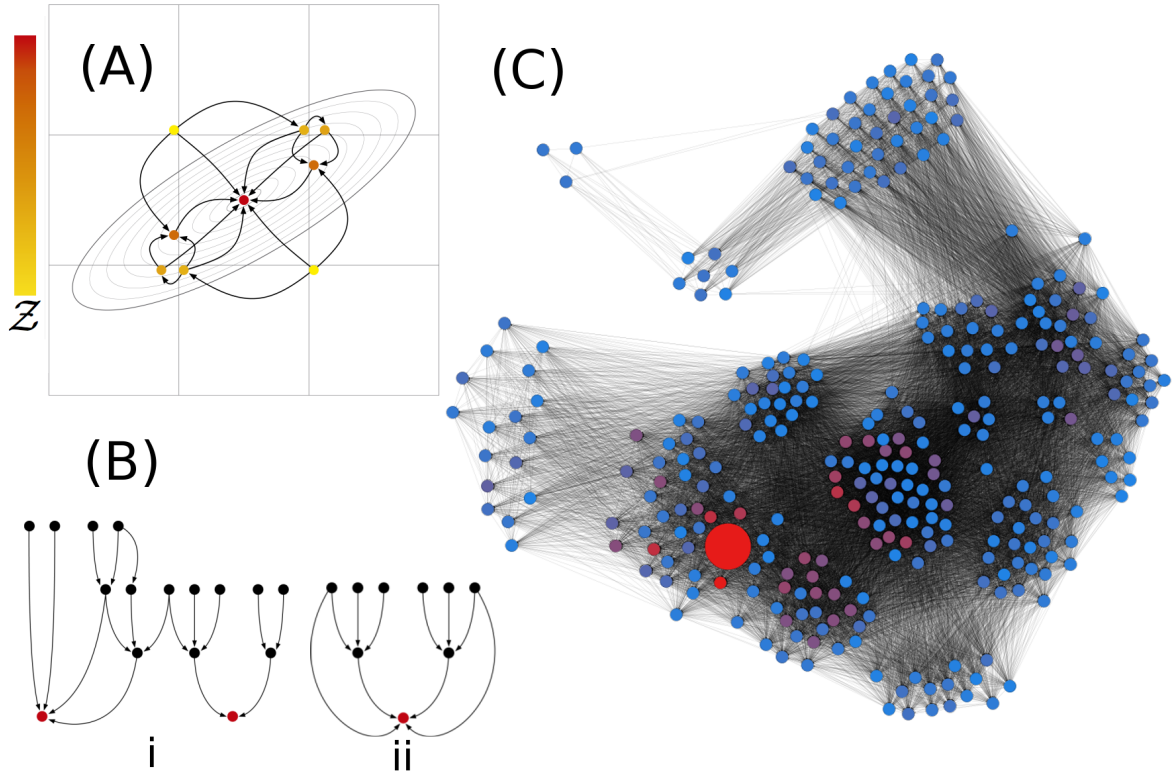


FIGURE 11.2: (A) Network reconstruction of error function $\mathcal{Z}$ landscape. A local minimum of $\mathcal{Z}$ is found in the interior of each of the hyper-cubes (six-dimensional) schematized in the figure. Then, directed links are stablished pointing from point of higher $\mathcal{Z}$ to points of lower $\mathcal{Z}$, provided they are closer than the size of the hypercube in the six axis. B) Two possible scenarios of the complex network: a unique solution (left) or more than one minima (right) (C) Error function $\mathcal{Z}$ landscape network for the masking-blocking problem. The network is is a fully connected graph with only one minimum-like node -representing with a bigger size-. Color of a node represents the value of P, red being the minimum and blue the maximum. Only points that corresponds to $P < 0.95$ are represented

$\vec{p}_0$ the surfaces of constant $\mathcal{Z}$ can be approximated to hyper ellipsoids -generally not aligned with the axis of the parameters [437]-, as it is shown in figure 11.3. To obtain an estimation of the uncertainties of the parameters, the following numerical procedure is performed.

First, we change the value of every parameter individually until a value of $\mathcal{Z} = 0.95$ is reached (see figure 11.3). We call this increment A_i ($i \in [1, 6]$). These values will not be symmetrical, and we will distinguish between A_i^+ and A_i^-. Then, we construct an asymmetrical Gaussian distribution for every parameter, centered in $p_{i,0}$ and with a variance given by $\sigma_i^{\pm} = cA_i^{\pm}$ at each side.c represents here a modulating factor equally affecting all the variances. Iteratively, we generate sets of points in the parameter space for which the coordinates in each parameter's axis are generated using the mentioned asymmetric gaussian distributions. Through an iterative process we search the value of c, for which a 95% of the points verifies $\mathcal{Z} < 0.95$. The uncertainty of each parameter will be given by $\pm 2\sigma_i^{\pm}$.

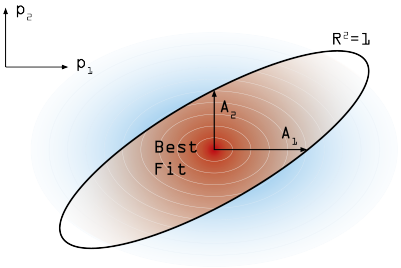


FIGURE 11.3: Schematic representation in a bidimensional parameters' space of the ellipsoids of constant $\mathcal{Z}$

Parameter	Value	Uncertainty range
$e(4.5)$	0.60	(0.54-0.64)
$e(15)$	0.40	(0.36-0.43)
$m_M(15)$	0.01	(0-0.08)
$m_S(15)$	0.07	(0-0.15)
$b_M(10.5)$	0.88	(0.70-1)
$b_S(10.5)$	0.41	(0.32-0.53)

TABLE 11.2: BCG masking, blocking and intrinsic efficacy estimations along with their confidence intervals

The parameters and their uncertainties obtained after this process are shown in table 11.2.3. As we can see, the levels of masking predicted are low -1% and 7%-, and their totally overlapping confidence intervals contain the no-masking scenario. On the contrary, blocking is remarkably higher -88% and 41%- as it was qualitatively foreseen by Barreto et al. [195]. It is also observed that the level of blocking is consistently higher in Manaus than in Salvador, as it was expected.

11.2.4 Universality of intrinsic efficacies

In addition to quantifying masking and blocking in the cities under study, our analysis of Barreto's trials has also allowed us to obtain two values of the intrinsic efficacy of BCG at two different times since vaccination -on average- $e(4.5) = 0.6$ and $e(15) = 0.4$. These intrinsic efficacies are defined as universal magnitudes, not subject to a particular location (i.e. Manaus or Salvador) according to our model. The question of whether these measures represent a global property of BCG that could be extrapolated to the results of other trials is pertinent.

In order to answer this question, we have studied an exhaustive meta-analysis of BCG clinical trials by Mangtani et al. [194], from which we can recover data from

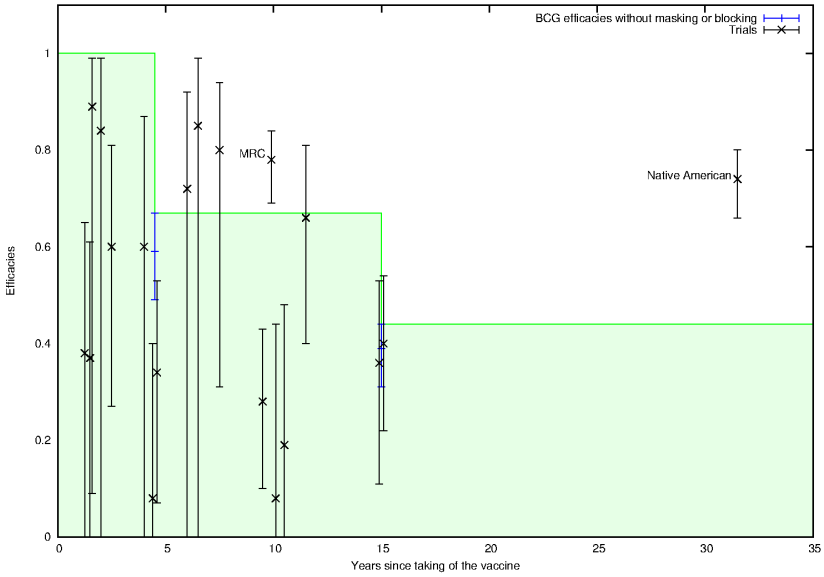


FIGURE 11.4: Results of different BCG clinical trials (black points) as a function of time since vaccination (average). The blue points are the two values of the base efficacy (without masking or blocking) obtained from the trials of Barreto et al [195]. The green line corresponds to the superior limit of the efficacy, if only blocking and masking occur.

19 different clinical trials aimed at determining BCG efficacy in different settings, countries and, very relevantly, in cohorts of individuals of different ages at different moments after vaccination. Using these data, we have represented, in a scatter plot, the efficacy determined by each trial vs. the time between vaccination time and the middle of the follow-up period (figure 11.4). If the only cause underlying variation of BCG efficacy were masking and blocking effects, and our estimation of the intrinsic efficacy from Barreto's data could be considered universal, then the two points determined by our model ($e(4.5) = 0.6$ and $e(15) = 0.4$) should constitute an upper bounds for the observed efficacies at similar ages.

In figure 11.4, the green line represents the superior possible value of BCG efficacy, defined by our base efficacy inferences ($e(4.5) = 0.6$ and $e(15) = 0.4$), coherent with such hypothesis. As we can see, we have found at least two trials whose results can not be explained by assuming that our intrinsic efficacies inferences are globally valid. This indicates that there exist other reasons behind the geographical variance of BCG efficacy additionally to blocking or masking, and that our estimations of the base efficacies have to be constricted to the local context of Barreto's trials.

11.3 Influence of blocking and immunity waning on novel vaccine's impact

In the last section, we studied the case of BCG vaccine, and concluded that, in a likely scenario, high levels of blocking are able to explain the variability observed for the efficacy of BCG vaccine in the trials by Barreto et al. [195], with residual (or null) levels of masking acting on the population.

When facing the problem of implementing a novel vaccine, the situation is different, as protection due to BCG has to be considered as a masking-like background protection acting on un-vaccinated people with the novel preventive drug. BCG is massively applied in virtually all the countries that face minimum levels of TB burden, and so, the efficacy and impact of any novel vaccine (substitutive or not) should be evaluated in comparison to BCG performance. Taking this in mind, our main purpose in this section will be that of offering a faithful description of two paradigmatic vaccines, when compared to BCG: a BCG-substitutive vaccine applied on newborns and a BCG booster applied on adolescents. Our intention is to use the model for vaccine efficacy estimations used in the last section to estimate plausible age-dependent efficacy profiles to introduce in our TB spreading model, developed in chapters 9 and 10, in order to compare the impacts of each vaccine.

First, we conceive the efficacy evaluation of a novel prime vaccine applied on newborns of base efficacy $e_{prime}(t)$, when compared to newborn vaccination with BCG, whose base efficacy we denote as $e_{BCG}(t)$. The argument of these functions is time after vaccination. In other words, we are comparing two identical cohorts of type 2 (i.e. newborn vaccinated, see figure 11.1), in each of which we apply a different vaccine, t years after vaccination. Since the vaccines we are modeling from chapter 9 on act by reducing infection probability, in order to measure their efficacy we will work with fractions of infected individuals. In this case, the relevant fractions will be:

$$I_{2,prime}(t) = (1 - e_{prime}(t))x \tag{11.16}$$

$$I_{2,BCG}(t) = (1 - e_{BCG}(t))x \tag{11.17}$$

and we will have that the measured efficacy of our novel vaccine as a function of age $\bar{e}_{prime}(t)$ is:

$$\bar{e}_{prime}(t) = 1 - \frac{I_{2,prime}(t)}{I_{2,BCG}(t)} = \frac{e_{prime}(t) - e_{BCG}(t)}{1 - e_{BCG}(t)} \tag{11.18}$$

Since, in this case, time after vaccination coincides with individuals' age a, we have:

$$\bar{e}_{prime}(a) = \frac{e_{prime}(a) - e_{BCG}(a)}{1 - e_{BCG}(a)} \tag{11.19}$$

The case of booster vaccination on adolescents (i.e. at average age of 17.5 y.o.) is subtler. First, we have to assure a proper map between the base efficacy of the newborn vaccine $e_{prime}(t)$ and the corresponding base efficacy of the booster vaccine $e_{boost}(t)$ so as

to guarantee that the two vaccines being compared are intrinsically identical. In order to do so, first step is to calculate the corresponding fractions of infected individuals for the booster case t years after second vaccination in a blocking-free scenario, taking into account that, this time, we compare a cohort of type 4 (i.e. newborn first vaccination and posterior booster, see figure 11.1) to a cohort of type 2 (newborn vaccination):

$$\begin{aligned} I_{4,BCG+boost}(t) &= (1 - e_{BCG}(t'))(1 - e_{boost}(t))x & (11.20) \\ I_{2,BCG}(t') &= (1 - e_{BCG}(t'))x & (11.21) \end{aligned}$$

Where t' is the time after BCG vaccination, which is $t' = t + 17.5$, so we have:

$$\begin{aligned} I_{4,BCG+boost}(t) &= (1 - e_{BCG}(t + 17.5))(1 - e_{boost}(t))x & (11.22) \\ I_{2,BCG}(t + 17.5) &= (1 - e_{BCG}(t + 17.5))x & (11.23) \end{aligned}$$

To assure that the prime vaccine (BCG substitutive) and the booster are identical, we impose that, t years after the end of each vaccination setup, and in case we have no blocking, both must protect the same fraction of individuals, which implies $I_{2,prime}(t) = I_{4,BCG+boost}(t)$:

$$(1 - e_{BCG}(t + 17.5))(1 - e_{boost}(t))x = (1 - e_{prime}(t))x \quad (11.24)$$

which allows us to express $e_{boost}(t)$ as a function of $e_{prime}(t)$ as follows:

$$e_{boost}(t) = \frac{e(t)_{prime} - e_{BCG}(t + 17.5)}{1 - e_{BCG}(t + 17.5)} \quad (11.25)$$

At this point, we have that, in this blocking-free scenario, the novel prime vaccine, and the combination of BCG and its new booster are able to protect the same fractions of individuals from getting infected, when comparing at the same times after vaccination. This is a suitable basis for setting up the comparison between both vaccines, but, unfortunately, such a couple of intrinsically identical vaccines (they could ideally be the same one, if its application on both unvaccinated and BCG-immunized is safe) might present different observed efficacies as a consequence of the age of the individuals being vaccinated. As we saw in section 11.2.2 for the case of BCG, vaccine's blocking may act on the AFV booster. As a matter of fact, the effect of blocking has to be taken into account when acting on the booster. In case the booster is affected by a certain level of blocking b, we have:

$$\begin{aligned} I_{4,BCG+boost}(t) &= (1 - e_{BCG}(t + 17.5))(1 - (1 - b)e_{boost}(t))x & (11.26) \\ I_{2,BCG}(t + 17.5) &= (1 - e_{BCG}(t + 17.5))x & (11.27) \end{aligned}$$

And now, we can calculate the observed efficacy for the booster vaccine $\bar{e}_{boost}(t)$ as follows:

$$\bar{e}_{boost}(t) = 1 - \frac{D_{4,BCG+boost}(t)}{D_{2,BCG}(t+17.5)} = e_{boost}(t)(1-b) \tag{11.28}$$

And, substituting $e_{boost}(t)$ from eq. 11.25, we have:

$$\bar{e}_{boost}(t) = \frac{e_{prime}(t) - e_{BCG}(t+17.5)}{1 - e_{BCG}(t+17.5)}(1-b) \tag{11.29}$$

Finally, taking into account that the booster is applied when individuals are 17.5 y.o., we recover the following explicit age-dependency:

$$\bar{e}_{boost}(a) = \begin{cases} 0 & \text{if } a < 17.5 \\ \frac{e_{prime}(a-17.5) - e_{BCG}(a)}{1 - e_{BCG}(a)}(1-b) & \text{if } a \geq 17.5 \end{cases} \tag{11.30}$$

Equations 11.19 and 11.30 define the observed efficacies of two identically performant vaccines, one of which is intended to substitute BCG, being applied on newborns, and the other one intended to boost BCG, being used on adolescents. The differences between the observed efficacies –which are measured in counterposition to BCG's performance– arise as a consequence of the ages on which they are being applied, more specifically, as a consequence of 1) the pattern of immunity waning over time of both BCG and the novel vaccines ($e_{BCG}(t)$, $e_{prime}(t)$, and $e_{boost}(t) = f(e_{boost}(t))$) and 2) the differential effect introduced by blocking, which only acts on the booster as a consequence of the older age of the individuals receiving it. In the following subsections we explore different scenarios that contemplate the effects on observed efficacy and vaccine impacts of dealing with vaccines following different patterns of immunity waning and/or subject to different levels of blocking.

11.3.1 Fast waning vaccines

As a first case scenario, let us suppose that both BCG ad the novel vaccines follow the same temporal waning pattern identified in section 11.2.2 for BCG in Manaus and Salvador. In section 11.2.2 we found: $e_{BCG}(4.5) = 0.6$ and $e_{BCG}(15) = 0.4$. As a first approximation, we consider a linear age-dependence to get, for BCG:

$$e_{BCG}(a) = max(0, -0.02 \cdot a + 0.69). \tag{11.31}$$

When a stands for the individuals' age, supposed they received BCG at birth. Then, let us assume that the intrinsic efficacies of the novel vaccines are greater at $t = 0$, but they wane at the same linear pace of BCG. For the prime vaccine, we have $e_{prime}(a) = max(0, -0.02 \cdot a + 0.94)$, when the independent term has been calculated so as to make the vaccine have a 70% observed efficacy at the first age group $a \in [0, 5]$. In what regards $e_{boost}(a)$, we calculate it from $e_{prime}(a)$, following equation 11.25.

The intrinsic and observed efficacies for each vaccine are represented in figure 11.5, panels A and B. In that figure, for the case of booster vaccines, three scenarios have been explored: high blocking (i.e. $b = 0.84$, like BCG in Manaus), medium blocking ($b = 0.4$, like BCG in Salvador) and no blocking. In turn, in panel C we observe the impact

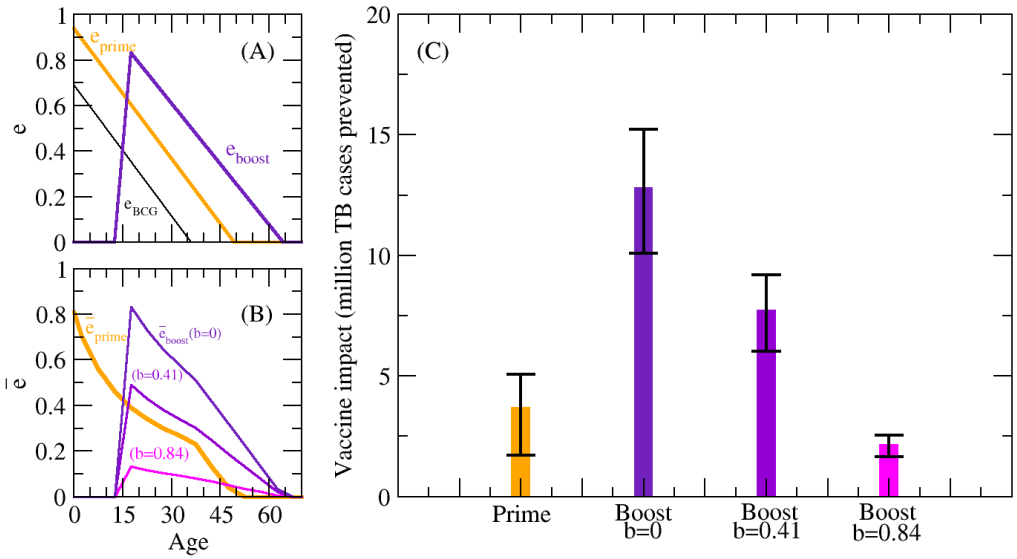


FIGURE 11.5: A. Intrinsic efficacies of different waning vaccines (BCG: black, novel prime vaccine: orange, novel booster vaccine: purple) B. Observed efficacies of novel vaccines (orange: prime, purple, violet and magenta: booster). Different levels of blocking reduce the observed efficacy of booster vaccines, when applied on adolescents. C: forecasted impacts of each vaccine, as predicted by our TB spreading model.

foreseen by our spreading model, in which, ϵ parameters are assimilated to 1 minus the observed efficacies $\bar{e}$. At this point, it is worth mentioning that this assimilation $\epsilon = 1 - \bar{e}$ makes sense because of that the non-immune branch of our spreading model is assimilated to the regular population, so it includes BCG vaccinated individuals, who virtually constitute the entire population in most high burden countries. This means that our fitting procedure of the spreading model is embedding the BCG effects, and so, the parameter ϵ represents the outperformance of any novel vaccine with respect to BCG, just as desired. Additionally, it is worth mentioning that, in this case, we are describing immunity waning by a progressive decay in vaccine's efficacy (i.e. an increasing of ϵ), and so, no recovery fluxes from VS to W classes are needed in the spreading model (i.e. $\gamma = 0$). Of course, if we consider a vaccination strategy focused on a traget age group a_o, a hypothetical subsequent age-group $a_1 = a_o + \tau$ will only be protected by the vaccine τ years after the vaccination campaign, which we set in 2020.

Many relevant observations can be noted from figure 11.5. First, after a careful observation of the intrinsic efficacies (panel a), we see that, leaving apart the 15 years shift among them, they are not perfectly identical, which would seem to be contradictory with the premise of both vaccines being intrinsically identical. What is actually happening is that, while the prime vaccine acts alone, the booster vaccine cooperates with BCG in order to provide protection to the same fraction of the population (i.e. to be identical, see equation 11.24), and therefore, so as to protect the same fraction of people, its efficacy e_{boost} must be lower than e_{prime}.

Additionally, even if prime and booster vaccines have similar intrinsic efficacies, if no blocking is acting, the pattern of time waning for the observed efficacies (panel b) is remarkably different. The reason, this case, is a consequence of the different protection levels provided by BCG when displacing 15 years in the horizontal axis. Indeed, observed efficacies are different because these are calculated from the comparison provided against sole BCG, which is different in both cases because of the 15 years shift. In this case, BCG efficacy decays with age, which supposes that it is more difficult to overcome in the first age groups: this translates in higher observed efficacies for the booster vaccine than for the prime when comparing age groups separated 15 years.

The final impact for each vaccine, in terms of the number of active TB cases prevented up to 2050 in SEAR after the introduction of a vaccine in 2020 is represented in panel C. As we see, when no blocking is present, an adolescent-focused booster vaccine will offer a larger impact than a prime vaccine on newborns. This is the result of two concurrent factors. First, as we discussed in chapter 10, adolescents and young adults are the main contributors to TB transmission, and so, it is of utmost importance to immunize them as fast as possible, something that is achieved by the AFV booster. Additionally, BCG is more difficult to overcome for the youngest segments of the population, which works in detriment to the NFV, again.

However, these effects, both advising the use of AFVs instead of NFVs, may be deactivated by vaccine's blocking. As we see in panel C, if the novel booster suffers from similar levels of blocking to those found for BCG in the trials by Barreto, the advantages mentioned before may be lost.

