Universidad Zaragoza
1542

Master Proyect

Title:
Functional analysis of responses to stress in distant prokaryotes: comparison between *Mycobacterium tuberculosis* and *Escherichia coli*

Author

Laura Solanas Laguna

Director

Yamir Moreno Vega

Supervisor
Joaquín Sanz Remón

Faculty: Faculty of science
Year: 2018

Index

Abstract

This project combines life -i.e., biological- sciences methodologies with physical and computational analyses of protein expression for two differentiated microorganisms with a completely different lifestyles: *E.coli*, a well-known bacteria, and *M.tuberculosis*, a deathly human pathogen. In other to do that, we build two folding change multilayer networks of protein expression and analyze them. The multilayer networks have six layers which are equivalent to six stress conditions: acid, cell damage wall, hypoxia, ion deprivation, oxydative stress and starvation. To do the analysis and comparison between the networks corresponding to the two bacteria, we employed several tools. Regarding bioinformatics: GEO, metasoft; softwares as R-studio, ClueGO; statistical measures like strength, overlap and partition coefficient and statistical tests such as the Mann-Whitney and Peacock tests. Our results show that the differences in lifestyles are captured by the network approach and the proposed metrics. This work could open the path to obtain further insights about protein-protein interactions and relevant challenges such as protein function determination.

Keywords

E.coli, M.tuberculosis, lifestyle, network, multilayer, stress, acid, cell damage wall, hypoxia, ion deprivation, oxydative stress, starvation, strength, overlap, partition coefficient.

1.Introduction

One might say: “the eternal mystery of the world is its comprehensibility” (Einstein A. Physics and Reality,1936), however, we see how a scientist’s goal is always to try getting that comprehensibility. So, what science means is the first issue I would like to reflect on. If you just look up the word in the dictionary, you find something like “ the intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation an experiment”[1], you can also read “a branch of such knowledge, e.g. biology, chemistry, physics, etc”[2]. The concept of science was has been used for a long time. Etymologically, it comes from Latin “scientia” understood as “knowledge, expertness”, from “scire” meaning “to know”, probably originally “to separate one thing from another, to distinguish” related to “scindere” “to cut, divide” [3].

Nowadays, science advances quickly. The growing complexity of scientific studies requires ever-increasing cross-disciplinarity when it comes to particular methodologies used in the quest to reach the goals of these studies [4]. Interdisciplinarity is considered the best way to face practical research topics since synergy between traditional disciplines has proved very fruitful [5]. So, remarkable enough, this project is a combination of biological life studies and physical and computational analysis. The idea is to study two multiplex networks of protein expression for two differentiated microorganisms and how those proteins interact in the organism, when it is subject to different stresses that are equivalent to life conditions. To this end, several statistical methods are employed.

Therefore in this introduction, we first make a short exposition about *Escherichia coli* (*E.coli*) and *Mycobacterium tuberculosis* (*M.tuberculosis*) and compare them to show the differences between both lifestyles: from a facultative anaerobic commensal to highly adaptation to a specific environmental niche like the case of *M.tuberculosis*, which has the human as exclusive reservoir. Finally, we also present a brief summary of what a network is, as well as, multiplex networks and biological ones.

1.1 Biological microorganisms

In biological taxonomy, according to the Woese system, introduced in 1990, the tree of life consists of three domains: Archaea, Bacteria, and Eukarya[6]. Considering the organisms of interest, both *E.coli* and *M.tuberculosis* belong to Archea domain. A comparation of taxonomy classification is assembled in Table_1 to visualize easily how different they are.

Table_1. Taxonomy *E.coli* and *M.tuberculosis* including phylum, class, order, suborder, family and genus[37].

Microorganism	*E.coli*	*M.tuberculosis*
Phylum	Proteobacteria	Actinobacteria
Class	Gammaproteobacteria	Actinobacteria
Order	Enterobacteriales	Actinomycetales
Suborder	-	Corynebacterine
Family	Enterobacteriaceae	Mycobacteriacea
Genus	*Escherichia*	*Mycobacterium*

1.1.2 *E.coli*

E.coli is one of the best characterised organisms and has served as a model to study many aspects of bacterial physiology and genetics of fundamentals and applied interest [7]. The resulting knowledge and molecular methods for investigating and manipulating its biology have since led to *E. coli*'s prominence in academic and commercial genetic engineering, pharmaceutical production, and experimental microbial evolution, not to mention the biotechnology industry, which contributed 500 billion dolars to the global economy in 2011[8].

The proteobacteria belong to Enterobacteriaceae family and *Escherichia* genus was first isolated by Theodor Escherich from a human stool sample in 1886. It was called Bacterium coli commune and some years later *Escherichica coli* name was established[8-10]. There are multiple strains of this kind of enterobacteria that have adapted to diverse environmental conditions and lifestyles. While the typical *E.coli* genome constains roughtly 4800 genes, only approximately 2000 are shared by every *E.coli* strain. (See Figure_1) The genomic plasticity of various *E.coli* isolates provides *E.coli* the ability to proliferate and survive in an array of environments. The major niche of *E.coli* is the lower intestine tract of mammals, birds, and reptiles [11]. It also can be found in soil, water and food [12]. Some of them can be pathogenic, associated to human enteritis, urinary tract infection and septicaemia or diarrhea in pet and farm animals[13]. However, most of the strains are not pathogenic as it is the case of *E.coli* K12, which we focus on.

E.coli K12 strain was first isolated in Stanford in 1922 [11]. It is a Gram-negative, rod-shared bacteria belonging to the K serogroup of *E.coli*. It lives as a harmless inhabitant on the human large intestine and is widely used in medical and genetic research[8]. Physiologically, it is a facultative anaerobe and despite that it can not grow at extremes temperature or pH, its metabolic flexibility allows it to adapt to hard external conditions [14].

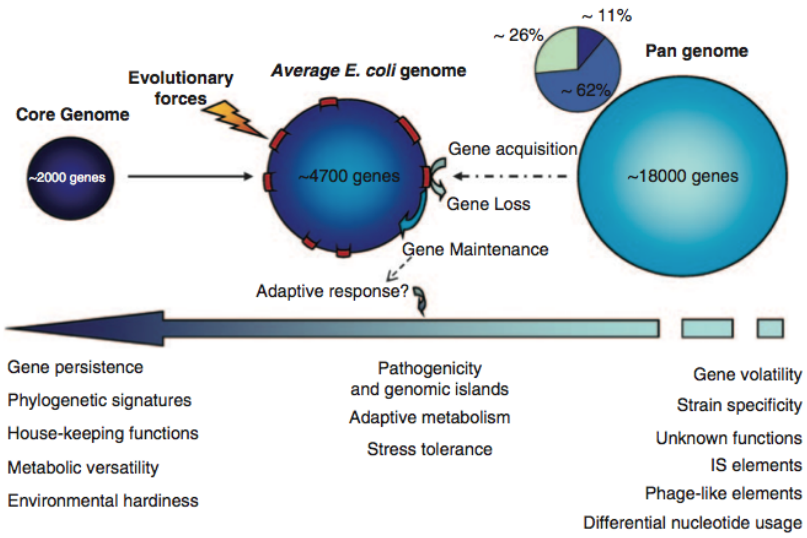

Figure_1. By Jan Dirk et al., 2011. The average *E. coli* genome is shaped by a multitude of evolutionary forces derived from its primary (host) and secondary habitats, in which both biotic (predators, competitors, cheaters, host defense mechanisms) and abiotic (pH, temperature, UV, mineral depletion and so on) pressures are present. *E. coli* strains possess a core of about 2000 genes, which equip them with a versatile metabolism. The *E. coli* pan genome consists of about 18000 genes, of which 11% belong to the core (dark blue), a large portion (62%, blue) is composed of so-called 'persistent' genes, and 26% can be considered as 'volatile' genes (pale blue) (Touchon et al., 2009). Events of gene acquisition and loss are consistently linked to insertion/deletion hotspots (red), and cooperatively shape the *E. coli* genome with selectively maintained core/persistent genes. These events may result in the evolution of gene clusters defining specific *E. coli* phenotypes [15].

1.1.2 *M. tuberculosis*

We focus now in *M.tuberculosis*, a microorganism detected by Robert Koch in 1882 [16]. This actinobacteria belongs to Mycobacteriaceae family and *Mycobacterium* genus (see in Table_1), and, furthermore, it is the causal pathogen of tuberculosis. Every minute, 3 people in the world die of this disease. With more than 8 million new cases of active disease and nearly 1.5 million deaths annually, tuberculosis is a global health emergency [17].

Regarding its cycle of life, the micobacteria concentrates on its unique reservoir: humans. The pathogen is transmitted by inhalation, enters in the alveolar space of the lungs and encounters the resident alveolar macrophages. If the defense fails, the bacteria invades the lungs interstitial tissue. Either dendritic cells or inflammatory monocytes transports *M.tuberculosis* to pulmonary lymph nodes for T cell priming. This event leads to the recruitment of immune cells, including T cells and B cells, and to the lung parenchyma to form a granuloma. Until now, the process described is called latent infection. Then, it can evolve to active disease. The bacteria replicate within the growing granuloma. If the number of bacteria is too high, the granuloma can not

support them and the pathogens disseminate to other organs [16]. At this point, there are symptoms such fever, fatigue, hemoptysis, coughing up blood in advanced disease [16]. (See in Figure_2).