11.3.2 Novel persistent vaccines vs. fast waning BCG

In this case, we will explore a more optimistic scenario according to which the novel vaccine is able to maintain its persistence during the entire lives of vaccinated individuals; while the same behavior of section 11.3.2 is reserved for BCG (eq. 11.31). In this case, for the efficacy of new vaccines, we discard the linear terms that make them time-dependent to recover constant values for all ages, to get $e_{prime}(a) = 0.91$ and $e_{boost}(a) = 0.70$ for all individuals regardless how old they are. The intrinsic and observed efficacies for each vaccine are represented in figure 11.6, panels A and B, whilst the projected impacts for each vaccine are represented in panel C.

As we see, forecasted impacts are now higher due to the increased vaccine efficacies considered. Furthermore, the picture is very similar to that obtained in section 11.3.2, with the difference that now, as there is no efficacy waning for the novel vaccines, these attain an slightly improvement in the observed efficacy as BCG efficacy, to which they are compared, wanes. In this case, the out-performance of the booster vaccine with respect to the prime to be more modest than before: in this case, the impact by the booster with no blocking is 2.44 times larger the impact generated by the prime vaccine; while, if the vaccines show the same fast waning behavior of BCG, this coefficient is 3.5. This is due to the fact that, in this case, the prime vaccine preserves its entire performance once it reaches the age segments more affected by the disease (adolescents and young adults), on which BCG is not working any more: the advantage of the

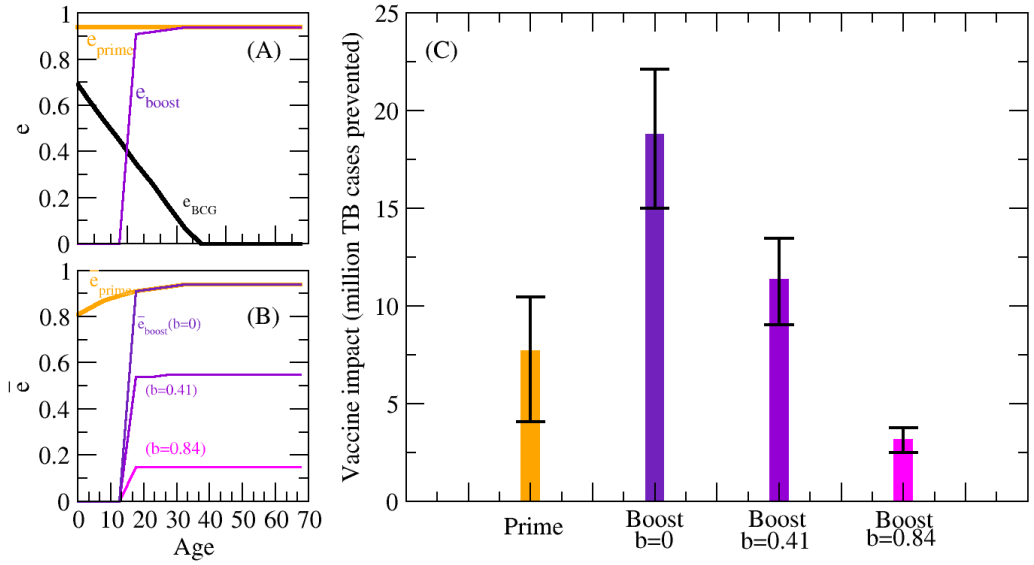


FIGURE 11.6: A. Intrinsic efficacies of different persistent vaccines (BCG: black (waning), novel prime vaccine: orange, novel booster vaccine: purple) B. Observed efficacies of novel vaccines (orange: prime, purple, violet and magenta: booster). Different levels of blocking reduce the observed efficacy of booster vaccines, when applied on adolescents. C: forecasted impacts of each vaccine, as predicted by our TB spreading model.

booster, now, is reduced to a consequence of immunizing faster these age-groups.

11.4 Discussion

In this chapter we have explored the influence of sensitization to environmental mycobacteria and immunity waning to the observed efficacy of anti-TB vaccines. Capitalizing on the results from recent clinical trials conducted in Brazil [195], we have quantified, through a simple mathematical model, the relevance of the principal mechanisms proposed, concluding that blocking of vaccine assimilation is the most likely mechanism responsible for BCG reduced efficacy when applied on school-age individuals, specially in lower latitudes.

Additionally, in chapter 10, we conducted impact comparisons between AFVs and NFVs, when both of the them showed identical observed efficacies. On that basis, in this chapter we have characterized the way in which two identical vaccines will present different levels of observed efficacies when used on different age groups. The reason for this is twofold. On the one hand, BCG background efficacy may be different in different age groups, and its protection may not be equally easy to overcome at all ages. Typically, BCG is considered to perform better on children, and losses its efficacy on older age groups, which would derive in AFVs having more observed efficacy than NFVs. On the contrary, AFVs may have to face vaccine blocking, an issue that is not expected to happen for NFVs. All these aspects make even less obvious the debate between AFVs and NFVs, as the result of the comparison between their observed efficacies will be the outcome of a final tradeoff between the two opposite factors mentioned before. In this chapter, we have developed a modeling platform for the quantitative evaluation of vaccine's impacts taking into account all these effects.

As a result, the major conclusions derived from our study are clear: first: a global effort to reliably measure age-dependent patterns for the efficacy of BCG are urgently needed, if novel vaccine impacts have to be undertaken, as the efficacy of them crucially depends on the background levels of protection provided by BCG. Furthermore, blocking effect, that solely affects to AFVs, may dramatically reduce their impacts, eventually causing AFVs to lose their initial advantage derived from reaching faster the most infectious age segments of the population. At this point, it is worth noticing that the vaccine impact evaluations accomplished in this chapter constitute just some relevant examples of how these effects may affect vaccines impacts, and other patterns could be observed when testing and implementing novel vaccines. For example, also the worst case in which time persistence of novel vaccines were lower than that of BCG could apply, and, being the persistence of BCG itself a controversial matter, in some areas, the fast-waning efficacy that we have used to describe BCG behavior may not apply.

As a conclusion from this chapter, as we can see from the studies by Barreto and Mangtani, the question of what are the causes for BCG efficacy variation among different parts of the world is probably complex enough so as not to admit a unique answer. In this section we have studied masking and blocking effects, capitalizing on the specifically designed trials by Barreto et al., concluding that blocking constitutes a major effect impairing BCG performance when applied on grown individuals, instead of masking, which is discarded as a plausible mechanism by our analysis. Notwithstand-

ing that, the extrapolation to these results to other trials is not easy, as particularly the intrinsic efficacies derived from the analysis of Barreto's data are not totally compatible with some of the outcomes of the trials meta-analyzed in [194]. As a conclusion, further efforts and clinical trials have to be conducted in order to achieve a deep understanding of the actual protective role currently played by BCG at different age segments in the different parts of the world. Provided that any future novel vaccine's performance should be compared to that of BCG, that unarguably constitutes an unavoidable task to accomplish if reliable impact estimations for novel anti-TB vaccines are intended, as we have seen in this chapter.

Part VI
Conclusions

Metáforas que apuntan hacia esa vaga, incitante dirección: el latigazo de la triple carambola, la jugada del alfil que modifica las tensiones de todo el tablero.

Julio Cortázar.
Territorios, 1978.

Chapter 12
Conclusions and prospects

Tuberculosis is an extraordinarily old companion of mankind. It was with us before we were able to write, and it remains with us now, in the digital era [92]. The discovery of its microbiological cause, by Robert Koch in 1882, constituted one of the most relevant milestones of the young discipline of medical microbiology, whose revolutionary paradigm have had a tremendous impact in public health, from Koch and Pasteur days to the end of XXth century. Notwithstanding that, it is precisely the long endurance of our relationship with *MTB*, the cause of most misconceptions about the disease, which difficult the labour of both the researcher and the clinician. As Susan Sontag wrote [431], when infectious diseases reach epidemic proportions before a proper description of its nature is available, peoples hit by it provide an explanation for its causes and enable it with complex, metaphoric meanings through non scientific narratives collectively constructed. This commonly includes supernatural elements and a component of judgement associated to punishable conducts, that contribute to the creation of stigma. This has happened many times in History, from syphilis and TB to AIDS. Beyond the dramatic repercussions of the stigmatization of sick individuals, the collective construction of these *ideas of disease* has a negative impact on the social perception of risk associated with them, as they cease to constitute a threat of unknown reach to be an issue that *is somehow understood* and to which the community gets accustomed.

In this sense, the ancient nature of TB might invite us to rely in the apparent decline of the disease, and to think that current medical and epidemiological knowledge about it should be complete by now. However, as we have seen along this Thesis, the disease still remains a major cause of morbidity and death worldwide, and its complexity, at different scales of description, poses strong limitations that hinders not just our knowledge about the disease, but more importantly, our ability to fight it. In this sense, for example, the incomplete description of human immune response against TB hinders the task of finding precise immunological correlates for protection against infection, which in turn makes harder the evaluation of novel vaccines. Simultaneously, epidemiological studies and vaccine impact estimations are hindered by the complex life cycle of the disease, and the difficulty of its surveillance, –specially in the countries with a greater burden, which are those for which public health infrastructures are more fragile [432]– makes uncertain any burden measure. These drawbacks evidence that there is yet a long way to walk in TB research, at all scales from Biochemistry and cell Biology to immunology and epidemiology, and, –even more remarkable– that further developments in every scale will strongly influence each other. These observations, along with the enormous societal implications associated to TB burden worldwide, should be enough to motivate the idea, already developed by Young, Stark and Kirchner [2], among others [433] of that comprehensive, multidisciplinary research efforts

embracing all the disciplines mentioned around integrated approaches are needed to achieve a deeper knowledge about this old companion of mankind, and, which is more important, to maximize our chances of fighting, and ultimately eradicating it. This is the spirit in which this thesis finds its context.

Along the chapters 3 to 11, we have exposed the results of the research conducted in this thesis; in the following lines, we review the most relevant conclusions attained after these studies, the research threads they open and, remarkably, how these threads intertwine and how their simultaneous consideration around integrated approaches can contribute to address unsolved questions in TB research.

12.1 Cell interactomes in *MTB*

In chapters 3 to 6 of this thesis we have focused on the characterization of a bibliography-based TRN of *MTB* and on the study of the PPIN of the pathogen, and its adaptation to disparate environmental conditions. Furthermore, we have studied the effects of data incompleteness in the outcomes of different topological analysis conducted on TRNs, and how networks based methods can be used to deal with the uncertainty that is inherent to these and other complex directed networks.

The bibliographical revision presented in chapter 3 constituted, at the time the work was published, the most complete dataset on *MTB* transcriptional regulation, reporting more than three times the number of links than previous compilations [89]. This allowed us to perform a detailed topological analysis of the system, identifying most connected genes in the network and analyzing its overall topological properties, which resulted similar, widely speaking, to those of analogous systems in other organisms.

Remarkably enough, even if the heterogeneous character of the network compiled (which incorporates data of different quality and meanings) supposes a major limitation, the sole integration of the information available around a single dataset shed light about the roles played by each gene in the whole system. In this line, a network view can help to contextualize the relevance of each regulator, and, something which is more difficult to achieve otherwise, to identify genes whose expression has been reported to be affected by a greater number of transcription factors. In this sense, while it will not be surprising that *PhoP* appears as one of most prominent hubs in the network, it is remarkable that two of the genes receiving a higher number of links –*fadD26* and *icl*–, have been identified to play central roles in controlling *MTB* virulence and persistence, as we commented in chapter 2 (see figure 2.2). These genes show divergent expression in the attenuated strain *MTB*VAC, (*icl* is overexpressed and *fadD26*, as well as *PhoP*, deleted) which has been associated directly with its attenuated, immunogenic and persistent phenotype [129, 88, 186].

In the context of the topological analysis of the TRN of *MTB*, the analysis of motifs statistics in the network brought some interesting results. On the one hand, the TRN of *MTB*, in an first analysis, constituted an apparent exception to the classification proposed by Alon and collaborators in [267]. In that work, authors identify four main groups –or superfamilies– of complex networks according to their different contents in

network motifs, each of which formed by similar systems. Among the four superfamilies, two are formed by biological systems: a first group formed by TRNs of unicellular organisms –which *MTB* "should" have joined– and another family formed by regulatory networks of pluricellular organisms and synaptic networks, to which *MTB* seemed to belong. According to the interpretation of this division provided by Alon et al. in [267], the first family would be formed by "shallow" networks in which typical response times associated to single interactions are similar to the response times of the whole system, because of which they called it "rate-limited" network superfamily. By contrast, the situation is the opposite for the systems in the second family of "not rate limited" networks, in which systemic response times can be much longer than those of single interactions. According these observations it would seem that the fact that *MTB* was the first unicellular organism with a TSP more similar to those of the second family could be a consequence of its particular lifestyle, very particularly to its ability to survive within the host for long periods of time, which is a wrong conclusion as we later show in chapter 5. In fact, what we show in that second phase of our research is that it is the whole division of biological networks around two superfamilies of related motifs profiles which is probably wrong. In chapter 5 we see how, by analyzing updated versions of the TRNs of bacteria originally belonging to the family of rate limited networks (*B.subtilis* and *E.coli*) we see how their TSPs now become compatible to those of the non-rate-limited family. The cause of this phenomenon is methodological: since the null model used in [267] to determine triads significances in real networks consisted in random versions of the original systems preserving the number of in, out, and bi-directional links of each node; in case there are no bi-directional links in a real network, they will neither appear in the null-ensemble, which makes impossible to define the Z-score for motifs containing bi-directional edges. In this sense, TSPs of networks with no bi-directional links are affected by a lack of robustness against the addition of even a reduced number of them. This was the case of some of the networks analyzed in [267] as members of the first superfamily, that lacked bi-directional links in the datasets used by Alon and collaborators in 2004, gaining some of them during the following years as resulting from the progressive gain in data quality and completeness. In chapter 5 we show how a simple update of the systems analyzed in [267], according to the data publicly available on-line [314, 358] completely changed their TSPs, making them switch from the "rate limited" family to the "not-rate limited" one, which actually evidences that the division of biological networks of information processing around two families is an artifactual consequence of data incompleteness.

Admittedly, the former result does neither offer an answer to the causes behind the observed motif statistics in biological networks nor compromise the essential interpretation of Alon and co-workers about it, which is related to the existence of evolutionary pressures affecting networks' topologies at the level of motifs, related to their dynamical performance. In this sense, it has been hypothesized that the reason for which some motifs are more selected than others in real networks is related to the benefits associated to the dynamics they drive. For example, feed-forward loops act, depending on the signs and logical coupling of the links involved, as noise discriminators [261, 268] or response-accelerator devices [269], both useful to drive cell adaptation to unpredictable,

volatile environments. In order to indirectly test this hypothesis, we have analyzed the TRNs of *E.coli* and *MTB*, looking for motifs presenting more divergent statistics in both organisms. The results of our analysis show that *E.coli* shows a differential trend towards forming four nodes FFL-related structures, which is coherent with a lifestyle much more disparate and unpredictable than that of *MTB*. These "fine grain" observations would be compatible with the hypothesis of network motifs as a product of evolutionary pressures, and may constitute a feasible methodological resource to provide further testing of it, if similar results were robustly found on a substantial number of related systems (for example, groups of TRNs of intracellular parasites vs TRNS of free-life organisms).

It is worth mentioning that the hypothesis of networks motifs being the product of hidden evolutionary pressures have been largely discussed in the literature, and that there exist relevant alternative hypothesis to explain non-trivial motifs profiles, such as the existence of physical spaces on which networks are embedded [339], or the effect of systematic errors related to network's inference process [434]. In chapter 5, we test this last hypothesis in the precise context of TRNs, and the emergence of systematic correlations between deficient characterization of links and their topological local profiles. The conclusion of our analysis is that lack of robust experimental evidence of large fractions of the interactions in a gene network may cause Z-scores associated to the significance of some motifs to increase, even if overall significance profiles (i.e. the relative significance of each motif with respect to the others) remained remarkably stable.

In chapter 5 we also analyzed the mesoscale of the TRN of *MTB*, and its compatibility with the functional characterization of genes at different resolution levels. This way, we find how a greater correspondence between functional and topological partitions is attained when all the information available on the system (i.e. links directions and signs) is rendered available to the corresponding modularity optimization algorithms, no matter the size of the scales explored. This result lies in the same line that those recalled in previous paragraphs, and highlights the relevance of the quality of the datasets on the outcomes of different kinds of topological analysis.

Taking that in mind, we moved our focus to the analysis of current network-based approaches to distinguish spurious information in complex systems, a relevant topic in the field. More precisely, being interested in their use on TRNs, we developed a new method for finding missing and spurious links in directed networks, based on a previous work by R.Guimerá and M. Sales-Pardo devoted to the analysis of undirected systems [276]. Our model was compared to other methods currently available in the literature for solving the same problem, showing a slightly better performance when identifying randomly distributed errors, that may not justify the greater computational costs associated to our approach. However, these higher computational costs appear justified when the algorithms are used to detect systematic missing data, as those associated to inherently incomplete datasets, as *TRN*s built up upon series of independent experiments.

In this first half of the Thesis, devoted to the analysis of biological networks, we also present, in chapter 4, a study of the PPIN of *MTB*, and its transformations as a function

of gene expression. The results of our analysis shed light about the mechanisms of transcriptional adaptation that allow *MTB* to face different situations and stresses. On the one hand, we have shown how networks associated to expression profiles measured in different experiments can be analyzed through phylogenetic analysis to identify the influence on their topologies of the experimental procedures, strains used and stimuli types. This kind of analysis can help to discriminate among different experimental protocols those causing a bacterial state more tightly resembling a desired phenotype, something of relevance in a field in which lack of proper models of infection constitute a crucial drawback. On the other hand, our methodology is used to study how the PPIN adapts to some of the specific stresses that are thought to constitute the basis of biological hostility of phagosomal environment: hypoxia, nutrient starvation, generic oxidative stress, exposure to NO, key-ions deprivation and cell wall damage. Our analysis evidences strong similarities in the PPIN of *MTB* in response of all these stimuli, which suggests the existence of convergent regulatory pathways associated to the specific adaptation of the bacteria to its host, whereby the pathogen deploy highly similar responses to these stimuli, which appear highly correlated *in vivo*. This result can be interpreted in the context of a series of works in which these kind of convergent regulatory pathways are identified [227, 133].

Remarkably, network analysis allows the identification of a group of proteins that systematically act as hubs under all stresses mentioned. These proteins constitute six pairs of ESAT6-cfp10 analogous, whose central role in the PPIN under any of the stresses mentioned suggests a deep biological implication. The fact that our method is able to capture the central importance of the gene pair ESAT-6/cfp10 constitutes an encouraging argument evidencing the need of systemic analysis of global cellular adaptation beyond standard transcriptional characterization. Nevertheless, our method also predicts a comparable role for five additional gene pairs parents to ESAT-6/cfp10. These proteins have not received a comparable attention, and their biological role is assumed to be that of providing a source for antigenic diversity to the pathogen that could help it to control host' immune response [435, 436]. Summarizing, our results suggest that the proteins mentioned could play a role comparable to their most relevant cognates ESAT-6 and cfp10, as all of them are promoted to play equally relevant roles in the PPIN under any stress. This suggests that these dimeric proteins could play relevant roles as major antigens, as important as ESAT6-cfp10; an hypothesis that deserves further exploration.

Admittedly, several aspects of this work require further testing. On the one hand, the comparison between the consensus networks associated to different stresses must be reproduced in other organisms so as to elucidate whether their convergent topological profiles can be confidently related to its parasitic lifestyle or, instead, are found in other kinds of bacteria thus pointing to the existence of broader mechanisms of convergent stress sensing and adaptation valid regardless specific environmental adaptation strategies (or other effects). On the other hand, although, in general terms, genes' roles within the multi-layer system of stress response are a product of both their topological position in the original unweighted network and the gene expression profiles associated to each condition, it remains pendent to show what of these factors are more impor-

tant on defining, for example, the strengths and participation coefficients of each gene. These two issues are being addressed in the context of an on-going research.

12.2 Epidemic models

In the second part of this Thesis, in chapters 7 to 11, we focus on the development of epidemiological models to describe TB and other diseases sharing some of its principal characteristics, as these are long latency periods of persistence and strong interactions with other infections.

In this sense, in chapters 7 and 8, our aim was to extend the well stablished paradigm of mathematical epidemiology on complex networks towards epidemiological scenarios tightly related to TB. The complexity that characterizes the contact networks on top of which infectious diseases spread, as well as its effects on the epidemic phenomena, constitute matters extensively addressed in the literature. However, the puzzle can hardly be considered complete, as, for example, the interplay between complex topologies and disease persistence, or the effects of the co-existence of two (or more) interacting diseases spreading through complex networks constitute topics only partly addressed in the literature.

In this sense, in chapter 7, we present a mean-field model aimed at describing the dynamics of a disease that, like TB, simultaneously presents *primoinfection* and long latency periods; when spreading on both homogeneous and heterogeneous populations. By doing so, we are able to derive, numerical and analytically, the expression for epidemic thresholds, addressing the influence of the parameters related to disease's dynamics and that of those related to the demography. Doing so, we corroborate, within the range of application of our HMF, that the epidemic threshold also depends on the quotient $\langle k \rangle / \langle k^2 \rangle$. In the context of this work we have addressed a series of novel questions that arise from the interplay between network complexity and persistence, as for example, how the degree distribution of the system can be reshaped by the effect of the disease depending on how newborn individuals are incorporated to the system.

In turn, in chapter 8, we present a mathematical model for the description of the simultaneous spreading of two interacting diseases on one population, through two distinct complex networks of contacts. In our model, special attention is paid to the description of the different mechanisms of interaction that can be observed between two diseases, like variations in the infectiousness, recovery rate or susceptibility to the disease. In this sense, by exhaustively considering all the possible interaction routes, we have been able to identify overall conditions for "positive" or "negative" interactions -disease enhancement or impairment, in terms of epidemic threshold shifts-, and to foresee the possibility that one disease can enhance or impair the spreading of another one as a function of its prevalence level. In addition, our description, simultaneously valid for SIS and SIR models, show that epidemic thresholds differ for both models not just as a consequence of dynamical correlations, (intrinsically disregarded in the HMF formulation) but as an effect of disease-disease interactions. Additionally, our model provide a description of epidemic thresholds within and outside equilibrium. That, as

we show in Figure 8.8, implies an understanding of how the epidemic threshold for one disease changes as the outbreak of a conjugate infection evolves. Such a dynamical description of epidemic thresholds makes explicit the relevant influence of timing on the critical properties of a system of interacting diseases. Finally, we have shown how our approach, despite its simplicity, is able to reproduce, at least in a qualitative way, real burden data corresponding to a well known system of interacting diseases: the *syndemics* formed by HIV and TB, using data from the Republic of South Africa, arguably one of the places in the world more dramatically affected by the concurrence of both diseases. By performing a simple fitting procedure, our model is able to provide a suitable reproduction of the time series of TB and HIV prevalence, from 1990 to 2010, when the interaction between them is known to have constituted the main driving effect responsible of the enhancement in TB burden rates. By doing this exercise, more importantly than showing that our model is able to reproduce real data (which, as we say, could hardly be considered surprising), it is to observe that, despite our model is a strong simplification of actual natural histories of both TB and HIV, the parameters yielding the best fit consistently reflect the interaction mechanisms whereby HIV modifies TB spreading dynamics; including the enhancement of the susceptibility to TB disease of HIV infected individuals [157] and the reduced TB transmission rates of patients co-infected with HIV [146]. This indicates that collective fitting in epidemic models of interacting diseases –an approach which has been demonstrated of usefulness in other grounds [437]– may help us to discriminate between the actual mechanisms of interaction between them.