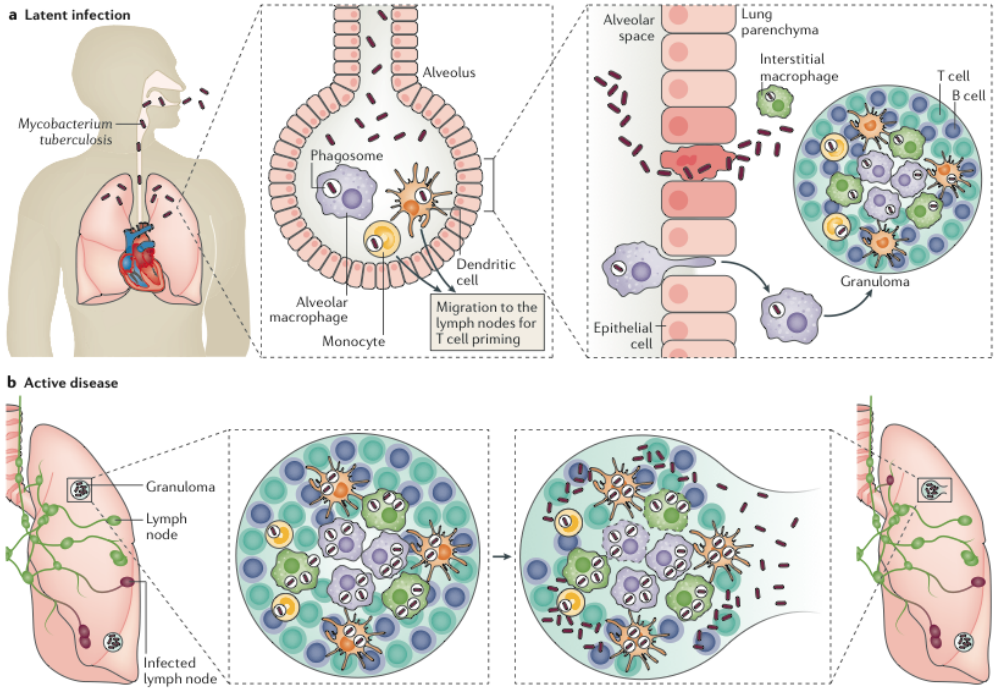

Figure_2. *Mycobacterium tuberculosis* and disease by Madhukar Pai et al. (Nature, 2016)[16]. a) Infection begins when *M.tuberculosis* enters in lungs via inhalation, reaches the alveolar space and encounters the resident alveolar macrophages. If this line of defense fails to eleminate the pathogen, it invades the lung interstitial tissue, either by the bacteria directly infecting the alveolar epithelium or the infected alveolar macrophages migrating to the lung parenchyma. Subsequently, either dendritic cells or inflammatory monocytes transport *M.tuberculosis* to pulmonary lymph nodes for T cell priming. This event leads to the recruitment of immune cells, including T cells and B cells, to the lungs parenchyma to form a granuloma. b) The bacteria replicate within the growing granuloma. If the bacterial load becomes too great, the granuloma will fail to contain the infection and bacteria will disseminate eventually to other organs. At this phase, the bacteria can enter the bloodstream or re-enter respiratory tract to released- the infected host is now infectious, symptomatic and is said to have active TB disease.

1.1.3 Differences between bacterial lifestyles

As we have seen before, both microorganisms are living inside a host and they also have a transit period outside the host for finding a new or the next host. On the one hand, *E.coli* is a mutualist and *M.tuberculosis* is an obligate pathogen. *E.coli* secures food and a comfortable enviroment and benefits its host, for instance, producing vitamin K and vitamin B12, both of which are required by mammalian hosts [8]. It lives mostly in gut and it can be excreted in fecal matter, but it also stands out for its high

survival in several niches. *E.coli* can also live in soil, manure and water [15]. On the other hand, *M.tuberculosis* has no known environmental reservoir; humans are its only known reservoir [16]. Its entire life cycle is determined by the context of human infection. It is transmited by aerosol and competes for nutrients with the host, disrupting its metabolic pathways and interacting with its immune system [18].

Therefore, the main difference can be stablished in the fact that *E.coli* is a generalist, in the sense that its metabolism is versatile and adapts to the environment. On the contrary, *M.tuberculosis* is a specialist as the pathogenic bacteria that it is, which means its metabolism causes the human disease.

1.2 Structure and function of complex networks

Biological and chemical systems, neural networks, social interacting species, the Internet and the World Wide Web, are examples of systems composed by a large number of hightly interconnected dynamical units [19]. This project is based on the analysis of two networks of two microorganisms. To understand how they work and what we have done, in what follows we comment the main networks features and their basic concepts.

As regards to the structure of complex networks, some general concepts appear below. A network can be represented as a graph, in accordance with Graph theory. A graph consists of two sets N, N ≡ {n1, n2,...,nN }, and L, L ≡ {l1, l2,...,lK}. N is the set of nodes and L that of the links or edges, which allow the connection between nodes (See Figure_3). Then, the features of the graph define which graph we are dealing with: undirected/directed, weighted/unweighted depending on the kind of information attached to the edges [19].

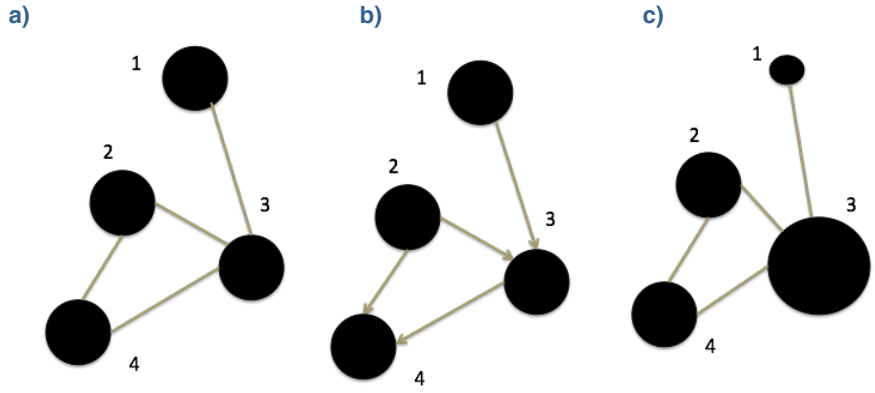

Figure_3. Illustrations of a graphic composed by N=4 nodes (black circles) and L=4 links (light connections). a) Undirected and unweighted. The set of nodes is P={1,2,3,4,5} and the edge or link set is E={{1,3},{2,3},{2,4},{3,4}}. b) Directed and unweight (P={1,2,3,4,5}, E={{1,3},{2,3},{2,4},{3,4}}).
c) Undirected and weight(P={1,2,3,4,5}, E={{1,3},{2,3},{2,4},{3,4}})

Some other interesting concepts are the following: the degree k_i of a node is related to the number of links to other nodes and the degree distribution, $P(k)$, is the probability that a node chosen uniformly at random has degree k or, equivalently, as the fraction of nodes in the graph having degree k [19]. Additionally, we note that real networks are often correlated in the sense that the probability that a node of degree k is connected to another one of degre k', depends on k', i.e., $P(k'|k)$ is not only a function of k [19].

If we focus now in distances, other features of the network can be mentioned as the shortest path length, the diameter, the closeness and the betweenness. The closeness is the inverse of the average distance from any node to all other nodes, while the betweenness measures node centrality, that is, the importance on a particular node in a network [19].

Clustering is also a network property which quantifies the number of triangles, i.e., three nodes that are conneted between them. A motif M is a pattern of interconnections occurring in a graph G at a number significantly higher than in randomized versions of the graph, (i.e. in graphs with the same number of nodes, links and degree distribution as the original one, but where the links are distributed at random). In addition, the community structure is an interesting concept as it characterizes how likely it is that there are groups of nodes that have a high density of edges within them, while having a lesser number of edges with other groups [20].

Regarding the topology of real networks, the development of data analysis tools has allowed to study a large variety of systems, reveling the fact that despite the existence of different kinds of real networks, they share roughly the same topological properties, for instance, relatively small characteristic path lengths, high clustering, fat tailed shapes in the degree distributions, degree correlations, and the presence of motifs and community structures [19].

On the other hand, while firstly networks were considered homogeneous, that is, topologically equivalent to random graphs or a regular lattice, it has been shown that real networks mostly show a scale-free degree distribution $P(k) \sim Ak^{-\gamma}$, with exponents varying in the range $2<\gamma<3$. The average degree $<k>$ in such networks is therefore well defined and bounded, while the variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$ is dominated by the second moment of the distribution that diverges with the upper integration limit k_{max} as

$$<k^2> = \int_{k_{min}}^{k_{max}} k^2 P(k) \sim k_{max}^{3-\gamma} \quad (1)$$

They are scale-free networks because power-laws have the property of having the same functional form at all scales [19]. See Figure_4.

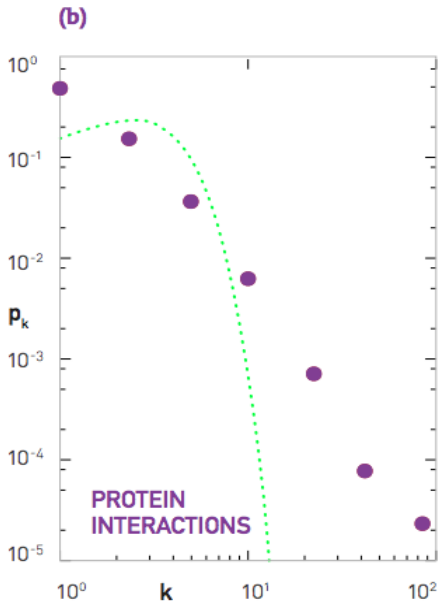

Figure_4. Adaptation [21]. Degree distribution of a Protein-protein interaction network which is Scale-free. The green dotted line shows the Poisson distribution with the same k as the real network, illustrating that the random network model cannot account for the observed $P(k)$.

1.2.1 Multiplex networks

A particular class of networks are multiplex networks, networks in which each node appears in a set of different layers, and each layer describes all the edges of a given type [22]. It is a better description of complex systems taking more aspects and conditions into consideration. Until now, the most frecuently approach to network description of complex systems consists of studying the graphs resulting from the aggregation of all the links observed between a certain set of elementary units. However, such aggregation procedure might discard important information about the structure and function of the original system, since in many cases the basic constituents of a system might be connected through a variety of relationships which differ for relevance and meaning [22]. That is the reason why in this project this kind of network has been used and studied. In particular, we have applied this framework to biological networks, which are explained in the next section.

1.2.2 Biological networks

Several elements at multiple scales in biology can be studied: cells, cellular components, metabolites, proteins, lipids, genes,... All of them are connected functionally in a complex way and complex networks can be used to provide insights into those relations [23].

In recent years, high-throughput experiments such as microarrays or yeast-two hybrids screens have produced large amount of information, which can be represented in

networks of interacting molecules. Lots of interacting molecules such as proteins and genes are treated and give rise to/generate biological networks, that have the main network properties commented before. So, most common types of biological networks are: protein-protein interaction networks, metabolic networks, genetic interaction networks, genes/transcriptional regulatory networks and cell signalling networks [24]. These networks have been studied since several years go and for instance, Metabolic networks are scale free (Jeong et al. Nature 2000; Wagner & Fell Proc. R. Soc. Lond. 2001) that explains how most metabolic substrates participate in only one or two reactions and there are few hubs. Other examples include Genetic regulatory networks (Featherstone & Brodie Bioessays 2002; Agrawal Phys. Rev. Lett. 2002), in which nodes are genes and links are derived from expression data; and finally Protein domain networks (Wuchty Mol. Biol. Evol. 2001; Apic et al. Bioinformatics 2001), where the network is constructed based on protein domain interactions.