Once we have explored the interplay between topological complexity of contact networks and different dynamical aspects of the epidemic spreading (like persistence or interactions between diseases), we specifically focused on TB spreading modeling in chapters 9 to 11. In this part of the Thesis we analyzed the state of the art in TB modeling and incorporated to the picture some new ingredients inspired from what was learnt in previous chapters and, in general, from what is customarily being done in other grounds of computational epidemiology. The aim is to test, at a quantitative level, the influence of these new hypothesis on model outcomes, as well as the feasibility of their incorporation to the models; with the final purpose in mind of increasing reliability of current modeling platforms for the impact evaluation of novel epidemic interventions; very remarkably preventive vaccines.

In this context, in chapter 9, we developed a data-driven model for the evaluation of TB burden, based upon a series of works by C.Dye et al. Our model processes two distinct sources of data. First, it uses demographical projections regarding populations' age distributions provided by UN population division [23]. Additionally, incidence and mortality rates are extracted from the WHO tuberculosis database [24]. The principal conceptual novelties of our model are four: 1) the use of heterogeneous contact patterns to drive disease transmission, 2) the explicit consideration of the coupling between disease dynamics and demographic evolution, 3) the abandon of the hypothesis of initial equilibrium and 4) The introduction of a fitting procedure (in chapter 10) that is able to reproduce age-distributed burden levels.

Our motivations to introduce these ingredients on the already sophisticated models

of TB spreading have to do with the current scientific context in the field. As we say, the importance of being able to provide reliable impact evaluations for the novel TB vaccines currently under development constitutes an strong reason advising the need of revising current models, and to assure that all the quantitative data resources currently on disposal are taken into account in model development. In this sense, as we have discussed in the text, the age at which novel vaccines are intended to be administered constitute a central question to address, and consequently, all possible age-dependencies must be exhaustively considered in the models. This includes the different population volumes at each age group (i.e. the demographic pyramids); the intrinsic variations of dynamical parameters depending on age; the distinct social behaviors, and connectivity profiles of the individuals as a function of age; the distribution of disease burden among the different age groups, and, finally, the addressing of the age dependence of vaccine efficacy, both for the case of BCG –that constitutes the background– and for the case of the novel vaccines which we want to test.

The new modeling ingredients introduced in chapters 9 and 10 contribute to provide a more consistent description of the mentioned age-dependencies, and, as we thoroughly discuss, they exert relevant influences on the impact forecasts at a quantitative level. Noteworthy, age-dependent contact patterns heterogeneities cause strong variations in these impact estimations, which makes evident the need of measuring them as reliably as possible. Contact matrices used in this work come from transnational surveys performed in a series of european countries [12], and, in this sense, the results presented in this Thesis should be considered an approximation, and these could vary if contact patterns of some of the regions differ from those of Polymod project.

Our results confirm the common hypothesis of that adolescents may constitute an optimal age-group to immunize in order to get faster, greater impacts up to 2050, for several reasons. On the one hand, adolescents and young adults constitute both the age groups affected by heavier levels of disease burden and the first segments for which infectious forms of pulmonary TB constitute the dominant form to disease. In addition, current BCG vaccine –even if subject to high uncertainty and variability– seems to present a rapid efficacy waning as time from vaccination passes by [214, 203, 215]. Given that the efficacy of novel vaccines is measured on the basis of a comparison to BCG, the protection by the old vaccine is probably easier to overcome in adolescents than in infants, if the waning of BCG vaccine is accepted as a general rule, something that is not totally clear though [197, 198, 199, 194, 195]. Finally, when heterogeneous contact patterns are introduced, the assortativity of these patterns makes more effective the immunization of these age-groups when compared to vaccination campaign focused on children, as disease transmission is being interrupted in the age groups that are, at the same time, responsible for a higher number of cases and contagions. Despite all these advantages of adolescent focused vaccination campaigns, these are, by definition, unable to protect children beyond BCG does, leaving a permanent reservoir of population at hand for the bacillus, something that could compromise the final goal of disease eradication.

Because of that reason, the complementation of adolescent focused vaccination campaigns with immunization of newborns should be an strategy to consider. Besides being

the only conceptual way to guarantee the eventual protection of all the population (if protection persistence is achieved), newborn immunization is, in certain aspects, *safer* than vaccinating adolescents. The reason, as we discuss in chapter 11, is the existence of prior sensitization to environmental mycobacteria that could compromise proper vaccine performance if it is applied on grown individuals. As we show in this Thesis, in case a novel vaccine applied on adolescents suffers of comparable levels of blocking to those estimated for BCG, the advantages related to directly immunizing adolescents could dissappear.

Furthermore, the discussion regarding the age of target populations in an optimized immunization campaign is tightly related to the type of vaccines to consider, as, as we have said in chapter 2, prime vaccines are initially intended to be applied on newborns substituting BCG while boosters are designed to be used on BCG individuals, remediating its immunity levels. In what regards BCG boosters, the negative results of PhaseIIb trials of MVA85A on infants advise in favor to change trial design for future vaccines to test them on adolescents, which hopefully will result in better protection and impact in case the effects of prior exposition to mycobacterial (or other) antigens are modest. Indeed, the interference of NTM on vaccine's performance are more unlikely for boosters because their biological mechanisms of dissemination throughout the body and their immunogenic mechanisms are different from those of a live vaccine like BCG, whose proper replication is thought to be blocked by the immune system previously sensitized. Taken together, these reasons advise for testing booster vaccines on adolescents as the safest strategy to follow in the context of vaccine development, although, for the case of boosters based on viral vectors, presence of viral antibodies in the individuals being vaccinated could also block vaccine assimilation in this case [139].

The situation is more problematic for live attenuated vaccines, for which vaccine's blocking is arguably more probable. As we have seen in chapter 11, according the analysis of the clinical trials recently conducted in Brazil by Barreto et al. [195], BCG is subject to strong levels of blocking when applied on school age individuals. If a novel live vaccine shares a similar behavior, its application on adolescents might result into residual observed efficacies, at least in low latitudes (where are, in turn, some of the countries with highest TB burden in the world). This is a likely scenario, not just because we talk about live vaccines like BCG, but also because, if a novel live vaccine is intended to be applied on adolescents, these will be previously immunized with BCG, which in this case would act as a source of prior sensitization contributing to vaccine blocking. Admittedly, the alternative strategy in this case is neither a guarantee of success, as BCG, as we know, is harder to overcome in infants. The most important property of a new TB vaccine, if it is intended to substitute BCG on infants, is a longer persistence than the old vaccine, rather than a better initial performance, and, for this reason, even a modest result for a novel vaccine could be interpreted positively in that phase. This complex situation would advise, at least for the case of live vaccines, to consider stratified trials conducted simultaneously in different age groups, as in ref. [195], as well as to foresee, from the beginning of the development process, further extensions of follow up periods from which protection waning patterns

could be estimated as fast as possible.

This is unarguably an exigent *wish list*, taking into account the high costs associated to the design of such clinical trials, and the difficulty to conduct them on high burden settings. Nevertheless, live vaccines constitute a conceptual complementary approach to boosters that shouldn't be disregarded, as they might become the only alternative to these in case the booster vaccines currently on the pipeline are not able to improve the results of MVA85A on infants [213]. Last, but not least, we mustn't forget that boosters could enhance the protection induced by any novel live vaccine the same way they are intended to enhance BCG.

In conclusion, the comparison between newborn and adolescents focused vaccination campaigns faced in this Thesis is subject to the precise compromise between the temporal patterns of protection waning (both of the new vaccines and of BCG, whose behavior is not fully understood yet), that advise for immunizing adolescents instead of infants, and the existence of relevant levels of sensitization of individuals to environmental mycobacterial agents that could be responsible of vaccine blocking, which may act in the opposite direction. However, it is not clear how these effects could affect to the different types of vaccines currently under development, and so, as a relevant part of the development process, the need of clinical trials specifically designed to estimate the impact of these phenomenologies on vaccine performance should be considered.

All that being said, beyond defining the age profile of target populations associated to each possible vaccine candidate, the pipeline for the development of new TB vaccines is committed to consider many other criteria to guarantee an adequate identification, among all the candidates, of those showing more promising protective features, compatible with safety concerns. What makes more insidious the task, -i.e. the seek of a safe candidate able to offer the greatest impact possible– is that the more time is devoted to the task, the worse will be the result (see figure 12.2).

The implementation of novel vaccines and other epidemiological measures to reduce tuberculosis burden at a global level is an urgent matter, and each year of delay is paid with hundreds of thousands of lives.

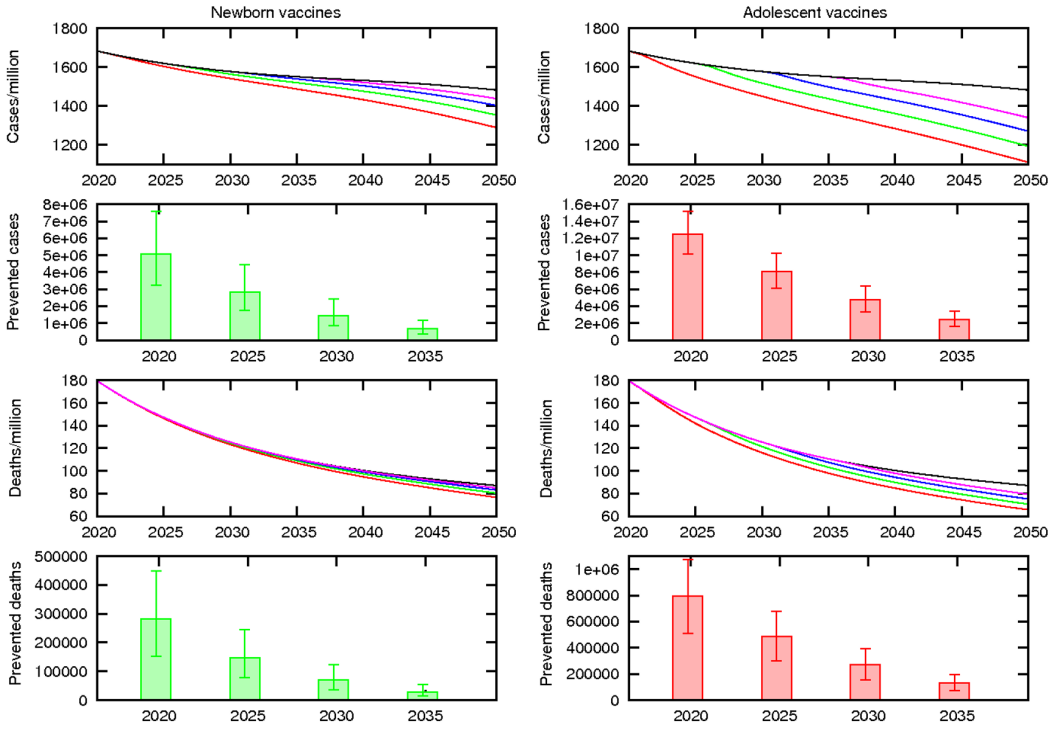


FIGURE 12.1: Effects of the delay in the time of application of a 70% efficacy vaccine. Left column: newborn focused vaccine. Right: adolescents focused vaccine. (a,b): Incidence rates after the application of vaccines in different moments from 2020 to 2035. (c,d): number of TB cases prevented, until 2050, as a function of the year of vaccine implementation. (e,f): Mortality rates after the application of vaccines in different moments from 2020 to 2035. (g,h) number of TB deaths prevented until 2050 by vaccines applied in different years.

12.3 Prospects

In the precedent lines, we have summarized the principal conclusions reached after the research completed around this Thesis, focused on different levels of description of TB infection process; from bacterial cells to human populations. As we have seen throughout this Thesis, the development of novel epidemiological tools against TB constitute an urgent need to reduce the affection caused by the disease worldwide, and, for that development to be possible, concurrent efforts must be undertaken from genetics and cell Biology to immunology and epidemiology; which is particularly true for the case of new preventive vaccines against TB. In this sense, the Thesis here presented, beyond the specific results summarized in the preceding sections, supposes a suitable springboard for the development of a deeper research program aimed at assisting the development of novel vaccines from the perspective of complex systems sciences.

In this sense, as we have discussed before, one of the most limiting caveats for TB vaccine's development is the lack of reliable immunological correlates of vaccine's in-

duced protection against disease that could be used to test vaccines before performing large, expensive and long-lasting clinical trials. As we say, most used readouts for addressing vaccine's performance are *MTB* reduced intracellular replication or *IFN*γ production by T-cells; which have been shown to constitute defective correlates of vaccine's protection [212, 210, 186]. Noteworthy, transcriptional analysis constitutes one of the most feasible approaches to the problem. In a recent work, Berry et al. found a transcriptional footprint for active disease in TB patients [438], which constitutes a promising result that suggests that transcriptional analysis and modeling may constitute a fruitful way to discriminate between different host's status relative to TB infection. In order to explore this possibility, dynamical modeling of gene response to phagocytosis –both of host and pathogen– constitute a fundamental task to undertake.

Admittedly, the TRN of *MTB* compiled in chapter 3 of this Thesis, –despite constituting a valuable resource by itself, as we have shown in this Thesis–, may not constitute a suitable basis for these modeling tasks. As we have already noted, the data compilation underneath our TRN includes interactions reported under very disparate conditions and through disparate experimental methods, which yields the appearance in the network of links that are only present under specific environmental conditions (particularly, links susceptible to be absent in the contexts mentioned before), links reflecting indirect dynamical effects and even links with no dynamical consequences at all. Similarly, even if the interactions in our network were reliable enough, the task of inferring the way multiple inputs couple on a single target is hard to achieve from such heterogeneous dataset. That would make any genome-wide dynamical model performed on our network very uncertain, and thus, the seek of transcriptional correlates of immunological states respect to TB infection will require the characterization of specific regulatory networks, both of host and pathogen, based on specific datasets compiled with that purpose.

A promising approach to achieve that goal is based on expression-quantitative trait loci (eQTL) mapping methods, which have already been shown useful to identify genes associated to susceptibility to TB [439]. EQTL are defined as single nucleotide polymorphisms (SNPs) significantly associated to variation in gene expression levels under a given experimental setup, and their characterization relies on the simultaneous analysis of genotypes and gene-expressions of extended series of different individuals, through multi-variate Bayesian methods [440, 441]. Furthermore, the use of multi-variate data from eQTL mapping to infer specific, highly reliable gene regulatory networks has turned into a hot topic in the field [442, 443, 444, 445]. The combined use of eQTL mapping and transcriptional modeling holds the promise of allowing us to go beyond the simple characterization of genetic factors correlated to disease susceptibility by addressing the dynamical pathways whereby the phenotypes related to TB disease are regulated.

Characterization of gene diversity in the response to *MTB* infection has further implications that hit directly the epidemiological part of the puzzle. Particularly, it has been demonstrated that the global geographical distribution of human ethnic profiles and the diverse lineages of *MTBC* are strikingly correlated [175], which has been related to the specific adaptation of the mentioned pathogenic clades to the ethnic

groups present in different geographical areas [174]. The relevance of these facts in the present time goes far beyond simple theoretical motivations. Instead, the dangers associated to the introduction of vaccines conferring heterogeneous levels of protection for different strains of a disease can not be overemphasized [177], as it could introduce differential evolutive pressures among lineages that could have undesired consequences, as it has occurred already in History, with other pathogens like *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Neisseria meningitides* and *Bordetella pertussis* [446]. For example, a vaccine offering less protection against infection by asian lineages than against others may have dramatic consequences, as it would cause an increase in the incidence of drug resistant tuberculosis associated to the proliferation of Beijing strains; which belongs to asian lineages and are known to be more prune to develop antibiotic resistances.

For these aspects to be properly modeled and understood, a suitable meta-populations model for TB spreading should be implemented; on top on which heterogeneous fitness landscapes between *MTB* strains and human ethnic groups could be explored and related to geographical location. In such a model mobility patterns are intended to be driven by migratory fluxes over extended time periods, and other phenomenologies, like antibiotic resistances and HIV interactions can and must be incorporated. Vaccine impacts evaluations should be faced within this context, and the TB spreading model developed in chapter 9 constitute a first step towards that goal.

Admittedly, the implementation of such modeling platform will necessarily depend on data availability, which, especially in what regards the effects of genetic diversity on TB spreading dynamics is still very limited. For this reason, once again, a deeper knowledge of the genetic mechanisms yielding different levels of susceptibility to disease, as well as how these are distributed among the populations of different countries, are needed to provide reliable descriptions of multi-strain spreading dynamics for TB.

Once again, tuberculosis reveals its complexity at multiple scales, each of which unavoidably attached to the rest. Once again, the advances made so far appear before us as conjoined pieces of a greater puzzle, which are in turn a further *source of error, hesitation, dismay, and expectation* to face in years of future work.

Chapter 13

Conclusiones y perspectivas

La tuberculosis es una vieja compañera del género humano. Estaba con nosotros antes de que fuésemos capaces de escribir y permanece a nuestro lado hoy, en plena era digital [92]. El descubrimiento de su origen microbiológico, por Robert Koch, en 1882, consti-

tuyó uno de los más relevantes hitos de la microbiología médica, una joven disciplina cuyo revolucionario paradigma estaba llamado a proporcionar un impacto excepcional en la Historia de la Salud Pública, desde los días de Koch y Pasteur hasta el final del siglo XX. No obstante, es precisamente lo dilatado de nuestra relación con el patógeno la causa de una gran cantidad de malentendidos y falsas interpretaciones acerca de la enfermedad que dificultan la tarea clínica e investigadora. Tal como escribió Susan Sontag [431], cuando las enfermedades infecciosas alcanzan proporciones epidémicas antes de que sus verdaderas causas sean comprendidas en profundidad, los pueblos golpeados por ellas elaboran explicaciones sobre sus causas y las dotan de complejos significados metafóricos a través de narrativas construidas colectivamente ajenas a cualquier método científico. Comúnmente, esto incluye elementos sobrenaturales y una cierta componente de juicio asociado a conductas punibles asociadas con la enfermedad que contribuyen a la creación de estigmas. Se trata de un proceso que ha ocurrido innumerables veces en la Historia; desde la sífilis hasta la tuberculosis y el VIH.

Más allá de la estigmatización del enfermo, la construcción colectiva de estas *ideas de enfermedad* arrastra un impacto negativo sobre la percepción social del riesgo asociado a la propia patología, que pasa de constituir una amenaza de alcance desconocido a un elemento familiar, conocido, al que la comunidad es capaz de acostumbrarse.

En este sentido, la antigua naturaleza de la tuberculosis parecería invitarnos a creer en el aparente declive de la enfermedad, y a pensar que el corpus de conocimiento médico y epidemiológico actual sobre esta vieja enfermedad debería estar completo ya. Sin embargo, tal como hemos visto a lo largo de esta Tesis, la tisis todavía constituye una de las mayores causas de morbilidad y muerte a nivel mundial, y su complejidad, en cualquiera de las escalas involucradas en el proceso de infección, impone severas limitaciones que dificultan no sólo nuestro conocimiento sobre la patología sino nuestra capacidad de combatirla. En este sentido, por ejemplo, nuestro incompleto conocimiento de los mecanismos de respuesta inmune frente a la infección tuberculosa dificulta la búsqueda de indicadores inmunológicos de protección ante la infección; lo cual, a su vez, lastra enormemente el desarrollo y evaluación de nuevas vacunas. Del mismo modo, la propia complejidad del ciclo de vida del bacilo dificulta la vigilancia epidemiológica, especialmente en los países más azotados por la enfermedad –que son precisamente aquellos con sistemas de salud pública más frágiles y vulnerables [432]–, haciendo muy complicada cualquier evaluación de la afección de la TB a nivel poblacional. En definitiva, todos estos problemas evidencian que todavía queda un largo camino por recorrer en investigación sobre TB a todos los niveles, desde la Bioquímica y Biología celular hasta la epidemiología; y lo que es más importante, que los avances que se produzcan en cada uno de dichos campos influirán potentemente en los demás. Estas observaciones, junto a las dramáticas implicaciones sociales que la TB arrastra tras de sí deberían ser suficientes para motivar la idea –ya desarrollada por Young, Stark and Kirschner [2], entre otros [433]– de que el desarrollo de nuevos enfoques multidisciplinares e integrados, abarcando el estudio de la enfermedad a todos los niveles de forma coordinada, son necesarios si pretendemos alcanzar un conocimiento más profundo de este viejo compañero del hombre, y, lo que es más importante, si pretendemos

maximizar nuestras oportunidades de erradicarlo. Este es el espíritu en el que esta Tesis encuentra su contexto.

En los capítulos 3 a 11 de este texto hemos presentado los resultados de la investigación llevada a cabo en esta Tesis; en las líneas siguientes repasaremos las principales conclusiones alcanzadas a partir de estos estudios, las líneas de investigación que estos abren y cómo éstas están llamadas a entrelazase alrededor de enfoques integrados para la resolución de ciertos problemas abiertos en el campo.

13.1 *Interactomas* celulares en *MTB*

En los capítulos 3 a 6 e esta Tesis nos hemos centrado en la caracterización de la red de regulación transcripcional de *MTB* y el estudio de su red de interacciones entre proteínas. Además, hemos estudiado los efectos de la incompletitud y falta de fiabilidad en los datos que constituyen las redes de transcripción a un nivel más general. Hemos estudiado cómo estos problemas afectan determinados análisis topológicos de las redes y hemos propuesto métodos para la identificación de errores en este tipo de redes y otros sistemas descritos como redes dirigidas.

La revisión bibliográfica presentada en el capítulo 3 constituyó, en el momento de la publicación de dicho trabajo, la compilación más completa sobre regulación transcripcional en *MTB*, conteniendo tres veces más interacciones que cualquier *dataset* previo [89]. Esto nos permitió llevar a cabo un análisis topológico detallado del sistema que condujo a la identificación de los genes más conectados en la red y la medición de las propiedades globales del sistema, que resultaron similares a las de redes análogas pertenecientes a otros organismos.

Notablemente, y pese a que la heterogénea naturaleza de la información compilada, que incluye datos de muy diferente naturaleza y significado, impone importantes limitaciones, la simple integración de toda esta información alrededor de un único *dataset* ha contribuido a arrojar luz acerca de los roles jugados por los distintos genes en la red. En esta linea, integrar los datos y analizar el sistema como una red ha servido de ayuda para contextualizar la importancia de cada regulador y, algo más difícil de conseguir de otro modo, identificar genes cuya expresión queda afectada por un número mayor de reguladores. En este sentido, aunque no resulta sorprendente que el regulador *PhoP* aparezca como uno de los nodos más conectados del sistema, resulta interesante comprobar que dos de los genes ecibiendo un número mayor de regulaciones –*fadD26* y *icl*–, hayan sido identificados como factores claves en el control de la virulencia del patógeno y su capacidad para persistir *in vivo*, tal cómo comentáramos en el capítulo 2 (ver figura 2.2). La modificación de la expresión de estos genes, de hecho, se encuentra en la base del fenotipo atenuado de la vacuna *MTB*VAC [129, 88, 186], *fadD26*, (como *PhoP*) se han eliminado, e *icl* aparece sobre-expresado a consecuencia de la ausencia de *PhoP*.