1.2.2.1 Protein-protein interaction networks

In this work, we focus on protein-protein interaction (PPI) networks. They are based on protein-protein interactions which are essetial to almost every process in a cell. Thus, understanding PPIs is crucial for understanding cell physiology in different states. It is also useful for developing new drugs by targetting specific elements of PPIs. Strictly speaking, PPI networks are mathematical representations of the physical contacts between proteins in the cell. And these contacts are specific, occur between defined binding regions in the proteins and have a particular biological meaning. In addition, PPI information can represent both transient and stable interactions and knowledge of PPIs can be adopted to assign putative roles to uncharacterised proteins, to add details about the steps within a signalling pathway or to characterise the relationships between proteins that form multi-molecular complexes [24].

2. Hypothesis and Objectives

The radical differences between lifestyles of the two bacteria: *Escherichia coli* and *Mycobacterium tuberculosis* should reflect in the way that stress-responding genes behave in responses to the different stresses. *M.tuberculosis* sees everything together, always in the phagosome, so, immunogenic proteins should response to many different stresses. On the contrary, in *E.coli*, our expectation is that we should see a richer dynamics, with a larger fraction of genes being turned on only upon specific stresses. Aimed at confirming this hypothesis, i.e., that the PPIs networks of the two organisms should be different thus reflecting their diverse lifestyles, two multilayer networks have been built from prior PPI networks. In these networks, each layer is associate to a kind of stress.

Thus, our objectives are:

1. To build the multilayer networks of *E. Coli* and *M. tuberculosis*
2. To compare both networks using network metrics

3. To extract orthologous information for both networks and compare them.
4. To study *E. Coli* antigens in relation to the network built.
5. To compare both representations in regard to genes function enrichment.

We next discuss the materials and methods used. In the next schematic figure, we summarize the work flow

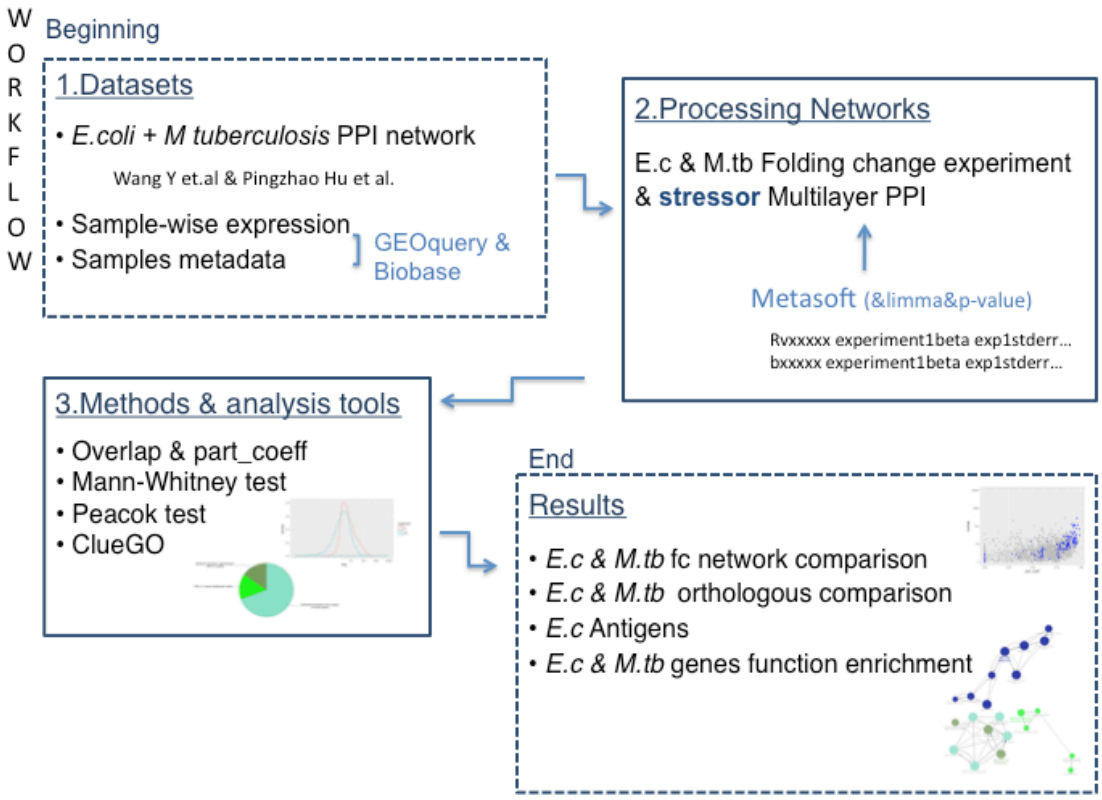

Figure_6. The work flow of the project. Firstly, we build *E.coli* and *M.tuberculosis* PPI folding change multilayer networks. To do that, we use three datasets: *E.coli* and *M.tuberculosis* PPI netowrks, sample-wise expression and samples metadata (these last two using GEOquery and Biobase). Secondly, metasoft provides multilayer network data (p-values, standard deviations…) by stress considering each stress a layer. At this point, we can analyse the multiplex measuring overlap and partition coefficient. We run also statistical test as Mann-Whitney and Peacock tests and use different sofwares like ClueGO. Finally we obtained and discute the results: *E.coli* and *M.tuberculosis* folding change multiplex network comparison, *E.coli* and *M.tuberculisis* orthologous comparison, *E.coli* antigens study and *E.coli* and *M.tuberculosis* genes functions enrichment.

3. MATERIALS

3.1 Multilayer networks: Folding Change based

The main objects of study of this project are both *E.coli* and *M.tuberculosis* folding change overexpression multilayer networks. They are composed by six layers and each layer represents a stress: acid, cell wall damage, hypoxia, ion deprivation and oxidative stress and starvation. They supposed to simulate different environmental conditions in which a microorganism has to live.

These networks have been built from previous networks, which are described in the articles: *Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv* by Wang Y et al. and *Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins* by Pingzhao Hu et al. From these researches, connections between proteins (links) are defined, then, coefficients of expression are needed to determine how important links are. (See the explanation of the procedure in 4.1. Network construction: Meta-analysis and GEO).

On the one hand, almost the entire ORFeome of *M.tuberculosis* was cloned. The pathogen's genome encodes about 4000 ORFs and around one third has unknown functions. The protein-protein interaction (PPI) network involves 2907 proteins linked via 8042 interactions.

On the other hand, the *E.coli* genome consists of 4339 protein-coding genes. 2667 of them have been purified and overexpressed, out of which 2337 were pairs of proteins and 330 were alone. The total number of identified protein-protein interactions were 16050 and after some filtration 11511.

The last aspect to mention is the fact that they are folding change networks, which means that they use Fold Change (FC) calculated as a ratio of averages from control and test sample values to select the differentially expressed genes in a microarray dataset with these two biological conditions [25]. Then, genes which appear over-expressed or down-expressed in the networks are those which exceed a threshold or cutoff that marks the level of change.

3.2. SOFTWARE_1: R-studio

R-studio is a free and open source data analysis software. Our scripts are developed using it to analyse the data and graph properties. Several packages have also been used, such as ggplot2 and Peacock.test, which can be loaded simply by

```
> library (ggplot2)
> library (Peacock.test)
```

3.3. SOFTWARE_2: Cluego plug-in (Cytoscape)

Cluego is a tool that improves biological interpretation of large lists of genes. It integrates Gene Ontology (GO) terms as well as KEGG/BioCarta pathways and creates a functionally organized GO/pathway term network. It can analyze one or compare two lists of genes and visualizes functionally grouped terms [26].

3.4. SET_1: Orthologous pairs *M.tuberculosis-E.coli*. (Along with dictionaries).

The set of genes which sequences have a common ancestor and have split due to speciation for *M.tb* and *E.coli*. They are collected in Annex Table_1, each column represent the id of those expressed genes for both microorganisms. These data come from *Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks* by M.Madam Baby, Sarah A.Teichmman and L.Aravind. As part of their analysis, they provide predictions of transcriptional regulatory interactions and transcription factors for the 175 prokaryotic genomes studied. The initial set is composed by 326 orthologous. (See Figure_7).

The work done with these data includes the following steps. Before being employed, the initial genes are treated. Firsly, their nomenclature is changed: Gene IDs are translated from GenBank Indentifiers (GI) to Ordered Locus Names (OLN) in both M.tb and *E.coli*. OLN are used to sequentially assign an identifier to each predicted gene or a completely sequenced genome or chromosome based on a prefix representing the organism followed by a number, which usually represents the sequencial ordering of genes on the chromosome (definition by Uniprot, the Universal Protein Resource). See a few examples below:

GI *M.tb*	OLN *M.tb*	GI *E.coli*	OLN *E.coli*
15607143	Rv0001	16131570	b3702
15607144	Rv0002	16131569	b3701
15607145	Rv0003	16131568	b3700

GI-OLN conversion was made by Uniprot's ID mapping tool which is useful for converting Uniprot IDs to NCBI Gene IDs and vice-versa [27]. After this step, there remain 274 *E.coli* and 211 *M.tuberculosis* genes. Out of these, only 85 are present in the *E.coli* folding change overexpression network and 94 in the *M.tuberculosis* one, in turn, 76 have overlap and participation coefficients (see below for the definition of these quantities) and 18 have overlap equal to zero and Not Available number (⟨NA⟩)as participation coefficient. These new sets are called pull genes. Finally, by filtering only orthologous pairs which share a function, 43 were selected and those are named Orthologous pairs or pair set.

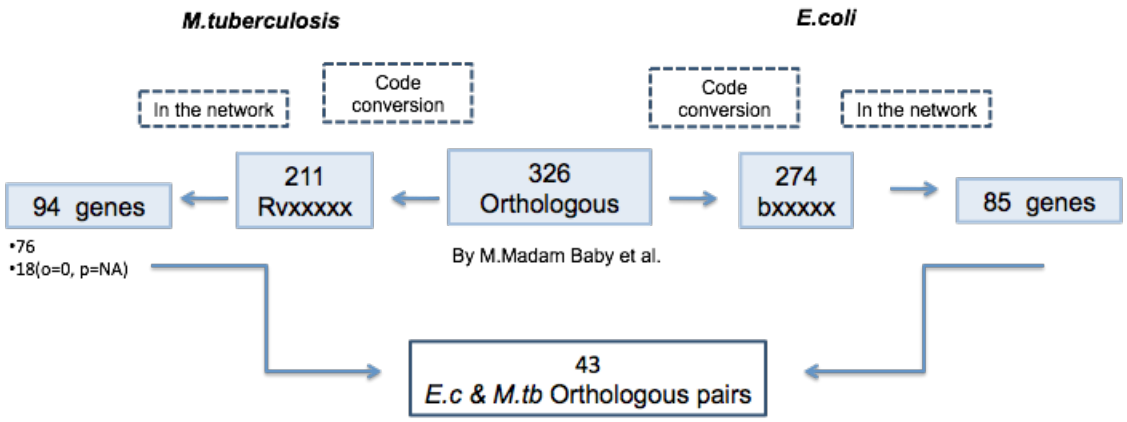

Figure_7. Procedure of orthologous selection to get final orthologous set used in orthologous analyisis (See in 5.2.Compartion *E.coli* and *M.tuberculosis* Orthologous genes: pull and pairs). 326 orthologous initial set from M.Madam Baby et al. After name transformation (from GI to OLN), 274 *E.coli* and 211 and 274 *M.tuberculosis* orthologous are obtained, of which 85 and 94 (14 of them cero overlap and <NA> partition coefficient) are part of *E.coli* and *M.tuberculosis* folding change overexpression multilayer networks respectively. They are named pull of orthologous. In turn, only 43 are pairs between them. These are the pair set.

Orthologs are studied in Section 5.2. Comparison *E.coli* and *M.tuberculosis* Orthologous genes: pull and pairs.

3.5. SET_2: Antigens of *E.coli*.

The set of antigens of *E.coli* K12 are grouped together in Annex Table_2. Antigens are foreign substances that trigger the production of antibodies when introduced into the body [28].The initial antigen set was composed of 40, but only 14 have been used because not all of them appear in the initial PPI from which the multilayer network was built. In the first place, the name was changed to OLN (24 antigens). The conversion was made using the PI network data table (S6) in *Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins* by Pingzhao Huand and the suplementary table “Protein-protein interaction data table” reported in *Large-scale identification of protein–protein interaction of Escherichia coli K-12* by M.Arifuzzaman et al. Finally, only 14 are part of the E.coli folding change overexpression multilayer network. (See Figure_8). This set is employed in 5.3. *E.coli* antigen analysis.

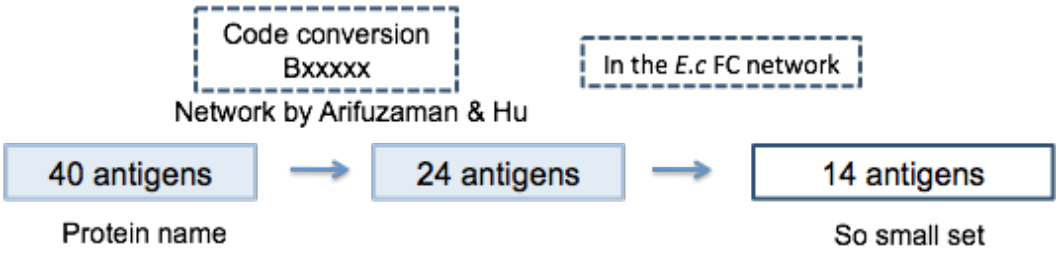

Figure_8. Procedure of the obtention of antigen set. 40 antigen made up the initial set. After name conversion in OLN, there remain 24 antigens, lastly, the final set comprises 14 antigens.

4. METHODS

4.1. Network construction: Meta-analysis and GEO

A meta-analysis is a set of techniques used "to combine the results of a number of different reports into one report to create a single, more precise estimate of an effect" (Ferrer, 1998). The aims of meta-analysis are "to increase statistical power; to deal with controversy when individual studies disagree; to improve estimates of size of effect, and to answer new questions not previously posed in component studies" (Hunter and Schmidt, 1990). There are several advantages to meta-analysis. It allows investigators to pool data from many trials that are too small by themselves to allow for secure conclusions [29].

Therefore, the folding change multilayer networks had been built by a meta-analytical process, which is explained some lines below. At the same time, it should be remarked that this work has been done in parallel with another named: "*A Multiplex approach to study the responses of Mycobacterium tuberculosis to in vitro induced stress, and its relation with epitope conservation*" by Miguel Baéz Martín. Miguel works with *M.tuberculosis* absolute expression multilayer network, being the main interest *M.tuberculosis* as a pathogen microorganism, while here, both multiplexes (*E.coli* and *M.tuberculosis*) are compared, focusing on ways of living, behaviours and evolution. However, these two researches have various common points such as some statistical analysis and the bioinformatic methodology, which were developed together and shared. They are explained later.

It should also be noted that these two studies are a consequence of a previous one by Fernando Cid who wrote "*Characterization of context-specific networks of protein-protein interaction in Mycobacterim tuberculosis*". Nonetheless, the latter had some statistical shortcomings and goals were not wide enough. Hence, these two new projects have broader objectives and seek to overcome the mentioned shortcomings. In what follows, we describe the process that allows building the networks and the way of doing the analysis. The procedure was developed in a simple way by Miguel and I, and then, corrected by Joaquín Sanz, Yamir Moreno and Sergio Arregui, due to the heterogenity of the inputs data and the difficulty of some scripts. (See Annex Code_3).

Let's first introduce the procedure used in *Characterization of context-specific networks of protein-protien interaction in M.tuberculosis* by Cid F. A large-scale meta-analysis of NCBI Gene Expression Omnibus (GEO) data was done, with the aim of knowing the expression levels for each gene. The GEO serves as a public repository for a wide range of high-throughput experimental data. These data include single and dual channel microarray-based experiments measuring mRNA, genomic DNA, and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), mass spectrometry proteomic data, and high-throughput sequencing data[30].

So, regarding experiment-wise, the data was extracted from GEO. In this database, samples (GSMxxx) are hierarchically grouped into arrays and experiments and stressors. A GEO serie (GSExxx) normally corresponds to one or more experiments, each of which corresponds to more than one sample or array [30]. For each array k of a given experiment, two important measures are obtained for gene-pair i-j in experiment k:

$$FC_{i-j}^{k} = FC_{i}^{k} - FC_{j}^{k} \quad \& \quad {E_{stim}}_{i-j}^{k} = {E_{stim}}_{i}^{k} + {E_{stim}}_{j}^{k} \tag{2}$$

Then, grouping all gene-pair-wise arrays within an experiment using the LIMMA package in R, we built model-based estimates and standard errors (along with the p-value and false discovery rates (fdrs)) for each link:

$$<FC>_{i-j}^{k}, SE_{FC(i-j)}^{k} \;\&\; <E_{stim}>_{i-j}^{k}, SE_{E_{stim}(i-j)}^{K} \tag{3}$$

Finally, links are grouped within experiments corresponding to the same broad stress type into a coarse-grained multiplex, see Figure_9.

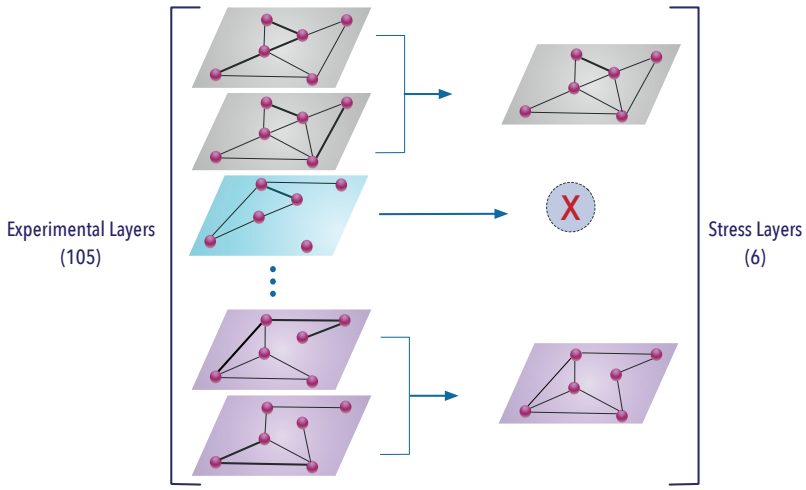

Figure_9. Multilayer per stress. From layers that can be each experiment from GEO, a classification of those per stress gives multilayer per stress (six layers, six stresses).

In short, the final pipeline was composed of three input tables: sample-wise expression (1) where columns were samples and rows were genes, a metadata table (2) in which rows collected samples, named the same way they were in the expression table, and, columns were sample attributes such as description, strain, experiment ID and stressor ID. The third input table was the PPI network (3) with the gene-pairs to filter. (See Figure_10).

Finally, the experiment-wise multi-layer network was built by using LIMMA and p.adjust. Limma allows to get $\langle FC \rangle_i$ values, sd's and t and p-values per experiment and p.adjust limits the false discovery rates (fdrs): 5% threshold.

The following Figure_10 summarizes how the two multilayer networks (experiment-wise multilayer PPI and a stressor-wise multilayer PPI -which depends on the six stresses previously mentioned-) were built.

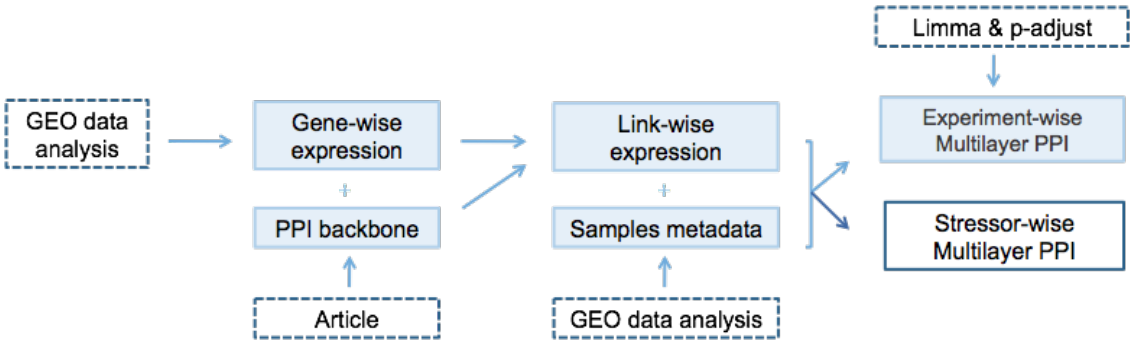

Figure_10. Steps to build multiplex networks. Firstly, expression data are retreived from GEO data analysis: both gene-wise expression and samples metadata. Apart from that, PPI network interactions are retreived from the bibliography. The gene-wise expression and PPI backbone together give link-wise expression summed over samples metadata generate both Experiment-wise Multilayer PPI and Stressor-wise Multilayer PPI, depending on which layers are represented: experiments or stress, respectively. Networks are built by limma and p-adjust which allow to get $\langle FC \rangle_i$ values, sd's and t and p-values.