En el contexto del análisis topológico de la red de transcripción de *MTB*, el estudio estadístico de los motivos de red conllevó algunos resultados de interés. Por un lado, la red transcripcional de *MTB*, tras un primer análisis, parecía constituir una excepción a

la clasificación propuesta por Alon y colaboradores en [267]. En el mencionado trabajo, los autores identifican cuatro grupos o superfamilias de redes en función de la presencia de los diferentes motivos (ver figura 1.3) en las mismas, formadas por redes de similar naturaleza. De entre los cuatro grupos, dos estarían formados por sistemas biológicos: un primer grupo formado por redes de regulación transcripcional de organismos unicelulares (en el cual *MTB debería* haberse integrado) y otro formado por redes de regulación de organismos pluricelulares y redes sinápticas; al cual, sorprendentemente, la red de transcripción de *MTB* parecía pertenecer. Según la intrepretación de esta división propuesta en [267], la primera familia estaría compuesta por redes "superficiales", de poca profundidad, donde el tiempo de respuesta del sistema completo es del orden del de una simple interacción. Por el contrario, el segundo grupo lo formarían redes más profundas para las cuales el tiempo de respuesta del sistema puede ser mucho mayor que el de una sola interacción.

A la luz de esta interpretación, podría resultar tentador atribuir el hecho de que los motivos de la red de transcripción de *MTB* presenten un perfil más compatible con el segundo grupo de redes a razones evolutivas, relacionadas con su habilidad para sobrevivir dentro del huésped durante largos periodos de tiempo, por ejemplo. Sin embargo, tal como demostramos en el capítulo 5, la propia división de los sistemas biológicos estudiados en dos familias diferentes constituye, muy probablemente, un error de origen metodológico. En el capítulo 5, analizamos las versiones actualizadas de algunas redes de transcripción de organismos que fueron ya estudiados en [267] como miembros de la primera familia: particularmente *B.subtilis* y *E.coli*. Curiosamente, tras actualizar los *datasets* 7 años después del análisis original, los motivos de red de estos sistemas pasan, como *MTB*, a presentar perfiles compatibles con el segundo grupo. La razón de esto es metodológica, y relacionada con la falta de enlaces bi-direccionales en los datasets originales y la aparición de algunos de ellos tras el proceso de actualización, del mismo modo que la red transcripcional de *MTB* presenta varios de estos *feedback loops*. En el modelo propuesto en [267] para el análisis de motifs, las redes aleatorias del *null ensemble* conservan el número de enlaces bi-direccionales del sistema original, lo cual hace que, cuando se analiza una red sin tales elementos, el Z-score asociado a motivos que contienen enlaces bi-direccionales no pueda ser definido. A causa de esta situación, los perfiles de los motivos de red pertenecientes a las redes del primer grupo en sus versiones antiguas –esencialmente carentes de enlaces bi-direccionales– se diferenciaban de los del segundo grupo; una situación que se revierte al añadir siquiera una reducida cantidad de enlaces bi-direccionales a consecuencia de la actualización de las redes.

De cualquier modo, el estudio realizado no pretende ofrecer una respuesta a la cuestión de porqué diferentes redes complejas presentan distintos tipos de motivos de red en sus estructuras. Del mismo modo, no compromete la interpretación general propuesta por Alon y colaboradores de acuerdo con la cual este fenómeno sería una consecuencia de la presencia de presiones evolutiva favoreciendo la aparición de unos motivos e impidiendo la de otros en función de sus características dinámicas. por ejemplo, los llamados *feed-forward loops* son estructuras capaces de actuar, dependiendo de los signos de los enlaces que los componen y el modo en que éstos se acoplan, como

discriminadores de ruido [261, 268] o como módulos aceleradores de respuesta [269], lo cual los hace particularmente útiles en redes de regulación genética de organismos adaptados a medios variados y difíciles de predecir. Con el objetivo de evaluar indirectamente esta hipótesis, hemos analizado las redes transcripcionales de *E.coli* y *MTB* en busca de motivos diferencialmente expresados en ambos sistemas (i.e. motivos que aparecen más veces de lo esperado en una de las redes, y menos en la otra). Los resultados de este análisis indican que *E.coli* presenta una tendencia mayor a presentar estructuras relacionadas con el *feed-forward loop*, lo cual es coherente con el hecho de que la bacteria es capaz de habitar en muy diversos medios, de modo mucho más predecible que *MTB*, cuya estrategia evolutiva ha consistido en adaptarse muy específicamente a un estilo de vida parasítico. Estas observaciones resultan coherentes con la hipótesis de que el contenido en motivos de una red específica obedece a presiones evolutivas, aunque puedan interpretarse, como mucho, como un indicio parcial de su existencia. La comparación de redes pertenecientes a organismos similares resulta una metodología plausible para verificarla, si, por ejemplo, un mayor número de párasitos intracelulares como *MTB* se comparan frente a un mayor número de organismos de vida libre como *E.coli*, y los resultados resultan robustos.

No obstante, conviene anotar que la hipótesis evolutiva para explicar los contenidos en motivos en redes dirigidas ha sido fuertemente cuestionada en la literatura. Existen, de hecho, relevantes interpretaciones alternativas de acuerdo a las cuales una red puede presentar un contenido en motivos no trivial sin necesidad de ningún efecto evolutivo. Tal es el caso de redes embebidas en espacios geométricos [339], o de redes conteniendo errores sistemáticos [434]. En el capítulo 5 hemos puesto a prueba esta última posibilidad en redes de transcripción, estudiando posibles correlaciones entre el nivel de fiabilidad con el que un enlace es caracterizado y su entorno topológico local. La conclusión de dicho análisis en que la caracterización defectuosa de algunos enlaces es capaz de aumentar el valor absoluto del Z-score asociado a diferentes motivos, dejando aproximadamente invariante la significación estadística relativa de unos frente a otros.

Por último, en el capítulo 5, también estudiamos la meso-escala de la red transcripcional de *MTB*, y su relación con la caracterización funcional de los distintos genes que la componen a diferentes niveles de resolución. En este punto hemos observado que el análisis de modularidad de la red transcripcional muestra una mayor correspondencia con la partición funcional cuando toda la información correspondiente a la dirección y signo de las interacciones es tenida en cuenta por los algoritmos. Este resultado, una vez más, resalta la influencia que la calidad y completitud de los datos ejerce sobre el resultado del análisis topológico de redes complejas.

Considerando todo lo anterior, en el capítulo 6 nos centramos en el estudio de distintos métodos para la estimación de la fiabilidad de los enlaces de una red, con el ánimo de identificar enlaces falsos y perdidos. Específicamente, a partir de un trabajo previo de R. Guimerá y M. Sales-Pardo [276], en el cual se propone un método útil a tal efecto en redes no-dirigidas, desarrollamos un método análogo para efectuar la misma labor en sistemas en los que, como las redes de transcripción, los enlaces tienen dirección. Nuestro método, comparado con otros algoritmos disponibles en la literatura,

ofrece un rendimiento ligeramente superior en la identificación de errores aleatorios; una mejora de rendimiento que podría ser insuficiente para justificar los elevados costes computacionales del método propuesto. La situación es probablemente distinta cuando se trata de identificar enlaces perdidos en un *dataset* real, donde nuestro método se muestra superior a cualquier otra alternativa.

Además de todo el trabajo anteriormente mencionado, en esta primera mitad de la tesis dedicada al análisis de redes biológicas hemos también estudiado, en el capítulo 4, la red de interacciones entre proteínas de *MTB*, y, más específicamente, como esta cambia a resultas de los cambios en los perfiles de expresión genética con los que la bacteria responde a diferentes estímulos externos. Por un lado, hemos estudiado, a través de un análisis filogenético, de qué modo las principales características ambientales asociadas a diferentes *setups* experimentales se reflejan en la topología de la red de proteínas del bacilo. Este tipo de análisis podría ayudar a discriminar, entre diferentes modelos *in vitro*, aquellos asociados a fenotipos más compatibles con un cierto estado de referencia, ayudando en la mejora de los modelos de infección, un aspecto fundamental en TB. Por otro lado, en este capítulo usamos la red de interacciones entre proteínas del bacilo para estudiar su respuesta ante diferentes stresses fundamentales en el ambiente fagosómico: hipoxia, escasez de nutrientes, estrés oxidativo genérico, exposición a NO, escasez de iones como Fe^{++} y daño celular. Nuestro análisis evidencia una fuerte coherencia en la red PPIN ante todos estos estímulos generadores de estrés, lo cual sugiere la existencia de mecanismos evolutivos convergentes a partir de los cuales la bacteria responde de manera muy similar a diferentes estímulos que aparecen altamente correlacionados *in vivo*. Este resultado de puede interpretar en relación a una serie de trabajos en los cuales este tipo de mecanismos de regulación convergentes son estudiados [227, 133].

Notablemente, este análisis nos ha permitido identificar una serie de proteínas que ejercen el papel de *hubs* en respuesta a todos los estreses mencionados. Estas proteínas constituyen seis pares de antígenos análogos a la pareja ESAT6-cfp10, cuyo rol central y persistente en la red de interacciones entre proteínas en respuesta a estrés sugiere una implicación biológica robusta. Más allá de que nuestro método destaque la propia pareja de proteínas ESAT6-cfp10 entre dichos *hubs* globales, resulta llamativo que sus análogos presenten un perfil similar en la red, puesto que, hasta ahora, es mucho menos lo que se sabe sobre ellos, más allá de la hipótesis de que constituyen una fuente de variedad antigénica que podría proporcionar al patógeno medios para mantener bajo control la respuesta inmune del huésped [435, 436]. Nuestro análisis sugiere que las proteínas identificadas en nuestro análisis podrían ejercer funciones biológicas tan relevantes como el propio par ESAT6-cfp10, una hipótesis que sin duda merece ser explorada en el futuro.

No obstante, diversos aspectos de este trabajo requieren de verificación adicional. Por una parte, la interpretación de las redes de respuesta a estrés en *MTB* debería someterse a comparación con otros organismos. El objetivo sería elucidar si la convergencia observada en los perfiles topológicos de las redes asociadas con los distintos estreses son una consecuencia de la adaptación del bacilo a un estilo de vida parasítico o no. Por otra parte, en términos generales, nuestro método implica que los nodos con mayor *strength* son aquellos que presentan una alta conectividad y, a la vez, tanto ellos

como su entorno se encuentran sobre-expresados. Una parte relevante, sin embargo, consiste en aislar el efecto de cada uno de los dos factores (i.e. topología y expresión) sobre los resultados obtenidos. Estas dos cuestiones, entre otras, están siendo investigadas actualmente.

13.2 Modelos epidemiológicos

En la segunda parte de la Tesis, en los capítulos 7 a 11, nos centramos en el desarrollo de modelos epidemiológicos para describir la tuberculosis y otras enfermedades que comparten con ella características como prolongados periodos de latencia e intensas interacciones con otras infecciones.

En este sentido, en los capítulos 7 y 8, pretendemos extender el paradigma de la epidemiología matemática en redes complejas hacia escenarios epidemiológicos estrechamente relacionados con la tuberculosis. La complejidad característica de las redes de contactos sobre las cuales se propagan las enfermedades infecciosas y sus efectos sobre los fenómenos epidémicos constituyen una materia recurrentemente tratada en la literatura científica. Sin embargo, el puzzle difícilmente puede considerarse resuelto cuando ciertos factores clave no han sido todavía explorados, como la interacción entre topologías complejas y la persistencia de la enfermedad, o los efectos de la coexistencia de dos (o más) enfermedades relacionadas transmitiéndose a través de redes complejas.

En el capítulo 7, presentamos un modelo de campo medio para describir la dinámica de una enfermedad que, al igual que la TB, presenta simultáneamente *primo-infección* y periodos de latencia prolongados al transmitirse tanto en poblaciones homogéneas como heterogéneas. Con ello logramos derivar, numérica y analíticamente, la expresión para los umbrales epidémicos, señalando la influencia de los parámetros relacionados con la dinámica de la enfermedad así como los relacionados con la demografía. De esta manera podemos corroborar, dentro del ámbito de aplicación de nuestro HMF, que el umbral epidémico depende del cociente $\langle k\rangle/\langle k^2\rangle$. En el contexto de este trabajo, hemos identificado fenomenologías específicas que surgen de la interacción entre las redes complejas y la persistencia, como por ejemplo el modo en que la distribución de grado del sistema puede ser alterada por efecto de la enfermedad en función de la forma en que los neonatos se incorporan al sistema.

Asimismo, en el capítulo 8, presentamos un modelo matemático para la descripción de la propagación simultánea de dos enfermedades propagándose sobre una misma población, a través de dos diferentes redes complejas de contactos, capaces de interaccionar entre sí. En nuestro modelo, prestamos especial atención a la descripción de los distintos mecanismos de interacción que se pueden observar entre dos enfermedades, como variaciones en la probabilidad de propagación, tasa de recuperación o susceptibilidad a la enfermedad. En este sentido, considerando exhaustivamente todas las posibles vías de interacción, somos capaces de identificar las condiciones globales necesarias para que se produzcan interacciones positivas o negativas - favorecimiento u obstaculización de la enfermedad, en términos del desplazamiento de los umbrales epidémicos-, previendo, por vez primera, la posibilidad de que una enfermedad pueda

facilitar o dificultar la transmisión de otra en función de su nivel de prevalencia. Nuestra descripción, válida simultáneamente para los modelos SIS y SIR, muestra además que los umbrales epidémicos difieren para ambos modelos, no sólo como consecuencia de correlaciones dinámicas (no consideradas en nuestro campo medio heterogéneo) sino por efecto de la interacción entre las enfermedades. Adicionalmente, nuestro modelo ofrece una descripción de los umbrales epidémicos tanto dentro como fuera de equilibrio. Esto, tal y como se muestra en la Figura refinteracting-lambda-2-c,supone ser capaces de describir cómo el umbral epidémico de una enfermedad cambia mientras evoluciona el brote epidémico de una infección conjugada, lo cual evidencia la importancia del factor temporal en este tipo de problemas. Finalmente, hemos mostrado cómo nuestro enfoque, a pesar de su simplicidad, es capaz de reproducir, al menos cuantitativamente, datos reales de prevalencia correspondientes a un sistema bien conocido de enfermedades interactuantes: el sistema sindémico formado por el VIH y la TB, utilizando datos procedentes de la República Sudafricana, sin duda uno de los lugares más afectados por la concurrencia de las dos enfermedades. Llevando a cabo un procedimiento simple de ajuste, nuestro modelo puede proporcionar una fiel reproducción de la serie temporal de la prevalencia de la TB y del VIH, desde el año 1900 hasta el 2000, periodo durante el cual la interacción entre las mismas ha constituido el principal factor responsable de incremento de las tasas de prevalencia de TB. Haciendo este ejercicio, más allá de mostrar que nuestro modelo es capaz de reproducir datos reales (lo cual, como hemos comentado, difícilmente puede considerarse sorprendente), podemos observar que, aunque nuestro modelo supone una simplificación significativa de historias reales tanto de la TB como del VIH, los parámetros obtenidos en el ajuste reflejan de forma consistente los mecanismos de interacción por los que el VIH modifica la dinámica de propagación de la TB: tanto el incremento de la susceptibilidad a la TB de los individuos afectados por VIH [157] como las reducidas tasas de transmisión de TB de los pacientes co-infectados con VIH [146]. Esto indica que el ajuste colectivo en modelos epidémicos de enfermedades interactuantes –un enfoque que ha mostrado su utilidad en otros campos [437]- puede ayudarnos a discriminar entre los mecanismos reales de interacción entre ellas.

Una vez explorada la relación entre la complejidad topológica de las redes de contactos y diferentes aspectos dinámicos de la transmisión epidémica (como la persistencia o la interacción entre enfermedades), nos centramos en la modelización de la propagación de la TB en los capítulos 9 a 11. En esta parte de la Tesis analizamos el estado del arte en modelización matemática de TB e incorporamos nuevos ingredientes basados en lo aprendido en capítulos anteriores y, en general, de lo que está siendo utilizado en otros campos de la epidemiología computacional. El objetivo es examinar, a nivel cuantitativo, la influencia de estas nuevas hipótesis sobre los resultados de los modelos, así como la posibilidad de incorporarlos a los mismos. El objetivo último es el de incrementar la fiabilidad de las plataformas de modelado actuales para el caso de la TB, y aplicar dichas mejoras en la evaluación del impacto de nuevas intervenciones epidémicas, muy especialmente vacunas preventivas.

Así, en el capítulo 9, desarrollamos un modelo cuantitativo para la descripción de la propagación de la TB, basado en una serie de trabajos de C. Dye y colaboradores. Nue-

stro modelo procesa dos fuentes diferentes de datos. Por un lado, utiliza proyecciones demográficas sobre distribuciones de edad de la población facilitadas por la división de población de Naciones Unidas [23]. Adicionalmente, las tasas de incidencia y mortalidad se extraen de la base de datos de tuberculosis de la Organización Mundial de la Salud [24]. Las principales novedades conceptuales de nuestro modelo son cuatro: 1) la utilización de patrones de contactos heterogéneos para simular la transmisión de la enfermedad, 2) la consideración explícita del acoplamiento entre la dinámica de la enfermedad y la evolución demográfica, 3) el abandono de la hipótesis de equilibrio inicial y 4) la introducción de un procedimiento de ajuste distribuido por edades (en el capítulo 10) capaz de reproducir niveles de incidencia por separado en cada grupo de edad.

Nuestras motivaciones para introducir estos ingredientes en los ya de por sí sofisticados modelos de propagación de TB tienen que ver con el contexto científico actual en este campo. Tal y como hemos comentado, la capacidad de ofrecer evaluaciones de impacto fiables para las nuevas vacunas de TB actualmente en desarrollo constituye un objetivo lo suficientemente relevante para justificar la necesidad de refinar los modelos actuales y de asegurar que todas las fuentes de datos actualmente disponibles son tenidas en cuenta en el desarrollo de nuevos modelos. En este sentido, como ya se ha señalado en el texto, la edad a la que las nuevas vacunas se dirigen supone una cuestión central que debe atenderse, y en consecuencia, todas las posibles dependencias relacionadas con la edad deben ser consideradas de forma exhaustiva en los modelos. Esto incluye los diferentes volúmenes de población para cada grupo de edad (las pirámides demográficas); las variaciones intrínsecas de parámetros dinámicos de la enfermedad; los diferentes comportamientos sociales y perfiles de conectividad de los individuos; la distribución de la afección de la enfermedad y, finalmente, la caracterización de la efectividad de la vacuna; todos ellos factores fuertemente dependientes de la edad de los individuos.

Los nuevos ingredientes introducidos en los modelos, en los capítulos 9 y 10 contribuyen a proporcionar una descripción más consistente de las mencionadas dependencias con la edad y, como hemos discutido en profundidad, ejercen influencias relevantes sobre las previsiones de impacto a un nivel cuantitativo. Las heterogeneidades de los patrones de de contacto dependientes de la edad causan fuertes variaciones en estas estimaciones de impacto, lo que evidencia la necesidad de medirlas de la forma más fiable posible. Las matrices de contactos utilizadas en este trabajo provienen de sondeos transnacionales realizados en varios países europeos [12] y, en este sentido, los resultados presentados en esta Tesis deberían considerarse una aproximación, que podría variar si los patrones de contacto de algunas de las regiones fueran medidos específicamente y difirieran de aquéllos del proyecto Polymod.

Nuestros resultados confirman la hipótesis común de que los adolescentes podrían constituir un grupo de edad óptimo de inmunización para conseguir mayores y más rápidos impactos hasta el 2050 por varias razones. Por un lado, los adolescentes y los adultos jóvenes son los grupos que sufren los más altos niveles de afección de la enfermedad y los primeros segmentos para los cuales la TB pulmonar es la forma dominante de la enfermedad. Además, la actual vacuna BCG, aún estando sujeta a

una alta incertidumbre y variabilidad, parece presentar una rápida pérdida de eficacia a medida que pasa el tiempo desde el momento de la vacunación [214, 203, 215]. Dado que la eficacia de las nuevas vacunas se mide por comparación con la BCG, la protección ofrecida por la antigua vacuna es probablemente más fácil de superar en adolescentes que en niños, si el decaimiento de la vacuna BCG es aceptado como regla general, algo que por otro lado no está totalmente claro [197, 198, 199, 194, 195]. Finalmente, cuando se introducen patrones de contacto heterogéneos, la asortatividad de estos patrones hace más efectiva la inmunización de estos grupos de edad, al compararla con campañas de vacunación infantiles, ya que la transmisión de la enfermedad se interrumpe más rápido en los grupos de edad que son, al mismo tiempo, responsables de un mayor número de casos sufridos y de contagios causados. A pesar de todas las ventajas de las campañas de vacunación para adolescentes, éstas son por definición incapaces de proteger a los niños más de lo que lo hace la BCG, dejando una reserva permanente de población al alcance del bacilo, algo que podría comprometer el objetivo final de la erradicación de la enfermedad.

Por esta razón, la complementación de las campañas de vacunación destinadas a la población adolescente con la inmunización de neonatos debería ser una estrategia tomada en consideración. Además, siendo la única vía conceptual para garantizar la eventual protección de toda la población (si se consiguiera la persistencia temporal de la eficacia en las nuevas vacunas), la inmunización de los recién nacidos es, en determinados aspectos, más segura que la vacunación de adolescentes. El motivo, como se ha discutido en el capítulo 11, es la existencia de exposición previa a mycobacterias ambientales que podría comprometer el buen funcionamiento de la vacuna al aplicarse en individuos adultos. Tal y como se muestra en esta Tesis, en caso de que una nueva vacuna aplicada en adolescentes sufriera niveles comparables de bloqueo a aquéllos estimados para la BCG, las ventajas relacionadas con la inmunización directa de adolescentes podrían desaparecer.