Regarding the stressor-wise network, in which we are interested, we wanted to get a global estimate or error that produces a p-value from a series of experiments that are in turn characterized by an average and a significance level. In the average-of-averages situation, we need a way to treat, in the same way, cases where the within-experiment uncertainty is larger or smaller to the one observed between-experiments. Figure_11 illustrates two possible scenarios: in one of them the previous approach (Cid. F.) would work, but in the other, it would fail. Therefore, we need to find a solution whatever the scenario is or, in other words, we need to treat in the same way cases in which the within-experiment uncertainty is larger/smaller to the one observed between-

experiments. To this end, we chose as an alternative the random effects model. The idea is based on the sum of two aspects. Firstly, weight experiments according to within-experiment variance, so that less noisy experiments were assigned more weight.

$$<FC>^{s}_{i\leftrightarrow j} = \frac{\sum w_k <FC>^{k\in s}_{i\leftrightarrow k}}{\sum w_k} \quad where \quad w_k = \frac{1}{\left(\sigma^{h}_{FC(i\leftrightarrow j)}\right)^2}$$

$$And\ also: \ = \left(\sigma^{k}_{FC(i\leftrightarrow j)}\right)^2 = \frac{1}{\sum w_k} \qquad (4\&5)$$

That's far better than before, but still inter-experiment variance is not accounted for. This is done by

$$w_k = \frac{1}{\left(\sigma^{k}_{FC_{(i\leftrightarrow j)}}\right)^2 + \hat{\tau}^2} \qquad (6)$$

where $\hat{\tau}^2$ is the so-called DerSimonian-Lard heterogeneity estimator (it is higher the larger the inter-experiment variance is):

If $(\sigma^{k}_{FC}\ (i\Leftrightarrow j)^2 >> \hat{\tau}^2$ then $w_k \approx 1/(\sigma^{k}_{FC}\ (i\Leftrightarrow j))$ and then everything gets the same as if there no inter-experiment Variance. However, if $(\sigma^{k}_{FC}\ (i\Leftrightarrow j)^2 << \hat{\tau}^2$ for all experiments, then $w_k \approx 1/\hat{\tau}^2$ for all, and $\langle FC\rangle^{S}_{i\Leftrightarrow j} = \langle\langle\langle FC\rangle^{k}_{i\Leftrightarrow j}\rangle\rangle^{k\in l}_{i\Leftrightarrow j}$ and $\sigma^{s}_{FC}\ (i\Leftrightarrow j)^2 = \hat{\tau}^2$

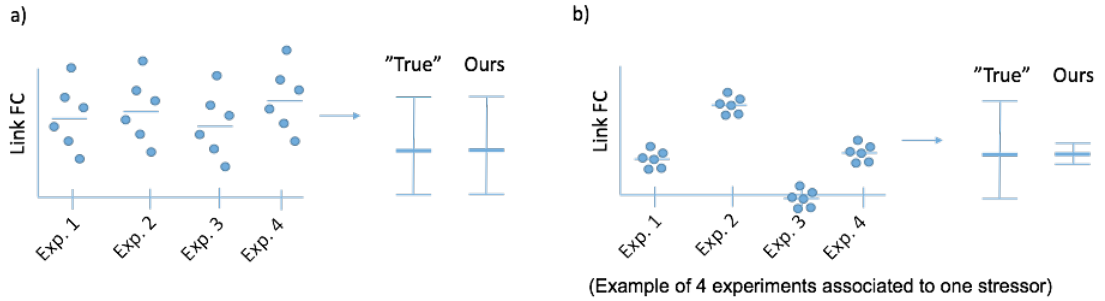

Figure_11. In the average of averages situation, when series of experiments has associated an estimate and a significance, two cases could be possible: a) The uncertainty is mainly driven by within-experiment variance (in this situation Cid.F research would work), b) The uncertainty is mainly driven by across-experiment variance (previous approach, by Cid.F, would fail).

The previous approach, due to DerSimonian-Lard (5), was however not used here for several reasons. Firstly, it was apparently more conservative than it should (one has a hard time to retrieve significant hits from it). Secondly, there is a newer and better method around: Metasoft.

Metasoft is a free, open-source meta-analysis software tool for genome-wide association study analysis, designed to perform a range of basic and advanced meta-analytic methods in an efficient manner [31]. Metasoft provides methods and estimates which are gathered in Table_3 and Table_4, such as Fixed Effect model (FE), Random Effects model (RE), Binary Effect model (BE) and M-values.

The format of the input file should has test gene-pairs as rows, the first column is gene-pairs IDs and the second and third columns are effect size (beta) and its standard error of experiment 1, the forth and fifth columns are effect size (beta) and its standar error of experiment 2, and so on. For missing beta and its standard error we write “NA”. (See Table_2).

Table_2. Example format Metasolft input file, *M.tuberculosis*, layer acid.

Gene pairs IDs	Experiment 1		Experiment 2	
	Effect size	Standard error	Effect size	Standard error
Rv0001_Rv0058	-15538	0.9893248	-0.0775	0.98932482
Rv0001_Rv0262c	0.2786	0.7638243	-0.2720	0.76382434
Rv0001_Rv0301	-0.7821	0.7447254	-0.2783	0.74472539

Example Running Command:

```
java -jar Metasoft.jar -input example.txt
```

Table_3. Metasoft output file methods and estimates.

Column Num.	Column name	Description
1	RSID	SNP rsID
2	NUM_STUDY	Number of studies used in meta-analysis for the SNP
3	PVALUE_FE	FE P-value
4	BETA_FE	Estimated Beta under FE
5	STD_FE	Standard error of BETA_FE
6	PVALUE_RE	RE P-value
7	BETA_RE	Estimated Beta under RE
8	STD_RE	Standard error of BETA_RE
9	PVALUE_RE2	RE2 P-value
10	STAT1_RE2	RE2 statistic mean effect part
11	STAT2_RE2	RE2 statistic heterogeneity part
12	PVALUE_BE	BE P-value (“NA” if -binary_effects option is not used)
13	I_SQUARE	I-square heterogeneity statistic
14	Q	Cochran's Q statistic
15	PVALUE_Q	Cochran's Q statistic's p-value
16	TAU_SQUARE	Tau-square heterogeneity estimator of DerSimonian-Laird
17 … 17+NUM_STUDY-1	PVALUES_OF_STUDIES	P-values of each single studies
17+NUM_STUDY … 17+2*NUM_STUDY-1	MVALUES_OF_STUDIES	M-values of each single studies (“NA” if -mvalue option is not used)

Table_4. . Some rows of outfile in Metasolft for acid stress in *M.tuberculosis*. Columns are: Study, PValue FE, Beta FE, STD FE, PValue RE, Beta RE, STD RE, PValue RE2, STAT1 RE2, STAT2 RE2, PValue BE, I Square, Q, PValue Q, TAU Square

RSID	Study	PValue_FE	Beta_FE	STD_FE	PValue_RE	Beta_RE	STD_RE
Rv0001_Rv0058	2	0.243635	-0.81565	0.699558	0.269163	-0.81565	0.73815
Rv0001_Rv0262c	2	0.995125	0.00330	0.54010	0.995125	0.00330	0.54010
Rv0001_Rv0301	2	0.314014	-0.53020	0.52660	0.314014	-0.53020	0.52660

STD_RE	PValue_RE2	STAT1_RE2	STAT2_RE2	PValue_BE	I_Square	Q
0.73815	0.270396	135944	0.00000	NA	101830	111337
0.54010	0.997522	3,73	0.00000	NA	0.00000	0.25981
0.52660	0.344983	101372	0.00000	NA	0.00000	0.22882

PValue_Q	TAU_Square	Pvalues_studies (Tab delimitered)	Mvalues_studies (Tab_delimitered)		
0.291349	0.110967	0.116283	0.937561	NA	NA
0.610251	0.00000	0.715303	0.721764	NA	NA
0.632400	0.00000	0.293633	0.708631	NA	NA

At this point, it is worth remarking that a total of 4 multilayer networks were used in this work. There were two organisms, *E.coli* and *M.tuberculosis*, and each of those were studied in two ways getting multiplex per experiment-wise and a multiplex per stress-wise. In turn, two multiplexes have been built from each data classification: an absolute expression network and a folding change network, that are used to do the statistical analyses. In this work, only the folding change networks were analysed.

4.2. Analytical methods

4.2.1.Basic properties

For this analysis, we closely follow the work by F. Cid. Three metrics are computed in order to describe the topology of the networks: strength, overlap and partition coefficient. The strenght of gene i in layer k, $s_k(i)$, is the sum of all link weights in which that node participates in the layer k:

$$s_k(i) = \sum_{j'=1}^{genes} < FC >_k (i \leftrightarrow j) \quad (7)$$

where the sum over j' means that only statistically significant links (here, those with Z-score greater than 1.96, and therefore, present in the layer) are considered. Layer k is associated to a stress multiplex in the reduced consensus multiplex.

So, the strength is a local property of node i in layer k. Instead, the overlap $o(i)$ of a node is defined as the sum of all nodes' strengths over all layers of the multiplex:

$$o(i) = \sum_{k=1}^{layers} S_k\ (i) \tag{8}$$

Hence, the overlap defines a global property for each node.

Both parameters quantify the importance of each node in a specific layer or in the whole network, repectively. The higher these values are, the more important a gene is in this context, because it interacts more intensively with more genes. However, a gene could have a high overlap value but its contribution to one specific stress is not necessarily important if it does not have a high strength value in that layer as well. The dissonance of these two measures remarks the importance of a multiplex approach to analyse biological networks. In order to account for this fact, we measure participation coefficiente $p(i)$ for each node i. It is define as follows:

$$p(i) = \frac{M}{M-1}\left(1 - \sum_{k=1}^{layers}\left(\frac{S_k(i)}{o(i)}\right)^2\right) \tag{9}$$

where M is the number of layers in the multiplex. This magnitude is equal to 0 when the presence of a node in the multiplex is concentrated on a single layer, while it is 1 when it is homogeneously distributed among all layers. In our context, nodes of low p are associated to genes related with specific responses to a particular stress, while genes of high p correspond to generic-stress respondent genes, regardless of the precise nature of these stresses.

Since the overlap of a node represents its overall importance, the nodes of a multiplex could be classified by looking, at the same time, at their multiplex participation coefficient and at their overlap. By this analysis, nodes which are hubs on the whole network and other ones that are only hubs on a specific layer could be distinguished. These second ones are especially important in the sense that thery could serve as support of other protein function prediction methods, since if a node is crucial in one layer but have non-significant importance in the rest of them, it could be inferred that that gene function is related to the stress response that characterizes that layer.

4.2.2 Statistical analysis

4.2.2.1. Mann Whitney test

The Mann-Whitney test is used as an alternative to a t test when the data are not normally distributed. The test can detect differences in shape and spread as well as just differences in medians and differences in population medians are often accompanied by equally important differences in shape[36].

Mann-Whitney test:
In R studio, Wilcoxon Rank Sum and Signed Rank tests. Wilcox.test performs one- and two-sample Wilcoxon tests on vectors of data. The Script is:

```
setwd("~/Desktop/M.BiotecnologíaCuantitativa/TFM/PPI_MTB_meta_analysis")
ec_fc_over <- read.table("Inputs/ppi_interaction_pairs.txt", header = TRUE)
mtb_fc_over <- read.table("Inputs/ppi_interaction_pairs.txt", header = TRUE)

wilcox.test (ec_fc_over$overlap, mtb_fc_over$overlap, alternative="greater")
wilcox.test (mtb_fc_over$part_coeff, ec_fc_over$part_coeff, alternative="greater")
```

4.2.2.2 Permutation tests in Hypothesis contrast

Permutation tests are a class of non-parametric methods. They were pioneered by Fisher (1935) and Pitman(1938). Fisher demonstrated that the null hypothesis could be tested simply by observing, after permuting observations, how often the difference between means would exceed the difference found without permutation, and that for such test, no normality would be required. Pitman provided the first complete mathematical framework for permutation methods, although similar ideas, based on actually repeating an experiment many times with the experimental conditions being permuted, can be found even earlier. Permutation methods can provide exact control of false positives and allow the use of non-standard statistics, making only weak assumptions about the data [32].

4.2.2.3. Peacock test

This technique is available to test for consistency between the empirical distribution of data points on a plane and a hypothetical density law. Two statistical tests are available. The first is a two-dimensional version of the Kolmogorov-Smirnov test, for which the distribution of the test statistic is investigated using a Monte Carlo method. It is found in practice to be very nearly distribution-free, and empirical formulae for the confidence levels are given. Secondly, the method of power-spectrum analysis is extended to deal with cases in which the null hypothesis is not a uniform distribution. These methods are illustrated by application to the distribution of quasar candidates found on an objective-prism plate of the Virgo Cluster [33].

This test is applied by using R-studio. There is a package called ⟨Peacock test⟩ [34]. After running it, the function gives the Peacock statistics D. D is a generalization of the Kolmogorov-Smirnov statistics, which, in 1D tests, measures the largest difference between the cumulative probabiliy distributions being compared, $P(X,Y,\ldots)$ as the probability $x<X$, $y<Y$,.... However, we applied the 2D Peacock test, where D corresponds to recognizing that all four quadrants of the plane defined by $(x<X, y<Y)$, $(x<X, y>Y)$, $(x>X, y < Y)$ and $(x>X, y >Y)$ are equally valid areas for the definition of the cumulative probability distribution. In this case, the procedure adopted here is to consider each in turn, and adopt the largest of the four differences in empirical and theoretical cumulative distributions as the final statistic [33].

Once D is obtained, the next step is to transform it into a Z-score by multipliying by the effective sample size:

$$Z_n = n * D \quad n = \sqrt{\frac{n_1 * n_2}{(n_1 + n_1)}} \tag{10}$$

Then, Z_n is corrected for finite sample size to get Z_∞.

$$Z_\infty = \frac{Z_n}{1 - 0.53^n} \tag{11}$$

Finally, the P-value is obtained as follows:

$$p(Z > Z_\infty) = 2 * e^{-2(Z_\infty - 0.5)^2} \tag{12}$$

Thus, it is possible to get a Z-score following the prevoious steps. We also note that our first idea to get a P-value was to run the Peacock test a high number of times. This is possible following a bootstrap algorithm for *E.coli* and *M.tuberculosis* overlap and partition coefficient data. See the script which was programmed in Annex Code_2. Finally, this option was discarded because it was computationally really expensive compared with the alternative already explained.

This method is used in Analysis 1 to compare two distributions, both for *E.coli* and *M.tuberculosis* networks (the overlap and the participation coefficient data are the two dimensions, see 5.1. Comparison *E.coli* versus *M.tuberculosis*). In Analysis 2, the Peacock test evaluates orthologous distributions (5.2. Compartion *E.coli* and *M.tuberculosis* Orthologous genes: pull and pairs).

4.2.2.4. Gene ontology enrichments

Gene ontology enrichments analysis (see 5.4. Gene ontology enrichement analysis comparing both *E.coli* and *M.tuberculosis*) was done with ClueGO-cytoscape. Firstly, input files were created. For each layer, the set of genes which had strenght different to zero were selected. In total, for both microorganisms, there were 12 sets: acid, cell wall damage, hypoxia, ion deprivation, oxydative stress and starvation of 141 genes each one, being selected those of higher overlap.

In order to get coherent results some conditions where stablished in the software: pathway Biological Process-EBI-UniProt-GOA (type GO), using GO Term Fusion, showing only Pathways with pV 0.05; and GO Tree Interval: 3 min level - 6 max level, Go Term/Pathway Selection (%Genes): 5 Min genes - 5% Genes. Regarding statistical options, Enrichment (Right-sided hypergeometric test), Benjamine Hochberg and finally Selected Ontologies Reference Set.

There was a second analysis only for *M.tuberculosis* in which some conditions changed: no p-value limitation (significance level: pV ≤ 0.05) and Custom reference selected (*M.tuberculosis* background file which is formed by all the *M.tuberculosis* folding change overexpression multilayer network).

4.2.2.5. Papers data sources.

In this work, the decision to build multilayer *M.tuberculosis* PPI keeping in mind several stress conditions was relatively easy. As alsready noted, the work Global protein-protein interaction network in the human pathogen *Mycobacterium tuberculosis* H37Rv by Wang Y et al. (2010) was used as a reference. This was also studied in a parallel report: *Systems Biology of Mycobacterium tuberculosis. Immunogenicity, natural selection & gene expression in response to stress* by Miguel Báez Martín and previously used by Cid. F.

However, choosing papers reporting the PPI network for *E.coli* K12 was not so easy. Three achievable articles were found: Interaction network containing conserved and essential protein complexes in *Escherichia coli* by Gareth Burland et al. (Nature, 2005), *Large-scale identification of protein-protein interaction of Escherichia coli K-12 by Mohammad Arifuzzaman et al.* (Genome Research, 2006), and *Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins by Pingzhao Hu.* (PLOS Biology, 2009). They had different information related to gene protein expression and established interactions between nodes. The first dismissed article was the Nature 2015 obe, because it used *E.coli* as a vehicle to study essential protein complexes. Then, two PPI *E.coli* networks were completed but they did not have the same data. Mohammand Arifuzzaman et al. studied a set of 4339 proteins and only 2667 proteins were purified and overexpressed. In turn, 2337 were copurified with other proteins and 330 without anyone. Lastly, a total of 16050 protein-protein interactions were identified. On the other hand, Pingzhao Hu. et al. studied a total of

4225 proteins and interactions between 1757 of those (451 were orphans). They reported 5993 physical interactions, and 74446 functional interactions.

Finally, we selected the PLOS Biology because it had more methodological details, reported a larger number of interactions between nodes and it was a more recent publication.

5. RESULTS

Results of the folding change network *E.coli* and *M.tuberculosis* analysis are showed in this section. Four different analyses were done by the combination of materials and methods previously described: analysis 1 is a comparation between two folding change microorganisms completed networks, the second analysis shows the comparison of orthologous genes of *E.coli* and *M.tuberculosis*, the third analysis is focused on *E.coli* antigens and finally, the fourth analysis is based on gene ontolgy enrichment comparing *E.coli* and *M.tuberculosis*.

5.1. Comparison *E.coli* versus *M.tuberculosis*.

Several analyses were done to compare both microorganisms. Firstly, the proportion of genes within each layer was contrasted. Then, the statistical tests (Mann Whitney and Peacock test) were runned to see the relation between the overlap and the participation coefficient between bacterias.

As far as nodes of the multiplex networks are concerned, the number of genes per layer in each network as well as its proportion are shown in Tabla_5. The *E.coli* folding change multilayer network has a total of 1757 genes or nodes and 5993 interactions or links, while the *M.tuberculosis* folding change multilayer network has 2907 genes and 8042 interactions. So, we see here two networks that do not have the same size, which could affect subsequent statistical analyses as disussed later on. It is also worth remarking the value of the genes in the starvation layer. Samples and series of experiments found in GEO database for this stress were higher than in the others, this fact could also influence the results.

Tabla_5. Number of genes and proportion of them in each layer in both *E.coli* and *M.tuberculosis* folding change multilayer network for overexpressed and repressed situation. Data filtrated using significance level of 0.05 in both cases.

Layers	Nodes							
	Overexpressed				Repressed			
	E.coli		*M.tuberculosis*		*E.coli*		*M.tuberculosis*	
	Genes	Proportion	Genes	Prop.	Genes	Prop.	Genes	Prop.
Acid	141	0,08	328	0,112	205	0,116	437	0,044
Cell wall damage	211	0,12	1388	0,477	493	0,28	1458	0,122
Hypoxia	150	0,085	1834	0,63	233	0,132	2384	0,065
Ion deprivation	345	0,196	586	0,201	581	0,33	1637	0,214
Oxydative stress	268	0,152	1549	0,532	76	0,043	2095	0,017
Starvation	1304	0,742	1157	0,398	186	0,105	1621	0,037

To get a better visualization of the components of the networks, some bar diagrams were built. On the one hand, there are four histograms which represent the proportion of genes in each layer for both networks in cases of overexpression and repression. On the other hand, there are other four plots which are the proportion of links in each layer for each network in the situation of overexpression and repression. We see that, as it was said before, in the overexpression case, the gene proportion shows some differences while for the repressed scenario this quantity is more similar.