Además, la discusión sobre la edad de la población a inmunizar en una campaña de vacunación optimizada está estrechamente relacionada con el tipo de vacunas a considerar, puesto que, tal y como se ha comentado en el capítulo 2, las vacunas *prime* están inicialmente orientadas para su aplicación en recién nacidos en lugar de la BCG, mientras que las vacunas *booster* están diseñadas para su uso en individuos mayores, previamente vacunados con BCG, a fin de *restaurar* los niveles de inmunidad de la vieja vacuna. En lo que respecta a las *boosters*, los resultados negativos de los ensayos de fase IIb de MVA85A en niños aconsejan la prueba de futuras vacunas en adolescentes, lo cual, con suerte, resultará en una mejor protección en caso de que los efectos de la exposición previa a mycobacterias sean modestos. De hecho, la interferencia de las mycobacterias ambientales en el funcionamiento de las vacunas es menos probable para las *boosters* ya que sus mecanismos biológicos de diseminación a través del cuerpo y sus mecanismos inmunogénicos son diferentes de los de una vacuna viva como la BCG, cuya adecuada reproducción es bloqueada por el sistema inmunitario a resultas de la exposición previa a antígenos mycobacterianos. Por todas estas razones es recomendable evaluar las vacunas *booster* en adolescentes como la estrategia más segura para continuar con el desarrollo de vacunas, aunque para el caso de *boosters*

basadas en vectores víricos, la presencia de anticuerpos virales en los individuos a vacunar podría también bloquear la asimilación de la vacuna [139]. Esta situación es más problemática para vacunas vivas atenuadas, para las cuales el bloqueo de vacunas es más probable. Tal y como hemos visto en el capítulo 11, de acuerdo con el análisis de los ensayos clínicos recientemente realizados en Brasil por Barreto y colaboradores [195], BCG sufriría intensos niveles de bloqueo cuando se aplica a individuos en edad escolar. Si una nueva vacuna viva comparte un comportamiento similar, su aplicación sobre adolescentes podría resultar en observadas eficacias residuales, al menos en latitudes bajas (donde se encuentran, asimismo, algunos de los países con los mayores niveles de TB en el mundo). Este es un escenario probable, no sólo porque nos referimos a vacunas vivas como la BCG, sino porque además, si una nueva vacuna viva está dirigida a una población adolescente, dicha población se encontrará previamente inmunizada con BCG, lo cual en este caso actuaría como una fuente adicional de exposición antigénica que contribuiría al bloqueo de la vacuna. Hay que admitir que la estrategia alternativa en este caso no es garantía de éxito, ya que la BCG, como sabemos, es más difícil de superar en niños. La propiedad más importante de una nueva vacuna de TB, si pretende sustituir la BCG en niños, es una persistencia más prolongada que la de la antigua vacuna más que un mejor funcionamiento inicial y, por esta razón, incluso un modesto resultado para una vacuna nueva podría ser interpretado positivamente en esta fase. Esta compleja situación haría conveniente, al menos para el caso de vacunas vivas, considerar ensayos estratificados llevados a cabo simultáneamente en diferentes grupos de edad, como en la referencia [195], así como prever, desde el principio del proceso de desarrollo, ulteriores extensiones de los periodos de seguimiento a fin de que los patrones temporales de decaímiento de la protección vacunal puedan ser estimados tan pronto como sea posible. Sin duda alguna, ésta es una exigente lista de deseos, teniendo en cuenta los altos costes asociados al diseño de dichos ensayos clínicos y la dificultad de llevarlos a cabo en escenarios de alta incidencia de la enfermedad. Sin embargo, las vacunas vivas constituyen un en enfoque conceptual complementario a las *boosters* que no debería ser obviado, puesto que podrían convertirse en la única alternativa a estas últimas en caso de que los *booster* actualmente en desarrollo no sean capaces de mejorar los resultados de MVA85A en niños [213]. Por último y no por ello menos importante, debemos recordar que las *boosters* podrían potenciar la protección inducida por cualquier nueva vacuna viva de la misma forma que podrían hacerlo con la BCG. En conclusión, la comparación entre las campañas de vacunación dirigidas a recién nacidos y las orientadas a adolescentes afrontada en esta Tesis está sujeta al compromiso preciso entre los patrones temporales del deterioro de la protección (tanto de las nuevas vacunas como de la BCG, cuyo comportamiento no ha sido aún completamente descifrado), que aconsejan la inmunización de adolescentes en lugar de niños; y la existencia de niveles relevantes de exposición a agentes mycobacterianos en la población, que podrían ser responsables del bloqueo de la vacuna, lo cual tendría el efecto contrario. Sin embrago, no está claro cómo estos aspectos podrían afectar a los distintos tipos de vacuna actualmente en proceso de desarrollo y por ello, como parte relevante del proceso, la necesidad de ensayos clínicos específicamente diseñados para estimar el impacto de estos factores debería tenerse en consideración.

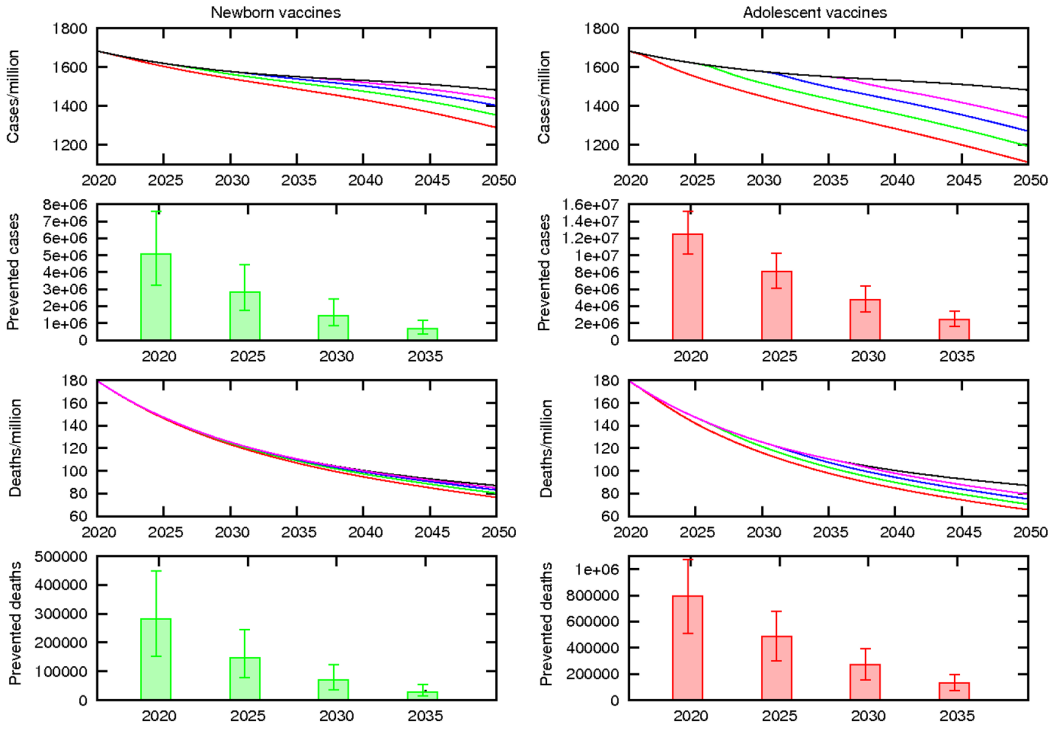


FIGURE 13.1: Efectos sobre el retraso en el tiempo de aplicación de una vacuna de un 70% de eficacia. Columna izquierda: vacuna dirigida a neonatos. Derecha: vacuna dirigida a adolescentes. (a,b): Las tasas de incidencia después de la aplicación de vacunas en momentos diferentes desde 2020 a 2035. (c,d): número de casos evitados hasta 2050 por vacunas introducidas en distintos momentos. (e,f): Tasas de mortalidad después de la aplicación de vacunas en diferentes momentos desde 2020 a 2035. (g,h): Número de muertes evitadas hasta 2050 en función del momento de implementación de la vacuna.

Dicho todo esto, más allá de definir el perfil de edad de la población objetivo asociada a cada posible vacuna candidata, el desarrollo de nuevas vacunas de TB está comprometido a considerar muchos otros criterios para garantizar una adecuada identificación de aquéllas que muestren características óptimas siendo compatibles con criterios de seguridad. Lo que hace más insidiosa la tarea de buscar entre todos los proyectos una vacuna segura capaz de ofercer el mayor impacto es que cuanto más tiempo se dedique a la tarea peor será el resultado, tal como se evidencia en la figura 13.2. La implementación de nuevas vacunas para reducir llos niveles de Tb a escala global es un asunto urgente y el precio por cada año de retraso son cientos de miles de vidas.

13.3 Perspectivas

En los párrafos precedentes hemos resumido las principales conclusiones alcanzadas tras la investigación realizada para esta Tesis, centrada en diferentes niveles de descripción del proceso de infección de la TB; desde las células de la bacteria hasta las poblaciones humanas. Tal y como hemos visto a lo largo de este trabajo, el desarrollo de nuevas herramientas epidemiológicas contra la TB se destaca como una necesidad urgente para reducir la afección causada por la enfermedad en todo el planeta y, para que este desarrollo sea posible, deben aunarse esfuerzos desde la genética y la Biología celular a la inmunología y la epidemiología; lo que es particularmente cierto para el caso del desarrollo de nuevas vacunas preventivas contra la TB. En este sentido, la Tesis aquí presentada, más allá de los resultados específicos resumidos en las secciones anteriores, supone un apropiado trampolín para el desarrollo de un programa de investigación más profundo orientado a la asistencia para el desarrollo de nuevas vacunas desde la perspectiva de la ciencia de sistemas complejos. Tal y como hemos comentado anteriormente, una de las limitaciones principales en el desarrollo de vacunas para la TB es la ausencia de indicadores inmunológicos fiables de protección contra la infección a los que recurrir antes de pasar a realizar prolongados y costosos ensayos clínicos. Las mediciones más usadas hoy en día como indicadores de la eficacia de una vacuna se basan en la reducción de la replicación de la bacteria *in vivo* o en la producción de $IFN\gamma$ por células T; técnicas éstas capaces de ofrecer resultados sólo parcialmente fiables [212, 210, 186].

En este contexto, el análisis transcripcional constituye uno de los enfoques más prometedores para tratar de resolver este problema. En un trabajo reciente, Barry y colaboradores han identificado una huella transcripcional específica capaz de distinguir pacientes con TB activa del resto [438], lo cual supone un espaldarazo a la idea de usar análisis transcriptómico para discriminar entre diferentes estatus del huésped (por ejemplo, inmune, susceptible, infectado o protegido por una vacuna) en relación con la infección de TB. Para explorar esta posibilidad, el modelado dinámico de la respuesta genética tras la fagocitosis –tanto del huésped como del patógeno– es una tarea fundamental a llevar a cabo. No obstante, la red de regulación transcripcional de *MTB* compilada en el capítulo 3, aún constituyendo un valioso recurso por sí mismo, como se ha mostrado en la Tesis-, puede no constituir una base apropiada para estas tareas de modelización. Como ya hemos señalado, nuestra red incluye interacciones identificadas bajo condiciones muy diferentes y a través de muy diversos métodos experimentales, que causan la aparición en el sistema de enlaces presentes sólo bajo condiciones ambientales específicas, enlaces que reflejan efectos dinámicos indirectos e incluso enlaces sin ningún tipo de consecuencia dinámica. Igualmente, incluso con interacciones lo suficientemente robustas en nuestra red, inferir el modo en que múltiples inputs se emparejan en un único objetivo es difícil de conseguir partiendo de datos tan heterogéneos. Así, la implementación de cualquier modelo dinámico sobre nuestra red resultaría particularmente incierta y, por ello, la búsqueda de indicadores transcripcionales de estados inmunológicos en relación a la infección de TB requerirá la caracterización de redes reguladoras específicas, tanto para el portador como para el patógeno, basadas

en bases de datos reunidas a tal propósito. Un enfoque prometedor para conseguir este objetivo se basa en métodos de mapeado de *expression-quantitative trait loci* (eQTL) , un método que ya se ha revelado útil para identificar genes asociados a susceptibilidad a la TB [439]. Los eQTL se definen como *single nucleotid polymorphisms* (SNPs) significativamente asociados a la variación en niveles de expresión genética de un determinado gen bajo unas condiciones experimentales dadas, y su caracterización se apoya en el análisis simultáneo de genotipos y expresiones genéticas de un conjunto de distintos individuos, a través de métodos Bayesianos [440, 441]. En esta línea, el uso de conjuntos de datos multivariados de esta naturaleza para inferir redes reguladoras genéticas específicas y fiables se ha convertido en un tema de actualidad en este ámbito [442, 443, 444, 445]. El uso combinado de mapeado eQLT y modelos transcripcionales podría permitir ir más allá de la simple caracterización de factores genéticos relacionados con la susceptibilidad a la enfermedad identificando las vías dinámicas por las cuales los fenotipos asociados a la TB son regulados. La caracterización de la diversidad genética en la respuesta a la infección de *MTB* tiene mayores implicaciones que van a dar directamente con la parte epidemiológica del puzzle. Particularmente, se ha demostrado que la distribución geográfica global de los perfiles étnicos humanos y los diversos linajes en el *MTBC* están sorprendentemente relacionados [175], lo que ha sido asociado con la adaptación específica de dichos linajes patogénicos a los grupos étnicos presentes en diferentes áreas geográficas [174]. La relevancia de estos hechos va mucho más allá de motivaciones teóricas, y tiene que ver con los peligros asociados a la introducción de vacunas con niveles heterogéneos de protección contra los diferentes linajes de *MTB* [177]. Si algo así sucediera, podría suponer la introducción de presiones evolutivas diferenciales entre linajes, lo cual podría tener consecuencias graves, como ha ocurrido anteriormente a lo largo de la Historia con otros patógeneos como *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Neisseria meningitides* y *Bordetella pertussis* [446]. Por ejemplo, una vacuna que ofreciera menos protección contra la familia de linajes asiáticos que podría tener consecuencias dramáticas causando un incremento en la incidencia de tuberculosis resistente asociada a la proliferación de cepas Beijing; pertenecientes a los linajes asiáticos y más proclives a desarrollar resistencia a los antibióticos. Para que estos aspectos puedan ser modelados y comprendidos, debería implementarse un modelo apropiado de meta-poblaciones para la TB; sobre el cual estudiar los diferentes niveles de compatibilidad entre los grupos filogenéticos de patógenos y huéspedes, y relacionarlos con su dispersión geográfica. En un modelo tal, los patrones de movilidad serían dirigidos por flujos migratorios sobre periodos temporales extensos, y otras fenomenologías, como la resistencia a los antibióticos y las interacciones con el VIH podrán ser y serán incorporadas. Idealmente, las evaluaciones de impacto de las vacunas deberían ser afrontadas dentro de este contexto, para lo cual el desarrollo del modelo de propagación de la TB que se presenta en esta Tesis en el capítulo 9 constituye un primer paso. No obstante, la implementación de una plataforma de modelado tal dependerá necesariamente de la disponibilidad de datos; todavía muy limitada especialmente en lo que respecta a los efectos de la diversidad genética sobre la dinámica de propagación de la TB. Por este motivo, una vez más, se hace necesario un conocimiento de los mecanismos genéticos que conllevan diferentes

niveles de susceptibilidad a la enfermedad, así como una descripción de su distribución en las poblaciones de diferentes países, para poder ofrecer descripciones robustas de la dinámica de propagación de la TB, cuando este tipo de cuestiones se tienen en consideración. Una vez más, la tuberculosis revela su complejidad en múltiples escalas, cada una de ellas inevitablemente asociada a las demás. Los avances realizados hasta ahora aparecen ante nosotros como piezas entrelazadas de un puzzle mayor, que son asimismo una *fuente de error, de duda, de desazón y de espera* a afrontar en futuros años de trabajo.

Bibliography

[1] World Health Organization. Global Tuberculosis Report 2013.

[2] D Young, J Stark and D Kirschner. Systems biology of persistent infection: tuberculosis as a case study. *Nature Rev. Microbiol.*, **6**, 520-528, (2008).

[3] J Sanz, J Navarro, J Arbués, C Martín, P Marijuán and Y Moreno. The transcriptional regulatory network of *Mycobacterium tuberculosis*. *PLoS One*, **6**, 7, e22178, (2011).

[4] Y Wang *et al.* Global protein-protein interaction network in the human pathogen *Mycobacterium tuberculosis* H37Rv. *J Proteome Res.*, **9**(12): 6665-6677, (2010).

[5] T Barrett and R Edgar. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, **411**: p. 352-369, (2006).

[6] J Sanz, E Cozzo, J Borge-Holthoefer and Y Moreno. Topological effects of data incompleteness of gene regulatory networks. *BMC Sys. Biol.*, **6**, 110, (2012).

[7] J Sanz, E Cozzo and Y Moreno. Data reliability in complex directed networks. *J. Stat. Mech.* **P12008**, (2013).

[8] R Pastor-Satorras and A Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, **86**, 3200-3203, (2001).

[9] S Meloni, N Perra, A Arenas, S Gomez, Y Moreno and A Vespignani. Modeling Human Mobility Responses to the Large-scale Spreading of Infectious Diseases. *Sci. Rep.*, **1**, 62, (2011).

[10] D Brockmann, L Hufnagel and T Geisel. The scaling laws of human travel. *Nature*, **439**, 462-465, (2006).

[11] F Liljeros, CR Edling, LAN Amaral, HE Stanley and Y Aberg. The web of human sexual contacts. *Nature*, **411**, 907-908, (2001).

[12] J Mossong, N Hens, M Jit, P Beutels, K Auranen et al. Social Contacts and Mixing Patterns Relevant tothe Spread of Infectious Diseases. *PLoS Med*, **5**, 3, e74, 0381-0391, (2008).

[13] R Guimerá, S Mossa, A Turtschi and LAN Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Nat. Acad. Sci. USA*, **102**, 7794-7799, (2005).

[14] V Colizza, A Barrat, M Barthélemy and A Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Nat. Acad. Sci. USA*, **103**, 2015-2020, (2006).

[15] D Balcan, H Hu, B Gonçalves, P Bajardi, C Poletto, JJ Ramasco, D Paolotti, N Perra, M Tizzoni, W Van der Broeck, V Colizza and A Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, **7**, 1, 45, (2009).

[16] M Tizzoni et al. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine*, **10**, 165, (2012).

[17] J Sanz, LM Floría and Y Moreno. Dynamics of persistent infections in homogeneous populations. *Int. J. Bifourc. Chaos*, **22**, 7, 125164 (8 pp.) (2012).

[18] J Sanz, LM Floría and Y Moreno. Spreading of persistent infections in heterogeneous populations. *Phys. Rev. E*, **81**, 056108/1-056108/9, (2010).

[19] J Sanz, C Xia, S Meloni and Y Moreno. Dynamics of Interacting diseases. *Phys. Rev. X*, in press, (2014).

[20] AERAS website www.aeras.org

[21] C Dye, P Glaziou, K Floyd and M Raviglione. Prospects for Tuberculosis Elimination. *Annu Rev Public Health*, **34**, 271-286, (2013).

[22] LJ Abu-Raddad et al. Epidemiological benefits of more effective tuberculosis vaccines, drugs and diagnostics. *Prod. Nat. Acad. Sci.*, **106**, 13980-13985, (2009).

[23] United Nations population division database. http://esa.un.org/unpd/wpp/index.htm

[24] World Health Organization tuberculosis database. http://www.who.int/tb/country/en/index.html

[25] G Perec. La vida: instrucciones de uso. Traducción al castellano de J Escué. *Anagrama*, Barcelona, (1988).

[26] B Killworth and H Bernard. Informant accuracy in social network data. *Human Org.*, **35**, 269-286, (1976).

[27] A Arenas, L Danon, A Díaz-Guilera, P Gleiser, and R. Guimerá, *Eur. Phys. J. B* **38**(2), 373 (2004).

[28] L Euler. *Commentarii academiae scientiarum Petropolitanae* (1741)

[29] DJ Watts and SH Strogatz. Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442, (1998).

[30] AL Barabasi and R Albert. Emergence of scaling in random networks. *Science* **286**, 509-512, (1999).

[31] R Albert and AL Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**(1), 47–51, (2002).

[32] M Kochen. The Small World Ablex, Norwood, NJ, (1989).

[33] P Erdös and A Rényi. On Random Graphs I. *Publ. Math.(Debrecen)*, **6** 290, (1959).

[34] P Erdös and A Rényi. On the evolution of random graphs. *Bull. Inst. Int. Stat.*, **38** 343, (1961).

[35] G Fagiolo G Clustering in complex directed networks. *Physical Review E*, **76**(2), 026107, (2007).

[36] M Newman and M Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E.*, **69**, 026113, (2004).

[37] HW Shen, XQ Cheng and JF Guo. Quantifying and identifying the overlapping community structure in networks. *J. Stat. Mech.* **P07042**, (2009).

[38] WW Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473, (1977).

[39] A Clauset, MEJ Newman and C Moore. Finding community structure in very large networks. *Phys Rev E*, **70**, 066111, (2004).

[40] V Blondel, J Guillaume, R Lambiotte and E Lefebvre. Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp*, **10**, P10008, (2008).

[41] J Reichardt and S Bornholdt. Statistical mechanics of community detection. *Phys Rev E*, **74**, 016110, (2006).

[42] MEJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E*, **74**, 036104, (2006).

[43] RM Anderson, RM May and B Anderson. Infectious diseases of humans: Dynamics and Control. Oxford University Press, UK, Oxford, (1992).

[44] DJ Daley and J Gani. Epidemic Modelling. Cambridge University Press, UK, Cambridge, (1999).

[45] JD Murray. Mathematical Biology. Springer-Verlag, Germany, Berlin, (2002).

[46] S Strogatz. Exploring complex networks. *Nature*, **410**, 268-276, (2001).

[47] L Hufnagel, D Brockmann and T Geisel. Forecast and control of epidemics in a globalized world. *Proc. Nat. Acad. Sci. USA*, **101**, 15124-15129, (2004).

[48] MY Li, JR Graef, L Wang and J Karsai. Global dynamics of a SEIR model with varying total population size. *Mathematical Biosciences*, **160**, 191-213, (1999).

[49] J Gómez-Gardeñes, V Latora, Y Moreno and E Profumo. Spreading of sexually transmitted diseases in heterosexual populations. *Proc. Nat. Acad. Sci. USA*, **105**, 1399-1404, (2008).

[50] S Boccaletti, V Latora, Y Moreno, M Chavez and DU Hwang. Complex Networks: Structure and Dynamics. *Phys. Rep.*, **424**, 4-5, 175-308, (2006).

[51] S Eubank, H Guclu, VL Anil-Kumar, MV Marathe, A Srinivasan, Z Toroczkai and N Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, **429**, 180-184, (2004).

[52] V Colizza, R Pastor-Satorras and A Vespignani. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, **3**, 276-282, (2007).

[53] V Colizza and A Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of Theoretical Biology*, **251**, 450-467, (2008).

[54] S Meloni, A Arenas and Y Moreno. Traffic-driven epidemic spreading in finite-size scale-free networks. *Proc. Nat. Acad. Sci. USA*, **106**, 16897-16902, (2009).

[55] MC González, CA Hidalgo and AL Barabasi. Understanding individual human mobility patterns. *Nature*, **453**, 779-782, (2008).

[56] NM Ferguson, DAT Cummings, S Cauchemez, C Fraser, S Riley, A Meeyai, S Iamsirithaworn and DS Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, **437**, 209-214, (2005).

[57] AL Lloyd and RM May. How Viruses Spread Among Computers and People. *Science*, **292**, 1316-1317, (2001).

[58] Y Moreno, R Pastor-Satorras and A Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B.*, **26**, 521-529, (2002).

[59] Y Moreno, JB Gómez and AF Pacheco. Epidemic incidence in correlated complex networks. *Phys. Rev. E*, **68**, 035103/1-035103/4, (2003).