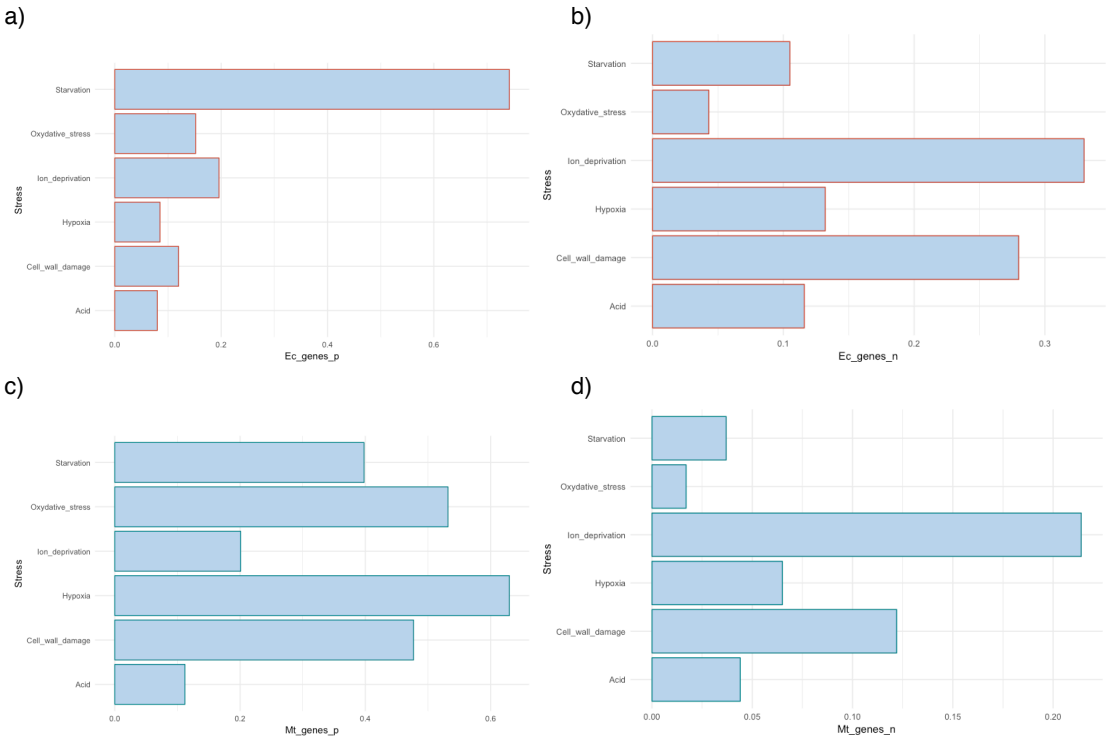

Figure_12. Bar diagrams proportion of genes per layer in both *E.coli* (red border line: a,c) and *M.tuberculosis* (blue border line: b,d) folding change multilayer network for overexpressed (a,b) and repressed (c,d) situations. Data filtered using a significance level of 0.05 in both cases. The layers correspond to six stress: acid, cell wall damage, hypoxia, ion deprivation, oxydative stress and starvation.

a) b)

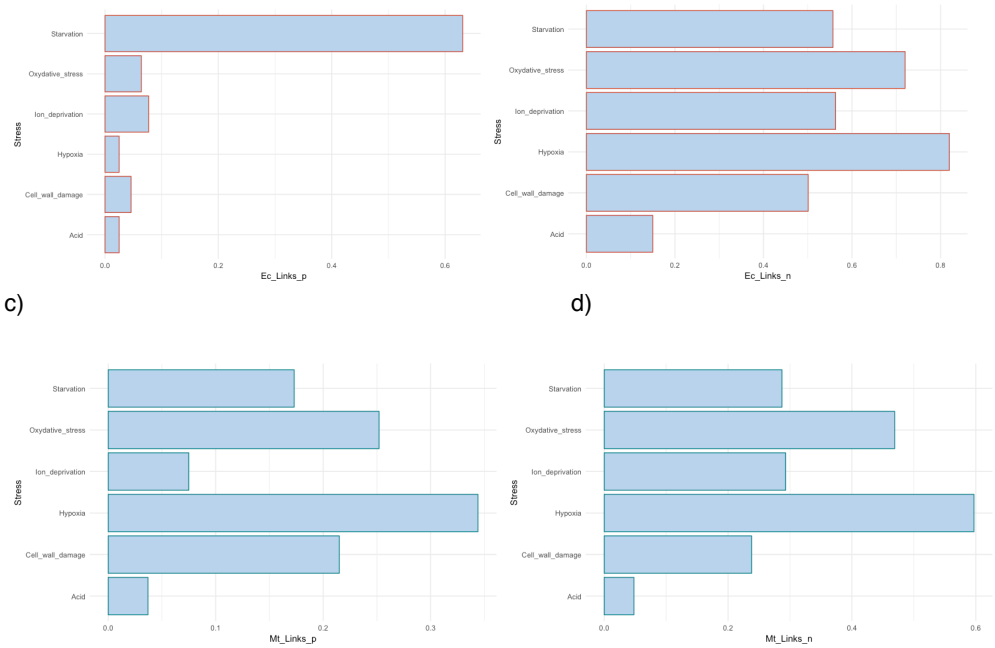

Figure_13. Bar diagrams proportion of links per layer in both *E.coli* (red border line: a,c) *and M.tuberculosis* (blue border line: b,d) folding change multilayer network for overexpressed (a,b) and repressed (c,d) situations. Data filtered using a significance level of 0.05 in both cases. The layers are associated to six stress: acid, cell wall damage, hypoxia, ion deprivation, oxydative stress and starvation.

The next results concern the analysis of the strenght, overlap and partition coefficient calculations. Interesting, we see that *E.coli* and *M.tuberculosis* networks do not show the same patterns. *E.coli* presents in general a higher overlap than *M.tuberculosis* but a lower partition coefficient, see the results in Figure_14. The figure shows two graphs of overlap versus partition coefficient for both *E.coli* and *M.tuberculosis* networks, respectively. Most of the genes in *E.coli* are located in the left part of the first quadrant, whereas in *M.tuberculosis* most nodes are located in the right part.

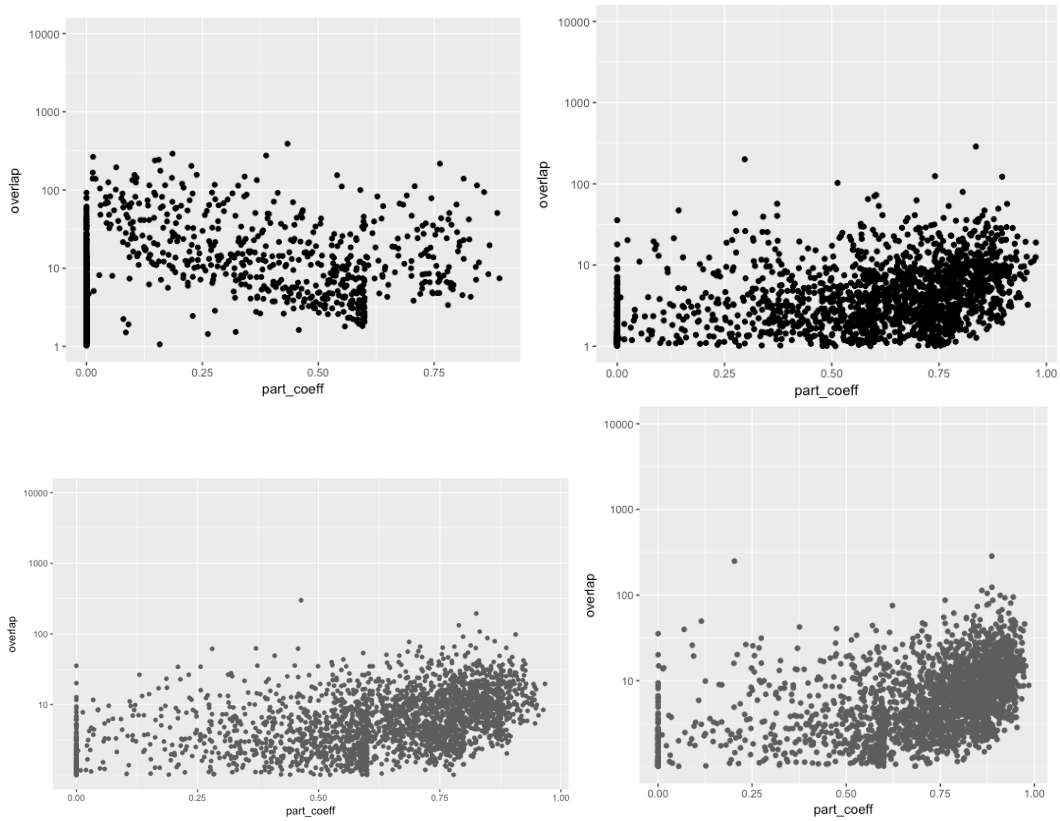

Figure_14. *E.coli* (a) *and M.tuberculosis* (b) overexpression overlap versus participation folding change multilayer networks. *E.coli* (c) *and M.tuberculosis* (d) repression overlap versus participation folding change multilayer networks. Significance level 0.05.

Regarding statistical analysis, first we compared *E.coli* and *M.tuberculosis* overlap medians using Mann-Whitney test and a p-value less than $2.2*10^{-16}$ was obtained. The null hypothesis supossed both equal medians and the alternative considers greater *E.coli* overlap median. That p-value was lower than the signicance level, so we rejected equality of overlap medians. Secondly, the same statistical method was used to compare partition coefficient medians between both *M.tuberculosis* and *E.coli*. Here, the alternative hypothesis says *M.tuberculosis* partition coefficient mean is higher than *E.coli*. The p-value obtained is less than $2.2^{*10\text{-}16}$. So, we rejected the null hypothesis. Finally, there is another statistics to compare 2D distributions overlap and partition coefficients: Peacock. After runninng it, the value of D is 0.4747 and after transformation to get a p-value, it was less than $2.2*10^{-16}$. So, we rejected the null hypothesis, meaning that we can say that both network distributions are not equal, see Figure_15 in which the density plots of overlap and partition coefficient are represented.

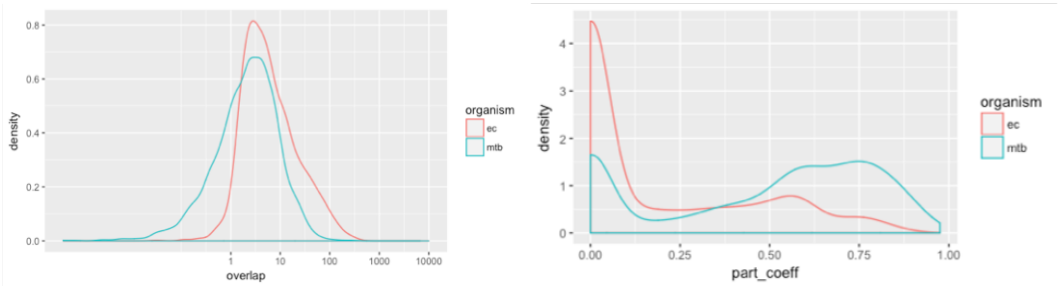

Figure_15. a) Density plot overlap *E.coli* (red) versus *M.tuberculosis* (blue) from overexpression folding change multilayer networks. b) Density plot partition coefficient *E.coli* (red) versus *M.tuberculosis* (blue) from overexpression folding change multilayer networks.