[60] BM Murphy, BH Singer, S Anderson and S Kirschner. Comparing epidemic tuberculosis in demographically distinct heterogeneous populations. *Mathematical Biosciences*, **180**, 161-185, (2002).

[61] WO Kermack and AG McKendrick A Contribution to the Mathematical Theory of Epidemics. *Proc. of the Royal Society A* **115**(772): 700, (1927).

[62] WO Kermack and AG McKendrick Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proc. Roy. soc. Lon. Ser. A*, **138**(834), 55-83, (1932).

[63] N Bacaër. The model of Kermack and McKendrick for the plague epidemic in Bombay and the type reproduction number with seasonality. *J Math Biol*, **64**, 403–422, (2012).

[64] C Poletto, S Meloni, V Colizza, Y Moreno and A Vespignani. Host mobility drives pathogen competition in spatially structured populations. *PLoS comp. Biol.*, **9**, 8, e1003169, (2013).

[65] B Karrer and MEJ Newman. Competing epidemics on complex networks. *Phys. Rev. E.*, **84**, 036106, (2011).

[66] S Funk and VAA Jansen. Interacting epidemics on overlay networks. *Phys. Rev. E.*, **81**, 036118, (2010).

[67] MEJ Newman. Threshold effects for two pathogens spreading on a network. *Phys. Rev. Lett.*, **95**, 108701, (2005).

[68] V Marceau, PA Noël, L Hébert-Dufresne, A Allard and LJ Dubé. Modeling the dynamical interaction between epidemics on overlay networks. *Phys. Rev. E.*, **84**, 026105, (2011).

[69] JA Nelson, P Ghazal and CA Wiley. Role of opportunistic viral infections in AIDS. *AIDS*, **4**, 1-10, (1990).

[70] LH Kasper and D Buzoni-Gatel. Some Opportunistic Parasitic Infections in AIDS: Candidiasis, Pneumocystosis, Cryptosporidiosis, Toxoplasmosis. *Parasitology Today*, **14**, 150-156, (1998).

[71] M Nuño, Z Feng, M Martcheva and C Castillo-Chavez. Dynamics of Two-Strain Influenza with Isolation and Partial Cross-Immunity. *SIAM J. Appl. Math.*, **65**, 3, 964-982, (2006).

[72] HL Mills, A Ganesh, C Colijn. Pathogen spread on coupled networks: Effects of host and network properties on transmission threshold. *Journal of Theoretical Biology*, **320**, 47-57, (2013).

[73] PA Noël, A Allard, L Hébert-Dufresne, V Marceau and LJ Dubé. e-print arXiv:1102.0987.

[74] FD Sahneh and C Scoglio. May the Best Meme Win!: New Exploration of Competitive Epidemic Spreading over Arbitrary Multi-Layer Networks. arXiv:1308.4880 v2, (2013).

[75] S Gómez, A Arenas, J Borge-Holthoefer, S Meloni and Y Moreno. Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *Europhys. Lett.*, **89**, 38009, (2010).

[76] C Granell, S Gómez and A Arenas. Dynamical Interplay between Awareness and Epidemic Spreading in Multiplex Networks. *Phys. Rev. Lett.*, **111**, 128701, (2013).

[77] V Colizza, R Pastor-Satorras and A Vespignani. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, **3**, 276–282, (2007).

[78] M Ajelli, B Gonçalves *et al.* Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *BMC Infectious Diseases*, **10**, 190, (2010).

[79] I Gudelj, KAJ White and NF Britton. The Effects of Spatial Movement and Group Interactions on Disease Dynamics of Social Animals. *Bulletin of Mathematical Biology*, **66**, 91–108, (2004).

[80] WL Langer. The black death. *Sci Am*, **2**, 114-121, (1964).

[81] JV Noble. Geographic and temporal development of plagues. *Nature*, **250**, 726-729, (1974).

[82] BT Grenfell, ON Bjornstadt and J Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, **414**, 695-696, (2001).

[83] D Mollison. Dependence of epidemic and population velocities on basic parameters. *Math Biosc*, **107**, 255-287, (1991).

[84] R Sorek and P Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, **11**, 9-16, (2010).

[85] DA Day and MF Tuite. Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J Endocrinol*, **157**, 361-371, (1998).

[86] MR Fabian, N Sonenberg and W Filipowicz. Regulation of mRNA Translation and Stability by microRNAs. *Annu Rev Biochem*, **79**, 351–379, (2010).

[87] A Sirbu, H Ruskin and M Crane. Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC Bioinformatics*, **11**, 59, (2010).

[88] J Gonzalo-Asensio, S Mostowy, J Harders-Westerveen, K Huygen, R Hernández-Pando, J Thole, M Behr, B Gicquel and C Martín. PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. *PLoS One*, **3**(10), e3496.

[89] G Balazsi, AP Heath, L Shi and ML Gennaro. The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. *Mol. Sys. Biol.*, **4**, 225, (2008).

[90] M Babu, S Teichmann and L Aravind. Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks. *J. Mol. Biol.*, **358**, 614-633, (2006).

[91] Z Bar-Joseph et al. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, **21**, 1337-1342, (2003).

[92] I Comas, M Coscolla *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature Genetics*, **45**, 10, 1176-1182, (2013).

[93] MW Borgdorff, K Floyd and JF Broekmans. Interventions to reduce tuberculosis mortality and transmission in low- and middle-income countries. *Bulletin of the World Health Organization*, **80**, 3, 217-227, (2002).

[94] C Lienhardt, P Glaziou, M Uplekar, K Lonnroth, H Getahun and M Raviglione. Global tuberculosis control: lessons learnt and future prospects. *Nature Revs. Microbiol.*, **10**(6), 407-416. (2012).

[95] C Dye, S and BG Williams. Criteria for the control of drug-resistant tuberculosis. *Proc. Nat. Acad. Sci.*, **97**, 14, 8180-8185, (2000).

[96] EL Korenromp, F Scano, BG Williams, C Dye and P Nunn. Effects of Human Immunodeficiency Virus Infection on Recurrence of Tuberculosis after Rifampin-Based Treatment: An Analytical Review. *Clinical Infectious Diseases*, **37**, 1, 101-112, (2003).

[97] CL Tseng, O Oxlade, D Menzies, A Aspler and K Schwartzman. Cost-effectiveness of novel vaccines for tuberculosis control: a decision analysis study. *BMC Public Healt*, **11**(1), 55, (2011).

[98] C Dye and PE Fine. A major event for new tuberculosis vaccines. *The Lancet*, **381**, 9871, 972-974, (2013).

[99] IM Orme. Vaccine Development for Tuberculosis: Current Progress. *Drugs*, **73**, 10, 1015-1024, (2013).

[100] JC Leemans, NP Juffermans, S Florquin, N van Rooijen, MJ Vervoordeldonk, A Verbon, SJ van Deventer and T van der Poll. Depletion of alveolar macrophages exerts protective effects in pulmonary tuberculosis in mice. *J. Immunol.*, **166**, 4604–4611, (2001).

[101] AJ Wolf, B Linas, GJ Trevejo-Nunez, E Kincaid, T Tamura, K Takatsu and JD Ernst. *Mycobacterium tuberculosis* infects dendritic cells with high frequency and impairs their function in vivo. *J. Immunol.*, **179**, 2509–2519, (2007).

[102] MC Tsai, S Chakravarty, G Zhu, J Xu, K Tanaka, C Koch, J Tufariello, J Flynn and J Chan. Characterization of the tuberculous granuloma in murine and human lungs: cellular composition and relative tissue oxygen tension. *Cell. Microbiol.*, **8**, 218–232, (2006).

[103] JL Flynn and J Chan. What is good for the host is good for the bug. *Trends Microbiol.*, **13**, 98–102, (2005).

[104] DG Russell. Who puts the tubercle in tuberculosis? *Nat. Rev. Microbiol.*, **5**, 39-47, (2007).

[105] DG Russell, PJ Cardona, KJ Kim, S Allain and F Altare. Foamy macrophages and the progression of the human tuberculosis granuloma. *Nat. Immunol.*, **10**, 943–948, (2009).

[106] T Ulrichs and SH Kaufmann. New insights into the function of granulomas in human tuberculosis. *J. Pathol.*, **208**, 261-269, (2006).

[107] AM Cooper, DK Dalton, TA Stewart, JP Griffin, DG Russell and IM Orme. Disseminated tuberculosis in interferon gamma genedisrupted mice. *J. Exp. Med.*, **178**, 2243–2247, (1993).

[108] RJ North and YJ Jung. Immunity to tuberculosis. *Annu. Rev. Immunol.*, **22**, 599–623, (2004).

[109] B Musellim, S Erturan, E Sonmez Duman and G Ongen. Comparison of extrapulmonary tuberculosis cases: factors influencing the site of reactivation. *Int. J. Tuberc. Lung Dis.*, **9**, 11, 1220-1223, (2005).

[110] CE Barry, HI Boshoff *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* **7**, 845–855, (2009).

[111] S Sturgill-Koszycki, PH Schlesinger, P Chakraborty, PL Haddix, HL Collins, AK Fok, RD Allen, SL Gluck, J Heuser and DG Russell. Lack of acidification in Mycobacterium phagosomes produced by exclusion of the vesicular proton-ATPase. *Science*, **263**, 678–681, (1994).

[112] RM Yates, A Hermetter and DG Russell. The kinetics of phagosome maturation as a function of phagosome/lysosome fusion and acquisition of hydrolytic activity. *Traffic*, **6**, 413–420, (2005).

[113] JD MacMicking, RJ North, R LaCourse, JS Mudgett, SK Shah and CF Nathan. Identification of nitric oxide synthase as a protective locus against tuberculosis. *Proc. Natl. Acad. Sci. USA*, **94**, 5243–5248, (1997).

[114] JD MacMicking, GA Taylor and JD McKinney. Immune control of tuberculosis by IFN-gamma-inducible LRG-47. *Science*, **302**, 654–659, (2003).

[115] UE Schaible, S Sturgill-Koszycki, PH Schlesinger and DG Russell. Cytokine activation leads to acidification and increases maturation of Mycobacterium avium-containing phagosomes in murine macrophages. *J. Immunol.*, **160**, 1290–1296, (1998).

[116] LE Via, RA Fratti, M McFalone, E Pagan-Ramos, D Deretic and V Deretic. Effects of cytokines on mycobacterial phagosome maturation. *J. Cell Sci.*, **111**, 897–905, (1998).

[117] NM Nesbitt, X Yang, P Fontán, I Kolesnikova, I Smith, NS Sampson and E Dubnau. A Thiolase of *Mycobacterium tuberculosis* Is Required for Virulence and Production of Androstenedione and Androstadienedione from Cholesterol. *Infect. Immun.* **78**(1), 275-282, (2010).

[118] J Gonzalo-Asensio et al. The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in *Mycobacterium tuberculosis*. *J. Biol. Chem.*, **281**, 1313, (2006).

[119] A Singh, DK Crossman, D Mai, L Guidry, MI Voskuil, MB Renfrow, and AJ Steyn. *Mycobacterium tuberculosis* WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS pathogens*, **5**(8), e1000545, (2009).

[120] A Brzostek, B Dziadek, A Rumijowska-Galewicz, J Pawelczyk and J Dziadek. Cholesterol oxidase is required for virulence of *Mycobacterium tuberculosis*. *FEMS Microbiol. Lett.*, **275**, 106–112, (2008).

[121] JC Chang, MD Miner, AK Pandey, WP Gill, NS Harik CM Sassetti and DR Sherman. igr Genes and *Mycobacterium tuberculosis* cholesterol metabolism. *J. Bacteriol.*, **191**, 5232–5239, (2009).

[122] NM Nesbitt, X Yang, P Fontan, I Kolesnikova, I Smith, NS Sampson and E Dubnau. A thiolase of *Mycobacterium tuberculosis* is required for virulence and production of androstenedione and androstadienedione from cholesterol. *Infect. Immun.*, **78**, 275–282, (2010).

[123] S Savvi, DF Warner, BD Kana, JD McKinney, V Mizrahi and SS Dawes. Functional characterization of a vitamin B12-dependent methylmalonyl pathway in *Mycobacterium tuberculosis*: implications for propionate metabolism during growth on fatty acids. *J. Bacteriol.*, **190**, 3886–3895, (2008).

[124] EJ Munoz-Elias, AM Upton, J Cherian and JD McKinney. Role of the methylcitrate cycle in *Mycobacterium tuberculosis* metabolism, intracellular growth, and virulence. *Mol. Microbiol.*, **60**, 1109–1122, (2006).

[125] M Jain, CJ Petzold, MW Schelle, MD Leavell, JD Mougous, CR Bertozzi, JA Leary and JS Cox. Lipidomics reveals control of *Mycobacterium tuberculosis* virulence lipids via metabolic coupling. *Proc. Natl. Acad. Sci. USA* , **104**, 5133–5138, (2007).

[126] AM Abdallah *et al.* Type VII secretion mycobacteria show the way. *Nat. Rev.*, textbf5, 883–891, (2007).

[127] AS Pym *et al.* Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nat. Med.*, **9**, 533–539, (2003).

[128] T Hsu *et al.* The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proc. Natl. Acad. Sci. USA*, **100**, 12420–12425, (2003).

[129] A Arbués, JI Aguilo, J Gonzalo-Asensio, D Marinova, S Uranga, E Puentes, *et al.*. Construction, characterization and preclinical evaluation of MTBVAC, the first live-attenuated *M. tuberculosis*-based vaccine to enter clinical trials. *Vaccine*, **31**(42), 4867-4873, (2013).

[130] AT Kamath *et al.* Differential protective efficacy of DNA vaccines expressing secreted proteins of *Mycobacterium tuberculosis*. *Infect. Immun.*, **67**, 1702–1707, (1999).

[131] Immunogenicity and efficacy of a tuberculosis DNA vaccine encoding the components of the secreted antigen 85 complex E Lozes, K Huygen *et al. Vaccine*, **15**, 8, 830–833, (1997).

[132] L Solans, J Gonzalo-Asensio *et al.* The PhoP-Dependent ncRNA Mcr7 Modulates the TAT Secretion System in *Mycobacterium tuberculosis*. *PLOS Pathogens*, **10**, 5, e1004183, (2014).

[133] JE Galagan, K Minch, M Peterson, A Lyubetskaya, E Azizi, *et al.* The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature* **499**, 178–183, (2013).

[134] St Cole et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544, (1998).

[135] ST Cole and BG Barrell. Analysis of the genome of *Mycobacterium tuberculosis H37Rv*. *Novartis Found. Symp.* 217, 160–172, (1998).

[136] SD Chaparas, CJ Maloney and SR Hedrick. Specificity of tuberculins and antigens from various species of mycobacteria. *Am. Rev. Respir. Dis.* **101**, 74–83, (1970).

[137] M Harboe, RN Mshana, O Closs, G Kronvall and NH Axelsen. Cross-reactions between mycobacteria. II. Crossed immunoelectrophoretic analysis of soluble antigens of BCG and comparison with other mycobacteria. *Scand. J. Immunol.* **9**, 115–124 (1979).

[138] CF von Reyn *et al.* Dual skin testing with *Mycobacterium avium* sensitin and purified protein derivative to discriminate pulmonary disease due to *M. avium* complex from pulmonary disease due to *M.tuberculosis. J. Infect. Dis.* **177**, 730–736, (1998).

[139] P. Andersen and T.M. Doherty The success and failure of BCG - implications for a novel tuberculosis vaccine *Nat Rev Microbiol.* **3**(8):656-62, (2005).

[140] J van Ingen, R de Zwaan *et al.* Region of Difference 1 in Nontuberculous *Mycobacterium* Species Adds a Phylogenetic and Taxonomical Character. *Journal of Bacteriology*, **198**, 18, 5865–5867, (2009).

[141] R Brosch, SV Gordon SV, M Marmiesse *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl Acad. Sci.* **99**(6), 3684–3689, (2002).

[142] AE Hirsh, AG Tsolaki, K DeRiemer, MW Feldman and PM Small. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc. Natl. Acad. Sci.* **101**, 4871–4876, (2004).

[143] I Comas, and S Gagneux. The past and future of tuberculosis research. *PLoS Pathog.* **5**, e1000600 (2009).

[144] I Comas, J Chakravartti *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nature Genetics*, **42**, 6, 498–503, (2010).

[145] Explaining microbial genomic diversity in light of evolutionary ecology. OX Cordero and MF Polz. *Nat. Rev. Microbiol.* **12**, 263–273, (2014).

[146] T Rodrigo *et al.* Characteristics of tuberculosis patients who generate secondary cases. *Int. J. Tuberc. Lung Dis.* **1**, 352–357, (1997).

[147] XY Han and FJ Silva. On the Age of Leprosy. *PLOS Neglected Tropical Diseases*, **8**, 2, e2544, (2014).

[148] LG Wilson. Commentary: medicine, population, and tuberculosis. *Int. J. Epidemiol.* **34**, 521–524 (2005).

[149] T Paulson. Epidemiology: A mortal foe. *Nature*, **502**(7470), S2-S3, (2013).

[150] Robert Koch, the Nobel Prize, and the Ongoing Threat of Tuberculosis SHE Kaufmann, *New Eng. J. Med.* **353**;23, 2423–2426, (2005).

[151] LS Farer, AM Lowell and MP Meador. Extrapulmonary tuberculosis in the United States. *Am J Epidemiol*, **109**: 205–217, (1979).

[152] OY Gonzalez, G Adams, LD Teeter, TT Bui, JM Musser and EA Graviss. Extrapulmonary manifestations in a large metropolitan area with a low incidence of tuberculosis. *Int J Tuberc Lung Dis*, **7**, 12, 1178—1185, (2003).

[153] MR Weir. The enigma of extrapulmonary tuberculosis. *N Y State J Med*, **89**: 251–252, (1989).

[154] HM Al-Freihi, SA Al-Mohaya, EM Ibrahim, HY Al-Idrissi and I Baris. Extrapulmonary tuberculosis: diverse manifestations and diagnosis challenge. *East Afr. Med. J.*; **64**: 295–301, (1987).

[155] NC Elder. Extrapulmonary tuberculosis. A review. *Arch Fam Med*; **1**: 91–98, (1992).

[156] K Lonnroth, E Jaramillo, BG Williams, C Dye and M Raviglione. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine*, **68**, 2240–2246, (2009).

[157] RE Chaisson and NA Martinson. Tuberculosis in Africa: Combating an HIV-Driven Crisis. *New. Eng. J. Med.*, **358**, 1089-1092, (2008).

[158] Mariano Íñiguez y Ortiz and Máximo Hercilla García. La tuberculosis en la provincia de Soria. Memoria presentada al Primer Congreso de la Tuberculosis verificado en Zaragoza en 1908. Soria. Imprenta Felipe de las Heras (1909)

[159] R Wood, SD Lawn, J Caldwell, R Kaplan, K Middelkoop LG Bekker. Burden of New and Recurrent Tuberculosis in a Major South African City Stratified by Age and HIV-Status. *PLoS One* **6**(10) e25098, (2011).

[160] M Lauzardo, and D Ashkin. Phthisiology at the Dawn of the New Century. A Review of Tuberculosis and the Prospects for Its Elimination. *Chest*, **117**, 5, 1455–1473, (2000).

[161] H Waaler, A Geser and S Andersen. The use of mathematical models in the study of the epidemiology of tuberculosis. *American Journal of Public Health* **52**, 1002, (1962).

[162] H Waaler. A dynamic model for the epidemiology of tuberculosis. *American Review of Respiratory Disease* **98** 591, (1967).

[163] S Ferebee. An epidemiological model of tuberculosis in the united states. *Bulletin of the National Tuberculosis Association*, **53**, 4, (1967).

[164] C ReVelle, W Lynn and F Feldmann. Mathematical models for the economic allocation of tuberculosis control activities in developing nations. *American Review of Respiratory Disease* **96**, 893, (1967).

[165] C Castillo-Chavez and B Song. Dynamical Models of Tuberculosis and Their Applications. *Mathematical Biosciences and Engineering*, **1**, 2, 361–404, (2004).

[166] C Ozcaglar, A Shabbeer, SL Vandenberg, B Yener and KP Bennett. Epidemiological models of *Mycobacterium tuberculosis* complex infections. *Mathematical Biosciences*, **236**, 77–96, (2012).

[167] C Dye, S Scheele, P Dolin, V Pathania and MC Raviglione. Global burden of tuberculosis. Estimated incidence, prevalence, and mortality by country. *J. Am. Med. Asoc.*, **282**, 7, 677-686, (1999).

[168] E Brooks-Pollock, T Cohen and M Murray. The Impact of Realistic Age Structure in Simple Models of Tuberculosis Transmission. *PLOS One*, **5**, 1, e8479, (2010).

[169] HL Mills, T Cohen and C Colijn. Modelling the performance of isoniazid preventive therapy for reducing tuberculosis in HIV endemic settings: the effects of network structure. *J R Soc Interface*, **8**, 1510–1520, (2011).

[170] T Cohen and M Murray. Modeling epidemics of multidrug-resistant *M. tuberculosis* of heterogeneous fitness. *Nature Medicine*, **10**, 10, 1117–1121, (2004).

[171] N Bacaër, R Ouifki, C Pretorius, R Wood and B Williams. Modeling the joint epidemics of TB and HIV in a South African township. *J. Math. Biol.*, **57**:557-593, (2008).

[172] DW Dowdy, RE Chaisson, LH Moulton and SE Dorman. The potential impact of enhanced diagnostic techniques for tuberculosis driven by HIV: a mathematical model. *AIDS*, **20**, 5, 751–762, (2006).

[173] S Bowong, JJ Tewa and J Kurths. Dynamics of the spread of tuberculosis in heterogeneous complex meta-populations. *International Journal of Bifurcation and Chaos*, **23**, 7, 1350128, (2013).

[174] S Gagneux, K DeRiemer *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS*, **103**, 8, 2869–2873, (2006).

[175] S Gagneux. Host-pathogen coevolution in human tuberculosis. *Phil Trans R Soc*, **367**, 850-859, (2012).

[176] B López, D Aguilar *et al.* A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin Exp Immunol*, **133**, 30–37, (2003).

[177] T Cohen, C Colijn and M Murray. Modeling the effects of strain diversity and mechanisms of strain competition on the potential performance of new tuberculosis vaccines. *PNAS*, **105**, 43, 16302–16307, (2008).

[178] NR Gandhi, P Nunn *et al.* Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet*, **375**, 1830–1843, (2010).

[179] T Frieden, T Sterling, A Pablos-Mendez, J Kilburn, G Cauthen G and S Dooley. The emergence of drug-resistant tuberculosis in New York City. *N Engl J Med*, **328**: 521–26, (1993).

[180] J Rullán, D Herrera, R Cano, *et al.* Nosocomial transmission of multidrug-resistant *Mycobacterium tuberculosis* in Spain. *Emerg Infect Dis*, **2**: 125–29, (1996).

[181] GB Migliori, J Ortmann, E Girardi E, *et al.* Extensively drug-resistant tuberculosis, Italy and Germany. *Emerg Infect Dis*, **13**: 780–82, (1996).

[182] GB Migliori, G Besozzi *et al.* Clinical and operational value of the extensively drug-resistant tuberculosis definition. *Eur Respir J*, **30**, 4, 623–626, (2007).

[183] N Shah, J Richardson, P Moodley, *et al.* Increasing second-line drug resistance among extensively drug-resistant tuberculosis patients in rural South Africa. *40th Union World Conference on Lung Health*; Cancun, Mexico; Dec 3–7, 2009.

[184] Z Ma, C Lienhardt, H McIlleron, AJ Nunn and X Wang. Global tuberculosis drug development pipeline: the need and the reality. *Lancet*, **375**, 2100–2109, (2010).

[185] HS Cox, M Morrow and PW Deutschmann. Long term efficacy of DOTS regimens for tuberculosis: systematic review. *Bmj*, **336**, 7642, 484-487, (2008).

[186] D. Marinova, J. Gonzalo-Asensio, N. Aguilo and C. Martin *Recent development in tuberculosis vaccines* Expert Rev. Vaccines 12(12), 1431-1448 (2013)

[187] L Barker, L Hessel and B Walker. Rational approach to selection and clinical development of TB vaccine candidates. *Tuberculosis (Edinb)* **92**(Suppl. 1), S25–29, (2012).

[188] MJ Brennan and J Thole. Tuberculosis vaccines: a strategic blueprint for the next decade. *Tuberculosis (Edinb)* 92(Suppl. 1), S6–13 (2012).

[189] Understanding BCG is the key to improve it. H McShane. *Clinical Infectious Diseases* **58** (2014), Editorial Commentary.

[190] AS Pym, P Brodin, R Brosch, M Huerre and ST Cole. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Molecular Microbiology*, **46**, 3, 709–717, (2002).

[191] NG Van Pittius, J Gamieldien, W Hide, GD Brown, RJ Siezen, and AD Beyers. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+ C Gram-positive bacteria. *Genome Biol*, **2**(10), 44-1, (2001).

[192] BR Bloom and PEM Fine. The BCG experience: implications for future vaccines against tuberculosis. In: Bloom BR, ed. *Tuberculosis: pathogenesis, protection and control.* Washington, DC: Am. Soc. of Microbiol., 531-557, (1994).

[193] PEM Fine and LC Rodrigues. Modern vaccines: mycobacterial diseases. *Lancet*, **335**: 1016–1020, (1990).

[194] P Mangtani et al. *Protection by BCG Vaccine Against Tuberculosis: A Systematic Review of Randomized Controlled Trials* Clinical Infectious Diseases 2014; 58(4): 470-80

[195] M.L. Barreto et al. *Causes of variation in BCG vaccine efficacy: Examining evidence from the BCG REVAC cluster randomized trial to explore the masking and the blocking hypotheses* In Press

[196] T Oettinger, M Jorgensen, A Ladefoged, K Haslov and P Andersen. Development of the *Mycobacterium bovis* BCG vaccine: review of the historical and biochemical evidence for a genealogical tree. *Tubercle and Lung Disease*, **79**, 4, 243–250, (1999).

[197] PEM Fine Variation in protection by BCG: implications of and for heterologous immunity. *The Lancet*, **346**, 1339–1345, (1995).

[198] TF Brewer. Preventing Tuberculosis with Bacillus Calmette-Guérin Vaccine: A Meta-Analysis of the Literature. *CID*, **31**, 63–67, (2000).

[199] SP Zodpey and SN Shrikhande The geographic location (latitude of studies evaluating protective effect of BCG vaccine and its efficacy /effectiveness against tuberculosis. *Indian J. Public health*, **51**, 4, 205–210, (2007).

[200] GF Black *et al.* Patterns and implications of naturally acquired immune responses to environmental and tuberculous mycobacterial antigens in northern Malawi. *J. Infect. Dis.* **184**, 322–329, (2001).

[201] S Floyd *et al.* Kinetics of delayed-type hypersensitivity to tuberculin induced by bacille Calmette–Guérin vaccination in northern Malawi. *J. Infect. Dis.* **186**, 807–814 (2002).

[202] RE Weir *et al.* Interferon-γ and skin test responses of schoolchildren in southeast England to purified protein derivatives from *Mycobacterium tuberculosis* and other species of mycobacteria. *Clin. Exp. Immunol.* **134**, 285–294 (2003).

[203] Hart, P. D. and Sutherland, I. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. Br. Med. J. 2, 293–295 (1977).

[204] I Miceli *et al.* Evaluation of the effectiveness of BCG vaccination using the case-control method in Buenos Aires, Argentina. *Int. J. Epidemiol.* **17**, 629–634, (1988).

[205] FA al-Kassimi, MS al-Hajjaj, IO al-Orainey EA Bamgboye. Does the protective effect of neonatal BCG correlate with vaccine-induced tuberculin reaction? *Am. J. Respir. Crit. Care Med.* **152**, 1575–1578 (1995).

[206] CE Palmer and MW Long. Effects of infection with atypical mycobacteria on BCG vaccination and tuberculosis. *Am. Rev. Respir. Dis.* **94**, 553–568 (1966).

[207] L Brandt *et al.* Failure of the Mycobacterium bovis BCG vaccine: some species of environmental mycobacteria block multiplication of BCG and induction of protective immunity to tuberculosis. *Infect. Immun.* **70**, 672–678 (2002).

[208] L Grode, P Seiler, S Baumann *et al.* Increased vaccine efficacy against tuberculosis of recombinant *Mycobacterium bovis* Bacille Calmette-Guérin mutants that secrete listeriolysin. *J. Clin. Investig.* **115**(9), 2472–2479 (2005).

[209] L Grode, CA Ganoza, C Brohm, J Weiner, B Eisele and SH Kaufmann. Safety and immunogenicity of the recombinant BCG vaccine VPM1002 in a phase 1 open-label randomized clinical trial. *Vaccine* **31**(9), 1340–1348, (2013).

[210] BM Kagina, B Abel, TJ Scriba *et al.* Specific T cell frequency and cytokine expression profile do not correlate with protection against tuberculosis after Bacille Calmette-Guérin vaccination of newborns. *Am. J. Respir. Crit. Care Med.* **182**(8), 1073–1079 (2010).

[211] BM Kagina, B Abel, M Bowmaker *et al.* Delaying BCG vaccination from birth to 10 weeks of age may result in an enhanced memory CD4 T cell response. *Vaccine*, **27**(40), 5488–5495, (2009).

[212] DA Hokey and A Ginsberg. The current state of tuberculosis vaccines. *Hum. Vaccin. Immunother.* **9**(10) (2013).

[213] MD Tameris et al. Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. *The Lancet*, **381**,9871, 1021-1028, (2013).

[214] JAC Sterne, LC Rodrigues and IN Guedes. Does the efficacy of BCG decline with time since vaccination? *Int J Tuberc Lung Dis*, **2**, 3, 200–207, (1998).

[215] GW Comstock, SF Woolpert, and VT Livesay. Tuberculosis studies in Muscogee County, Georgia.Twenty-year evaluation of a community trial of BCG vaccination. *Public Health Rep.* **91**, 276–280, (1976).

[216] R Rustomjee, B McClain, MJ Brennan *et al.* Designing an adaptive phase II/III trial to evaluate efficacy, safety and immune correlates of new TB vaccines in young adults and adolescents. *Tuberculosis (Edinb)* **93**(2), 136–142 (2013).

[217] M Tameris, SJ Gelderbloem and R Rustomjee. An urgent call for a stronger, louder voice for TB vaccine advocacy. *Tuberculosis (Edinb)* **93**(3), 277–278 (2013).

[218] MA Behr, MA Wilson, WP Gill and P Small. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science*, **284**:1520–1523, (1999)

[219] C Martín, A Williams *et al.* The live Mycobacterium tuberculosis phoP mutant strain is more attenuated thanBCG and confers protective immunity against tuberculosis in mice and guinea pigs. *Vaccine*, **24**:3408–3419, (2006).

[220] D Aguilar, E Infante, C Martín, E Gormley E, B Gicquel and R Hernández Pando. Immunological responses and protective immunity against tuberculosis conferred by vaccination of Balb/C mice with the attenuated *Mycobacterium tuberculosis* (*phoP*) SO2 strain. *Clinical and Experimental Immunology*,147:330–338, (2007).

[221] PJ Cardona, JG Asensio *et al.* Extended safety studies of the attenuated live tuberculosis vaccine SO2 based on *phoP* mutant. *Vaccine*, **27**: 2499–2505, (2009).

[222] FA Verreck, RA Vervenne, I Kondova, *et al.* MVA.85A boosting of BCG and an attenuated, *phoP* deficient *M. tuberculosis* vaccine both show protective efficacy against tuberculosis in rhesus macaques. *PLoS ONE*, 4(4):e5264, (2009).

[223] AT Kamath, U Fruth *et al.* Newlive mycobacterial vaccines: the Geneva consensus on essential steps towards clinical development. *Vaccine* **23**:3753–3761, (2005).

[224] KB Walker, MJ Brennan, *et al.* The second Geneva consensus: recommendations for novel live TB vaccines. *Vaccine*, 28:2259–2270, (2010).

[225] T Palmer and BC Berks. The twin-arginine translocation (Tat) protein export pathway. *Nature Reviews*, **10**, 483–496, (2012).

[226] E Infante, LD Aguilar, B Gicquel and RH Pando. Immunogenicity and protective efficacy of the *Mycobacterium tuberculosis* fadD 26 mutant. *Clinical and Experimental Immunology*, **10**, 1111, 1365–2249, (2005).

[227] DG Russell, BC VanderVen, W Lee, RB Abramovitch, M Kim, S Homolka, S Niemann, and KH Rohde. *Mycobacterium tuberculosis* Wears What It Eats. *Cell Host & Microbe* **8**(1), 68–76, (2010).

[228] R Siméone, M Légert *et al.* Delineation of the roles of *FadD22*, *FadD26* and *FadD29* in the biosynthesis of phthiocerol dimycocerosates and related compounds in Mycobacterium tuberculosis. *FEBS J.* **277**, 2715–2725, (2010).

[229] M Bastian, S Heymann and M Jacomy. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, **8**, 361-362, (2009).

[230] M Kivelä, A Arenas, M Barthelemy, JP Gleeson, Y Moreno, MA Porter. Multilayer Networks. Journal of Complex Networks, in press.

[231] F Battiston, V Nicosia and V Latora. Structural measures for multiplex networks. *Phys. Rev. E*, **89**(3), 032804, (2014).

[232] D Schnappinger, S Ehrt, MI Voskuil, Y Liu, JA Mangan, IM Monahan, G Dolganov, B Efron, PD Butcher, C Nathan and GK Schoolnik. Transcriptional adaptation of Mycobacterium tuberculosis within macrophages insights into the phagosomal environment. *J. exp. med.*, **198**(5), 693-704, (2003).

[233] MI Voskuil, D Schnappinger, KC Visconti, MI Harrell, GM Dolganov, DR Sherman and GK Schoolnik. Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. *J. exp. med.* **198**(5), 705-713, (2003).

[234] PS Jackett, VR Aber, and DB Lowrie. Virulence and resistance to superoxide, low pH and hydrogen peroxide among strains of Mycobacterium tuberculosis. *J. gen. microbiol.* **104**(1), 37-45, (1978).

[235] M Steinbach, G Karypis and V Kumar. A comparison of document clustering techniques. *Text mining workshop.* In: Proc.of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston,MA, pp. 20-23 (2000).

[236] R Guimera and LAN Amaral. Cartography of complex networks: modules and universal roles. *JSTAT*, **2005(02)**, P02001, (2005).

[237] JM Lew, A Kapopoulou, LM Jones, and ST Cole. TubercuList-10 years after. *Tuberculosis*, **91**(1), 1-7, (2011).

[238] S Homolka *et al.* Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathogens*, **6**(7): p. e1000988, (2010).

[239] D Young. Animal models of tuberculosis. *Eur. j. immun.* **39**(8), 2011-2014. (2009).

[240] AM Sherrid, TR Rustad, GA Cangelosi and DR Sherman. Characterization of a Clp protease gene regulator and the reaeration response in *Mycobacterium tuberculosis*. *PLoS One* **16**;5(7):e11622 (2010).

[241] L Shi, R North, and ML Gennaro. Effect of growth state on transcription levels of genes encoding major secreted antigens of *Mycobacterium tuberculosis* in the mouse lung. *Infect. immun.* **72**(4), 2420-2424, (2004).

[242] Y Haile, G Bjune and HG Wiker. Expression of the mceA, esat-6 and hspX genes in *Mycobacterium tuberculosis* and their responses to aerobic conditions and to restricted oxygen supply. *Microbiology*, **148**, 3881-3886, (2002).

[243] P Brodin, I Rosenkrands, P Andersen, ST Cole and R Brosch. ESAT-6 proteins: protective antigens and virulence factors? *Trends in microbiology*, **12**(11), 500-508, (2004).

[244] AT Kamath, CG Feng, M Macdonald, H Briscoe and WJ Britton. Differential protective efficacy of DNA vaccines expressing secreted proteins of *Mycobacterium tuberculosis*. *Infect. Immun.* **67**(4), 1702-1707, (1999).

[245] AS Pym *et al.* Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nat. Med.* **9**, 533-539, (2003).

[246] T Hsu *et al.* The primary mechanism of attenuation of BCG is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proc. Natl. Acad. Sci.* **100**(21), 12420-12425, (2003).

[247] SV Gordon, R Brosch, A Billault, T Garnier, K Eiglmeier and ST Cole. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.* **32**(3), 643-656, (1999).

[248] MA Behr, MA Wilson, WP Gill, H Salamon, GK Schoolnik, S Rane and PM Small. Comparative genomics of BCG vaccines by whole-genome DNA microarrays. *Science* **284** (5419), 1520-1523, (1999).

[249] P Andersen, ME Munk, JM Pollock and TM Doherty. Specific immune-based diagnosis of tuberculosis. *Lancet* **356**(9235), 1099-1104, (2000).

[250] World Health Organization. Global tuberculosis control. WHO report 2000. World Health Organization Document WHO/CDS/TB/2000.275, 1179. World Health Organization, Geneva, (2000).

[251] EL Corbett, CJ Watt, N Walker, D Maher, BG Williams et al. The Growing Burden of Tuberculosis: Global Trends and Interactions with the HIV Epidemic. *Arch. Intern. Med.*, **163**, 1009-1021, (2003).

[252] ST Cole, R Brosch, J Parkhill, T Garnier and C Churcher. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537-544, (1998).

[253] AM Dannenberg and GSW Rook. Pathogenesis of pulmonary tuberculosis: interplay of tissue-damaging and macrophage-activating immune responses - dual mechanisms that control bacillary multiplication. Bloom BR., editor. Tuberculosis: pathogenesis, protection, and control. ASM Press, Washington, 459-484, (1994).

[254] HI Boshoff and CE Barry. Tuberculosis-metabolism and respiration in the absence of growth. *Nature Rev. Microbiol.*, **3**, 70-80, (2005).

[255] J Gonzalo-Asensio, S Mostowy, J Harders-Westerveen, k Huygen, R Hernandez-Pando et al. The *Mycobacterium tuberculosis* phoPR Operon Is Positively Autoregulated in the Virulent Strain H37Rv. *PLoS ONE*, **3**, e3496, (2008).

[256] D Young and C Dye. The development and impact of tuberculosis vaccines. *Cell*, **124**, 4, 683-687, (2006).

[257] C Martín. Tuberculosis vaccines: past, present and future. *Curr. Opin. Pulm. Med.*, **12**, 3, 186-189, (2006).

[258] C Martin, A Williams, R Hernandez-Pando, PJ Cardona, E Gormley et al. The live *Mycobacterium tuberculosis* phoP mutant strain is more attenuated than BCG and confers protective immunity against tuberculosis in mice and guinea pigs. *Vaccine*, **24**, 17, 3408-3419, (2006).

[259] R Albert and AL Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 1, 47-97, (2002).

[260] MEJ Newman. The structure and function of complex networks. *SIAM Rev.*, **45**, 2, 167-256, (2003).

[261] S Shen-Orr, R Milo, S Mangan and U Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.*, **31**, 64-68, (2002).

[262] R Milo, S Shen-Orr, S Itzkovitch, N Kashtan, D Chklovskii et al. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, **298**, 5594, 824-827, (2002).

[263] P Roback, J Beard, D Baumann, C Gille and K Henry. A predicted operon map for *Mycobacterium tuberculosis*. *Nuc. Acid. Res.*, **35**, 15, 5085-5095, (2007).

[264] RC Taylor, AK Brown, A Singh, A Bhatt and GS Besra. Characterization of a β-hydroxybutyryl-CoA dehydrogenase from Mycobacterium tuberculosis *Microbiology*, **156**, 1975-1982, (2010).

[265] S Maslow and K Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, **296**, 5569, 910-913, (2002).

[266] R Milo, N Kashtan, S Itzkovitz, MEJ Newman and U Alon. On the uniform generation of random graphs with prescribed degree sequences. (2003). Available: http://arxiv.org/abs/condmat/0312028. Accesed 2011 Jun 21.

[267] R Milo, S Itzkovitch, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, M Sheffer and U Alon. Superfamilies of Evolved and Designed Networks. *Science*, **303**, 5663, 1538-1542, (2004).

[268] S Mangan, A Zaslaver and U Alon. The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks. *J. Mol. Biol.*, **334**, 2, 197-204, (2003).

[269] S Mangan, U Alon. Structure and function of the feed-forward loop network motif. *Proc. Nat. Acad. Sci.*, **100**, 21, 11980-11985, (2003).

[270] N Rosenfeld and U Alon. Response Delays and the Structure of Transcription Networks. *J. Mol. Biol.*, **329**, 4, 645-654, (2003).

[271] S Basu, S Mehreja, S Thiberge, MT Chen and R Weiss. Spatiotemporal control of gene expression with pulse-generating networks. *Proc. Nat. Acad. Sci. USA*, **101**, 17, 6355-6360, (2004).

[272] N Rosenfeld, MB Elowitz and U Alon. Negative Autoregulation Speeds the Response Times of Transcription Networks. *J. Mol. Biol.*, **323**, 5, 785-793, (2002).

[273] A Zaslaver, AE Mayo, R Rosenberg, P Bashkin, H Sberro et al. Just-in-time transcription program in metabolic pathways. *Nature Genet.*, **36**, 5, 486-491, (2004).

[274] M Ronen, R Rosenberg, BI Shraiman and U Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Nat. Acad. Sci. USA*, **99**, 16, 10555-10560, (2002).

[275] U Alon. Network motifs: theory and experimental approaches. *Nature Rev. Genet.*, **8**, 450-461, (2007).

[276] R Guimera and M Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Nat. Acad. Sci. USA*, **106**, 52, 22073-22078, (2009).

[277] J Gómez-Gardeñes, LM Floria and Y Moreno. Scale-free topologies and activatory-inhibitory interactions. *Chaos*, **16**, 1, 015114, (2006).

[278] R Albert. Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947-4957, (2005).

[279] Signed version of the transcriptional regulatory network of *M.tuberculosis* [`http://cosnet.bifi.es/research-lines/systems-biology/data`].

[280] SB Walters et al. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. *Mol. Microbiol.*, **60**, 312, (2006).

[281] T Parish et al. The senX3-regX3 two-component regulatory system of *Mycobacterium tuberculosis* is required for virulence. *Microbiology*, **149**, 1423, (2003).

[282] HD Park et al. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **48**, 833, (2003).

[283] M Santangelo et al. Mce2R from *Mycobacterium tuberculosis* represses the expression of the mce2 operon. *Tuberculosis*, **89**, 22, (2009).

[284] B Abomoelak et al. mosR, a Novel Transcriptional Regulator of Hypoxia and Virulence in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **191**, 5941, (2009).

[285] A Maciag et al. Global Analysis of the *Mycobacterium* tuberculosis Zur (FurB) Regulon. *J. Bacteriol.*, **189**, 730, (2007).

[286] M Guo et al. Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system. *Genome Res.*, **19**, 1301, (2009).

[287] H He et al. MprAB Is a Stress-Responsive Two-Component System That Directly Regulates Expression of Sigma Factors SigB and SigE in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **188**, 2134, (2006).

[288] P Fontan et al. *Mycobacterium tuberculosis* Sigma Factor E Regulon Modulates the Host Inflammatory Response. *J. Infect. Dis.*, **198**, 877, (2008).

[289] R Manganelli et al. The *Mycobacterium tuberculosis* ECF sigma factor SigE: role in global gene expression and survival in macrophages. *Mol. Microbiol.*, **41**, 423, (2001).

[290] EP Williams et al. *Mycobacterium tuberculosis* SigF Regulates Genes Encoding Cell Wall-Associated Proteins and Directly Regulates the Transcriptional Regulatory Gene phoY1. *J. Bacteriol.*, **189**, 4234, (2007).

[291] S Raman et al. *Mycobacterium tuberculosis* SigM Positively Regulates Esx Secreted Protein and Nonribosomal Peptide Synthetase Genes and Down Regulates Virulence-Associated Surface Lipid Synthesis. *J. Bacteriol.*, **188**, 8460, (2006).

[292] JH Lee et al. Role of Stress Response Sigma Factor SigG in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **190**, 1128, (2008).

[293] JH Lee et al. Roles of SigB and SigF in the *Mycobacterium tuberculosis* Sigma Factor Network. *J. Bacteriol.*, **190**, 699, (2008).

[294] Y Akhter et al. Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis. *Gene*, **407**, 148, (2008).

[295] JC Micklinghoff et al. Role of the transcriptional regulator RamB (Rv0465c) in the control of the glyoxylate cycle in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **191**, 7260, (2009).

[296] S Raghavan et al. Secreted transcription factor controls *Mycobacterium tuberculosis* virulence. *Nature*, **454**, 717, (2008).

[297] MCM Reddy et al. Crystal structure of *Mycobacterium tuberculosis* LrpA, a leucine-responsive global regulator associated with starvation response. *Prot. Sci.*, **17**, 159, (2008).

[298] MA Gazdik et al. Rv1675c (cmr) regulates intramacrophage and cyclic AMP-induced gene expression in *Mycobacterium tuberculosis*-complex mycobacteria. *Mol. Microbiol.*, **71**, 434, (2009).

[299] N Agarwal et al. Characterization of the *Mycobacterium tuberculosis* Sigma Factor SigM by Assessment of Virulence and Identification of SigM-Dependent Genes. *Infect. Inmun.*, **75**, 452, (2007).

[300] N Andreu Martín. Estudio de la implicación de los genes Rv0576-Rv0577 en la tinción con rojo neutro y la virulencia de *Mycobacterium tuberculosis*. *Doctoral Thesis*. Universidad Autónoma de Barcelona (UAB), (2007).

[301] GM Rodriguez et al. ideR, an essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect. Inmun.*, **70**, 3371, (2002).

[302] T Liu et al. CsoR is a novel Mycobacterium tuberculosis copper-sensing transcriptional regulator. *Nat. Chem. Biol.*, **3**(1), 60-68, (2006).

[303] KA Kantardjieff et al. Structure of pyrR (Rv1379) from *Mycobacterium tuberculosis*: a persistence gene and protein drug target. *Acta Cryst.*, **D61**, 355, (2005).

[304] H He and TC Zahrt. Identification and Characterization of a Regulatory Sequence Recognized by *Mycobacterium tuberculosis* Persistence Regulator MprA. *J. Bacteriol.*, **187**, 202, (2005).

[305] A Singh et al. mymA operon of *Mycobacterium tuberculosis*: its regulation and importance in the cell envelope. *FEMS Microbiol. Lett.*, **227**, 53, (2003).

[306] M Santangelo et al. Mce3R, a TetR-type transcriptional repressor, controls the expression of a regulon involved in lipid metabolism in *Mycobacterium tuberculosis*. *Microbiology*, **155**, 2245, (2009).

[307] SL Kendall et al. A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **65**, 684, (2007).

[308] RM Goldstone et al. The Transcriptional Regulator Rv0485 Modulates the Expression of a pe and ppe Gene Pair and Is Required for *Mycobacterium tuberculosis* Virulence. *Infect. Immun.*, **77**, 4654, (2009).

[309] C Sala et al. Genome-wide regulon and crystal structure of BlaI (Rv1846c) from *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **71**, 1102, (2009).

[310] RegulonDB website: http://regulondb.ccg.unam.mx/ Accesed 2011 Jun 21.

[311] S Gama-Castro, V Jiménez-Jacinto, M Peralta-Gil, A Santos-Zavaleta, M Peñaloza-Spinola et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nuc. Acid. Res.*, **36**, D120-D128, (2008).

[312] J Gonzalo-Asensio, CY Soto, A Arbués, J Sancho, MC Menéndez et al. The Mycobacterium tuberculosis phoPR Operon Is Positively Autoregulated in the Virulent Strain H37Rv. *J. Bacteriol.*, **190**, 21, 7068-7078, (2008).

[313] Uri Alon's web site: http://www.weizmann.ac.il/mcb/UriAlon/ Accesed 2011 Jun 21.

[314] S Gama-Castro et al. RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nuc Acids Res*, **39**, Database Issue D98-D105, (2011).

[315] N Sierro, Y Makita, MJL de Hoon and K Nakai. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nuc Acids Res*, **36**, Database Issue D93-D96, (2008).

[316] PE Jacques, AL Gervais, M Cantin, JF Lucier, G Dallaire, G Drouin, L Gaudreau, J Goulet and R Brzezinski. MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*. *Bioinformatics*, **21**, 2563-2565, (2005).

[317] E de Silva, T Thorne, P Ingram, I Agrafioti, J Swire, C Wiuf and MPH Stumpf. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol*, **4**, 39, (2006).

[318] S Fortunato. Community detection in graphs. *Phys. Rep.*, **486**, 75-174, (2010).

[319] A Arenas, A Fernández and S Gómez. Analysis of the structure of complex networks at different resolution levels. *New J. Phys.*, **10**, 053039, (2008).

[320] S Gómez, P Jensen and A Arenas. Analysis of community structure in networks of correlated data. *Phys. Rev. E.*, **80**, 016114, (2009).

[321] V Spirin, M Gelfand, A Mironov and L Mirny. A metabolic network in the evolutionary context: Multiscale structure and modularity. *Proc. Nat. Acad. Sci.*, **103**, 23, 8774-8779, (2006).

[322] S Fortunato and M Barthélemy. Resolution limit in community detection. *Proc. Nat. Acad. Sci.*, **104**, 1, 36-41, (2007).

[323] E Ravasz, AL Somera, DA Mongru, ZN Oltvai and AL Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 5586, 1551-1555, (2002).

[324] S Gómez, A Fernández, J Borge-Holthoefer and A Arenas.**radatools.php**, `[http://deim.urv.cat/~sgomez/]`.

[325] J Lew, A Kapopoulou, L Jones and S Cole. Tuberculist: 10 years after. *Tuberculosis Edinb.*, **91**, 1, 1-7, (2011).

[326] L Kuncheva, S Hadjitodorov. Using diversity in cluster ensembles. *Systems, Man and Cybernetics, IEEE International Conference on Systems, man and Cybernetics. Volume 2*, 1214-1219, (2004).

[327] WM Rand. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Asoc.*, **66**, 336, 846-850, (1971).

[328] L Hubert and P Arabie. Comparing partitions. *J Classif*, **2**, 1, 193-218, (1985).

[329] EB Fowlkes and CL Mallows. A Method for Comparing Two Hierarchical Clusterings. *J. Am. Stat. Asoc.*, **78**, 383, 553-569, (1983).

[330] M Meila. Comparing clusterings: an information based distance. *J Multivariate Anal*, **98**, 5, 873-895, (2007).

[331] B Serrour, A Arenas, and S Gómez. Detecting communities of triangles in complex networks using spectral optimization. *Comp. Comm.*, **34**(5), 629-634, (2011).

[332] LF da Costa, FA Rodrigues, G Travieso and PR Villas-Boas. Characterization of complex networks: A survey of measurements. *Adv Phy*, **56**, 1, 167-242, (2007).

[333] M Costanzo et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge Library, an integrated resource for protein information. *Nuc Acids Res*, **29**, 1, 75-79, (2001).

[334] Database of synaptic connectivity of C. elegans for computation. *Technical report of Cybernetic Caenorhabditis elegans Program*, (2003). `[http://ims.dse.ibaraki.ac.jp/ccep/]`.

[335] **In [267] –see note 12 there–, feedback loops are cancelled when supposing less than** 0.1% **of network links. According to that convention, the only feedback loop in yeast TRN –which obviously could not be rewired– is cancelled.**

[336] **An operon based representation is not available for the TRN of *Mycobacterium tuberculosis* because of that a global enough experimental characterization of its operon map has not been accomplished yet. To our knowledge, most relevant works in this area –see, for example: [263]– consist yet of general computational predictive tools.**

[337] S Mangan, S Itzkovitz, A Zaslaver and U Alon. The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli. J Mol Biol*, **356**, 5, 1073-1081, (2006).

[338] Z Burda, A Krzywicki, OC Martin and M Zagorski. Motifs emerge from function in model gene regulatory networks. *Proc Nat Acad Sci*, **108**, 42, 17263-17268, (2011).

[339] Y Artzy-Randrup, SJ Fleighman , N Ben-Tal and L Stone. Comment on Network motifs: Simple Building Blocks of Complex Networks and Superfamilies of Evolved and Designed networks. *Science*, **305**, 1107, (2004).

[340] P Dwight Kuo, W Banzhaf and A Leier. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, **85**, 3, 177-200, (2006).

[341] S Huang. Back to the biology in systems biology: What can we learn from biomolecular networks? *Briefings Funct. Genomics*, **2**, 4, 279-297, (2004).

[342] A Mazurie, S Bottani and M Vergassola. An evolutionary and functional assessment of regulatory network motifs. *Genome Biol.*, **6**, R35, (2005).

[343] N Banerjee and M Zhang. Functional genomics as applied to mapping transcription regulatory networks. *Curr. Op. Microbiol.*, **5**, 313-317, (2002).

[344] H Bolouri and EH Davidson. Modeling transcriptional regulatory networks. *BioEssays*, **24**, 12, 1118-1129, (2002).

[345] MM Babu, NM Luscombe, L Aravind, M Gerstein and SA Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr. Op. Struct. Biol.*, **14**, 283-291, (2004).

[346] JA Dunne, RJ Williams and ND Martinez. Food-web structure and network theory: The role of connectance and size. *Proc. Natl. Acad. Sci.*, **99**, 20, 12917-12922, (2002).

[347] M Conover, J Ratkiewicz, M Francisco, B Gonçalves, A Flammini and F Menczer. Political polarization on twitter. *Proc. 5th Intl. Conference on Weblogs and Social Media*, **89**, (2011).

[348] J Borge-Holthoefer et al. Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. *PLoS One*, **6**, 8, e23883, (2011).

[349] G Kossinets. Effects of missing data in social networks. *Soc. Networks*, **28**, 247-268, (2006).

[350] JL Schafer and JW Graham. Missing data: our view of the state of the art. *Psychol. Methods*, **7**, 2, 147-177 (2002).

[351] CT Butts. Network inference, error and informant (in) accuracy: a Bayesian approach. *Soc. Networks*, **25**, 103-140, (2003).

[352] S Draghici, P Khatri, AC Eklund and Z Szallasi. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genet.*, **22**, 2, 101-109, (2006).

[353] J Brettschneider, F Collin, BM Bolstad and TP Speed. Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics*, **50**, 3, 241-264, (2008).

[354] JP Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 2, 149-155, (2008).

[355] SA Bustin et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinic. Chem.*, **55**, 4, 611-622, (2009).

[356] RG Rutledge and D Stewart. Assessing the Performance Capabilities of LRE-Based Assays for Absolute Quantitative Real-Time PCR. *PLoS One*, **5**, 3, e9731, (2010).

[357] AA Margolin, T Palomero, P Sumazin, A Califano, AA Ferrando and G Stolovitzky. ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. *Proc. Nat. Acad. Sci.*, **106**, 1, 244-249, (2009).

[358] Y Makita, M Nakao, N Ogasawara and K Nakai. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nuc. Acid Res.*, **32**, Database Issue D75-D77, (2004).

[359] M Hecker, S Lambeck, S Toepfer, E Van Someren and R Guthke. Gene regulatory network inference: Data integration in dynamic models. A review. *Biosystems*, **96**, 1, 86-103, (2009).

[360] R Jansen et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449-453, (2003).

[361] A Clauset, C Moore and MEJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, **453**, 98-101, (2008).

[362] B Yan and S Gregory. Finding missing edges in networks based on their community structure. *Phys. Rev. E*, **85**, 056112, (2012).

[363] L Lü and T Zhou. Link prediction in complex networks: A survey. *Phys. A*, **390**, 1150-1170, (2011).

[364] M Kim and J Leskovec. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SIAM*, International conference on data mining, 47-58, (2011).

[365] QM Zhang, L Lü, WQ Wang, YX Zu and T Zhou. Potential theory for directed networks. *PLoS One*, **8**, 2, e55437, (2013).

[366] B Prud'homme, N Gompel and SB Carroll. Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci.*, **104**, 1, 8605-8612, (2007).

[367] Bill Cherowitzo's graph theory glossary: http://www-math.ucdenver.edu/w̃cherowi/courses/m4408/m4408f.html

[368] pajek datasets web page: http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm

[369] ME Monaco and RE Ulanowicz. Comparative ecosystem trophic structure of three US mid-Atlantic estuaries. *Mar. Ecol. Prog. Ser.*, **161**, 239-254, (1997).

[370] RE Ulanowicz, JJ Heymans and MS Egnotovich. Network analysis of trophic dynamics in South Florida ecosystems, FY 99: the graminoid ecosystem. *Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES]*, CBL 00-0176, (2000).

[371] C Espinosa-Soto, P Padilla-Longoria and ER Alvarez-Buylla. A Gene Regulatory Network Model for Cell-Fate Determination during *Arabidopsis thaliana* Flower Development That Is Robust and Recovers Experimental Gene Expression Profiles. *The Plant Cell*, **16**, 2923-2939, (2004).

[372] A Stathopoulos and M Levine. Genomic Regulatory Networks and Animal Development. *Developmental Cell*, **9**, 449-462, (2005).

[373] GM Suel, J García-Ojalvo, LM Liberman and MB Elowitz. An excitable gene regulatory circuit induces transient cellular differentiation. *Nat. Lett.*, **440**, 545-550, (2006).

[374] M Tumminello, T Aste, T Di Matteo and RN Mantegna. A tool for filtering information in complex systems. *Proc. Nat. Acad. Sci.*, **102**, 30, 10421-10426, (2005).

[375] MA Serrano, M Boguñá and A Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proc. Nat. Acad. Sci.*, **106**, 16, 6483-6488, (2009).

[376] F Radicchi, JJ Ramasco and S Fortunato. Information filtering in complex weighted networks. *Phys. Rev. E*, **83**, 046101, (2011).

[377] D Bleed, C Watt and C Dye. World Health Report 2001: Global Tuberculosis Control, Technical Report. *world health organization, WHO/CDS/TB/2001.287*, (2001). Available from <http://www.who.int/gtb/publications/globrep01/index.html>

[378] SM Blower, AR Mclean, TC Porco, PM Small, PC Hopewell, MA Sanchez and AR Moss. The intrinsic transmission dynamics of tuberculosis epidemics. *Nature Medicine*, **1**, 8, 815-821, (1995).

[379] GW Comstock. Epidemiology of Tuberculosis. *Am. Rev. Respirat. Dis.*, **125**(3), 8-15, (1982).

[380] T Daniel, J Bates and K Downes. Tuberculosis: Pathogenesis, Protection and Control. in History of tuberculosis BR Bloom, ed. Washington, DC: American Society for Microbiology press, pps 13-24 (1994).

[381] HW Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, **42**, 599-653, (2000).

[382] K Styblo. Recent Advances in Respiratory Medicine. DC Flenley and TL Petty, eds. Churchill Livingstone, Edinbourgh, Vol 4, 77-108, (1986).

[383] K Styblo, J Meijer and L Sutherland. Tuberculosis Surveillance Research Unit Report No. 1: the transmission of tubercle bacilli; its trend in a human population. *Bull.Int. Union Tuberc.*, **42**, 1-104, (1969).

[384] LG Wilson. The Historical Decline of Tuberculosis in Europe and America: Its Causes and Significance. *J Hist Med Allied Sci*, **45**(3), 366-396, (1990).

[385] M Faloutsos, P Faloutsos and C Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM '99, Comput. Commun. Rev.*, **29**(4), 251-262, (1999).

[386] G Caldarelli, R Marchetti and L Pietronero. The fractal properties of Internet. *Europhys. Lett.*, **52**, 386, (2000).

[387] R Albert, H Jeong and AL Barabási. Internet: Diameter of the World-Wide Web. *Nature*, **401**, 130, (1999).

[388] DS Callaway, MEJ Newman, SH Strogatz and DJ Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, **85**, 5468-5471, (2000).

[389] R Cohen, K Erez, D ben Avraham and S Havlin. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.*, **86**, 3682-3685, (2001).

[390] MEJ Newman. Spread of epidemic disease on networks. *Phys. Rev. E.*, **66**, 016128, (2002).

[391] A Liccardo and A Fierro A lattice model for influenza spreading. *PloS one*, **8**(5), e63935, (2013).

[392] E Vynnycky and PEM Fine. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiol. Infect.* **119**, 183-201, (1997).

[393] A van Rie, N Beyers, RP Gie,M Kunneke, L Zietsman and PR Donald. Childhood tuberculosis in an urban population in South Africa: burden and risk factor. *Arch. Dis. Child.* **80**:433-437, (1999).

[394] J Marro and R Dickman. Nonequilibrium phase transitions in lattice models. Cambridge University Press, Cambridge, (2005).

[395] LJS Allen. Some Discrete-Time SI, SIR, and SIS Epidemic Models. *Math. Biosciences*, **124**, 83-105, (1994).

[396] HE Stanley. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, Oxford, (1987).

[397] HW Hethcote. Qualitative analyses of communicable disease models. *Math. Biosciences*, **28**, 3-4, 335-336, (1976).

[398] R Pastor-Satorras and A Vespignani. Evolution and Structure of the Internet: a statistical physics approach. Cambridge University Press, Cambridge, (2007).

[399] SN Dorogovtsev, AV Goltsev and JFF Mendes. Critical phenomena in complex networks. *Rev. Mod. Phys.*, **80**, 1275-1336, (2008).

[400] R Pastor-Satorras and A Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E.*, **63**, 066117, (2001).

[401] C Castellano and R Pastor-Satorras. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.*, **105**, 218701, (2010).

[402] M Boguñá and R Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E*, **66**, 047104, (2002).

[403] M Barthélemy, A Barrat, R Pastor-Satorras, and A Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys. Rev. Lett.*, **92**, 178701, (2004).

[404] M Boguña, C Castellano and R Pastor-Satorras. Langevin approach for the dynamics of the contact process on annealed scale-free networks. *Phys. Rev. E.*, **79**, 036110, (2009).

[405] B Guerra and J Gómez-Gardeñes. Annealed and Mean-field formulations of disease dynamics on static and adaptive networks. *Phys. Rev. E.*, **82**, 035101(R), (2010).

[406] V Belik, T Geisel and D Brockmann. Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Phys. Rev. X.*, **1**, 011001, (2011).

[407] P Wang, MC González, CA Hidalgo and AL Barabási. Understanding the Spreading Patterns of Mobile Phone Viruses. *Science*, **324**, 3, 1071-1076, (2009).

[408] M de Domenico et al. Mathematical Formulation of Multilayer Networks. *Phys. Rev. X.*, **3**, 041022, (2013).

[409] We further assume $P(k, l)$ to be stationary, i.e., the composed distribution does not change.

[410] ZS AL-Salloum. Defensive computer worms: an overview. *J. of Security and Networks*, **7**, 1, 59-70, (2012).

[411] United Nations HIV datable UNAIDS. www.unaids.org

[412] MC Boily, C Lowndes and M Alary. The impact of HIV epidemic phases on the effectiveness of core group interventions: insights from mathematical models. *Sex Transm. Infect.*, **78**, i78-i90, (2002).

[413] JJ Moré. The Levenberg-Marquardt algorithm: implementation and theory. Numerical Analysis, Springer Berlin Heidelberg, 105-116, (1978).

[414] Au-Yeung et al. Tuberculosis mortality in HIV-infected individuals: a cross-national systematic assessment. *Clin. Epidemiol.*, **3**, 21-29, (2011).

[415] CC Huang, ET Tchetgen, MC Becerra, T Cohen, KC Hughes, Z Zhang, R Calderon, R Yataco, C Contreras, J Galea, L Lecca and M Murray. The effect of HIV-related immunosuppression on the risk of tuberculosis transmission to household contacts. *Clin. Infect. Dis.*, **58**(6):765-74. (2014).

[416] MA Espinal, EN Pérez, J Báez, L Henríquez, K Fernández, M López, P Olivo and AL Reingold. Infectiousness of Mycobacterium tuberculosis in HIV-1-infected patients with tuberculosis: a prospective study. *Lancet*, **355**(9200):275-280 (2000).

[417] AM Elliott, RJ Hayes, B Halwiindi, N Luo, G Tembo, JO Pobee, PP Nunn and KP McAdam. The impact of HIV on infectiousness of pulmonary tuberculosis: a community study in Zambia. *AIDS* **7**(7):981-987, (1993).

[418] F Tria, S Pompei and V Loreto. Dynamically correlated mutations drive human Influenza A evolution. *Sci. Rep.*, **3**, (2013).

[419] K Eames and MJ Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Nat. Acad. Sci.*, **99**, 20, 13330-13335, (2002).

[420] S H Kaufmann, G Hussey and PH Lambert. New vaccines for tuberculosis. *The Lancet*, **375**, 9731, 2110-2119, (2010).

[421] SH Kaufmann. Tuberculosis vaccines: time to think about the next generation. *Sem. immun.* **25**, 2, 172-181, (2013).

[422] C Dye, GP Garnett, K Sleeman and BG Williams. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *The Lancet*, **352**, 1886-1891, (1998).

[423] P Dornelles, SL Bassanesi, ML Avancini, RL Targa, CA Jarczewski and P Rodrigues de Borba. Risk factors for recurrence of tuberculosis. *J Bras Pneumol*, **33**, 5, 572-578, (2007).

[424] T Pillay, M Khan, J Moodley, M Adhikari and H Coovadia. Perinatal tuberculosis and HIV-1: considerations for resource-limited settings. *The Lancet Infect. Dis.*, **4**, 3, 155-165, (2004).

[425] SY Del Valle, JM Hyman, HW Hethcote and SG Eubank. Mixing patterns between age groups in social networks. *Social Networks*, **29**, 539-554, (2007).

[426] E Miller, K Hoschler, P Hardelid, E Stanford, N Andrews and M Zambon. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *The Lancet*, **375**, 9720, 1100-1108, (2010).

[427] PJ Birrell et al. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proc. Nat. Acad. Sci.*, **108**, 45, 18238-18243, (2011).

[428] S. H. Kaufmann, G. Hussey, and P. H. Lambert *New vaccines for tuberculosis* The Lancet, 375(9731), 2110-2119. (2010)

[429] C Dye and P.E. Fine *A major event for new tuberculosis vaccines.* The Lancet, 381(9871), 972-974. (2013)

[430] I.M. Orme *Vaccine Development for Tuberculosis: Current Progress* Drugs 73, Issue 10, pp 1015-1024

[431] S Sontag. Illness as Metaphor and AIDS and Its Metaphors. Ed. Picador, (2001).

[432] A Attaran. An Immeasurable Crisis? A Criticism of the Millennium Development Goals and Why They Cannot Be Measured. *PLoS Med*, **2**, 10, e318, (2005).

[433] I Comas and S Gagneux. A role for systems epidemiology in tuberculosis research. *Trends in microbiology*, **19**(10), 492–500, 2011.

[434] T Schaffter, D Marbach, and D Floreano. GeneNetWeaver: *In silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**(16): 2263–70, (2011).

[435] F Tekaia, SV Gordon, T Garnier, R Brosch, BG Barrell, and ST Cole. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber. Lung Dis.* **79**(6), 329-342, (1999).

[436] RLV Skjot, I Brock, SM Arend, ME Munk, M Theisen, TH Ottenhoff and P Andersen. Epitope mapping of the immunodominant antigen TB10.4 and the two homologous proteins TB10.3 and TB12.9, which constitute a subfamily of the esat-6 gene family. *Infect. Immun.* **70**(10), 5446-5453, (2002).

[437] RN Gutenkunst, JJ Waterfall, FP Casey, KS Brown, CR Myers, and JP Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS comp. biol.*, **3**(10), e189, (2007).

[438] MP Berry, CM Graham, FW McNab et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**(7309), 973–977, (2010).

[439] LB Barreiro, L Tailleux, AA Pai, B Gicquel, JC Marioni, and Y Gilad. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Nat. Acad. Sci.* **109**(4), (2012).

[440] T Flutre, X Wen, J Pritchard, and M Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS genetics*, **9**(5), e1003486, (2013).

[441] JC Maranville, F Luca, AL Richards, X Wen, DB Witonsky, S Baxter, M Stephens and A Di Rienzo. Interactions between glucocorticoid treatment and cis-regulatory polymor- phisms contribute to cellular response phenotypes. *PLoS Genet.* **7**:1002162, (2011).

[442] X Cai, JA Bazerque and GB Giannakis. Inference of Gene Regulatory Networks with Sparse Structural Equation Models Exploiting Genetic Perturbations. *PLoS Comp. Biol.* **9**:5 e1003068, (2013).

[443] Z Dong, T Song and C Yuan. Inference of Gene Regulatory Networks from Genetic Perturbations with Linear Regression Model. *PLoS One* **8**:12 e83263, (2013).

[444] L Zhang and S Kim. Learning Gene Networks under SNP Perturbations Using eQTL Datasets. *PLoS Comp. Biol.* **10**:2 e1003420, (2014).

[445] I Tur, A Roberato and R Castelo. Mapping eQTL networks with mixed graphical models. *Arxiv* preprint arXiv:1402.4547v2 (2014).

[446] M Martcheva, BM Bolker and RD Holt. Vaccine-induced pathogen strain replacement: What are the mechanisms? *J R Soc Interface* **5**:3–13, (2008).

List of Figures

List of Tables