These results are interesting and novel, as they show that the multilayer network approach is not only relevant to isolate PPIs according to stresses -that is, working with the aggregated network could lead to highly misleading conclusions about which genes are active or not under a given stress-, but also relevant because the network metrics reflect the lifestyle of the bacterium. This is the case shown in Figure_14 for *E.coli*, which is a generalist organism that adapts to different environments. In fact, the average of its genome is shaped by a multitude of evulutionary forces from its primary (host) and secondary habitats, in which biotec (predators, competitor, cheaters, host defense mechanisms) and abiotic (pH, temperature, UV, depletion and so on) pressures are present [15]. Its expression of genes in each layer is different and we actually see how genes do not show the highest partition coefficient values. This means that they are not present in all layers, and thus, that they are specific stress responders. On the other hand, the bacterium has genes that can have high overlap values, which means that response to environment stress is also acute. On the contrary, *M.tuberculosis* as obligate pathogen is a specialist microorganism, whose entire life cycle is driven in the context of human infection and its metabolism necessarily underpins both physiology and pathogenesis [18]. So, this actinobacteria seems to respond to all the stresses in the same way (the strenght of the nodes is similar between layers). According to this observation, we see high partition coefficients, genes are moved to the right part of panel b of Figure_14.

The previous interpretation is very feasible, but we shoul also note possible alternatives. For instance, could the different number of samples used to build the network affect the results? We have already seen that the *E.coli* folding change multilayer network is smaller than the *M.tuberculosis* network. Thus, it might be that the statistical significance could affect the results, due to less statistical power of *E.coli*. If this were true, some samples could not pass the thershold (0.05), because of the stricter statistics selection procedure and as a consequence information could have been lost.

Although the last suggestion could be possible, however, there are counter arguments against this interpretation of the results. Firstly, in the same way that information may be lost because of the lower number of samples, n, in *E.coli* than in *M.tuberculosis*, the

opposite could be possible too. The presence of a higher amount of data for *M.tuberculosis* could add more noise. So, a higher number of samples do not have to mean necessarily higher signal. Measurements should be coherent. Secondly, if *E.coli* had low statistical power, neither the partition coefficient nor the overlap would be high. As a matter of fact, we have seen that although the partition coefficient median is lower in *E.coli* than in *M.tuberculosis*, the overlap values are higher for *E. Coli* when compared to those for *M.tuberculosis*.

5.2. Comparison of *E.coli* and *M.tuberculosis* Orthologous genes: pull and pairs.

The reseach for this analysis was focused on orthologous, i.e., those genes that have diverged due to speciation from an ancestral one and in which biological function is supposed to be conserved by different species [35]. We worked with two set of genes as detailed in Section 3.4. SET_1: Orthologs pairs *M.tb-E.coli*: pull orthologous collection and pair orthologous collection. Here, we want to see what is their behaviour by studying the overlap and the participation coefficients for each microorganism.

On the one hand, pull orthologuos overlap and partition coefficients were calculated from *E.coli* and *M.tuberculosis* folding change of overexpression networks. In Figure_16, we see that these parameters show differences between them. Here again, *M.tuberculosis* genes appear in the right part of the graphic while those for *E.coli* are left to the first quadrant. The interpretation is similar to the one exposed previously for the gene-stress-wise networks: in general, a higher overlap in *E.coli* can be related to their overexpression in a few stress layers, thus showing a specific and acute response to the corresponding stress. However, the partition coefficient is higher for *M.tuberculosis*, which is could mean that genes are equally overexpressed for almost any environmental stress. Interstingly, despite of the conserved function of these genes, they have different overexpression, depending on both the microorganism and the stress. Some statistical tests were also applied to support this latter result. The Mann-Whitney test, assuming as null hypothesis that medians are equal, yielded a p-value of $4.08*10^{-7}$. So, H_0 was rejected, being the alternative: "*E.coli* overlap medians are greater than *M.tuberculosis* overlap median", which supports our previous phenomenological interpretation. As far as the participation coefficient is concerned, for H_0 we assume that both microorganisms have equal medians; and the hyphotesis H_1 would be that *M.tuberculosis* has a greater median than *E.coli*. The p-value after applying the Mann-Whitney test was 0.0004131, that is, H_0 was rejected. Finally, the Peacock test gave a D value of 0.8457523 and after statistical transformation, its p-value was smaller than $2.2*10^{-16}$. Once again, H_0 is rejected and thus the distributions for both microorganisms are considered to be different for the orthologous as well. The density plots showing the different distributions of overlap and partition coefficient are shown in Figure_17.

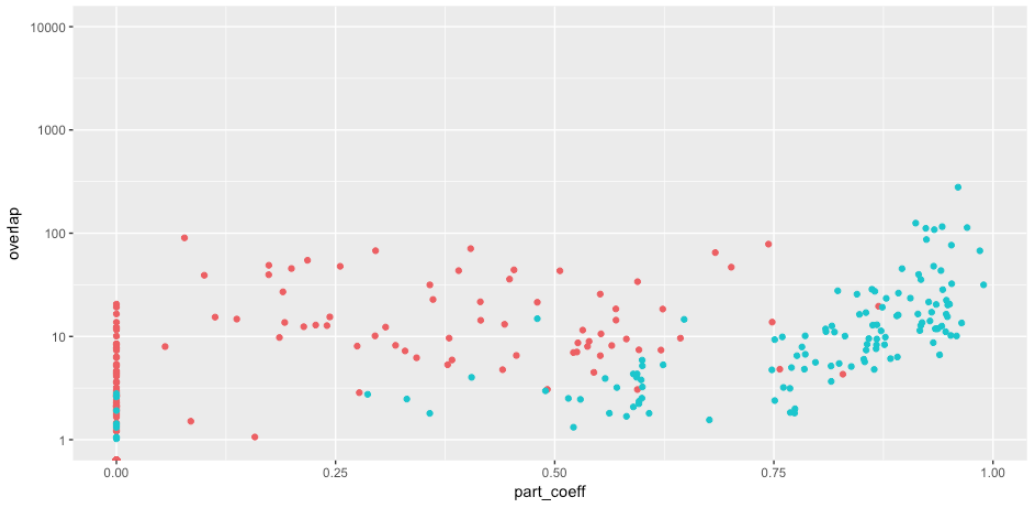

Figure_16. *E.coli* (pink dots) and *M.tuberculosis* (blue dots) pull orthologous overexpression overlap versus participations folding change multilayer network. Thershold/Significance level 0.05.

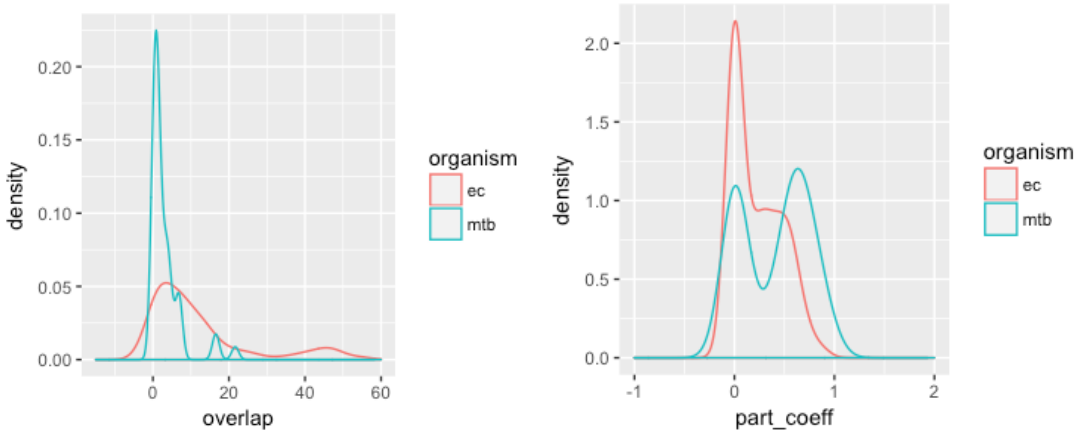

Figure_17. a) Density plot overlap *E.coli* (red) versus *M.tuberculosis* (blue) from both overexpression pull orthologous folding change multilayer networks. b) Density plot partition coefficient *E.coli* (red) versus *M.tuberculosis* (blue) from pull overexpression orthologous folding change multilayer networks.

On the other hand, the pair orthologous are only 43 (See Figure_18). They were treated in a special way. First, we test whether *E.coli* overlap is smaller than *M.tuberculosis* overlap (Mann-Withney test, p-value $5.007*10^{-6}$). In other works, the hyphotesis H_0, which assumes no difference and thus that medians deviation is equal to 0 was rejected, in favour of H1, i.e., that the median of the difference is greater than zero, see Figure_18). Secondly, we test whether *M.tuberculosis* partition coefficient is smaller than *E.coli* partition coefficient. Again H_0 considered the difference of the medians equal to zero and in the alternative scanario, that it would be greater than zero. As the p-value obtained was $1.911*10^{-7}$, H0 was rejected.

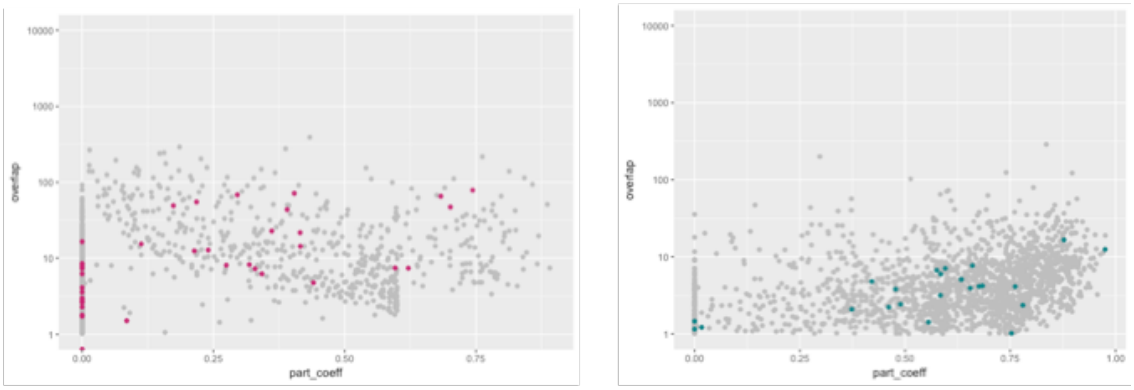

Figure_18. Antigen pair collections versus theirs completed folding changed overexpressed network for both microorgamisn a)*E.coli* (violetred) b)*M.tuberculosis* (turquoise)

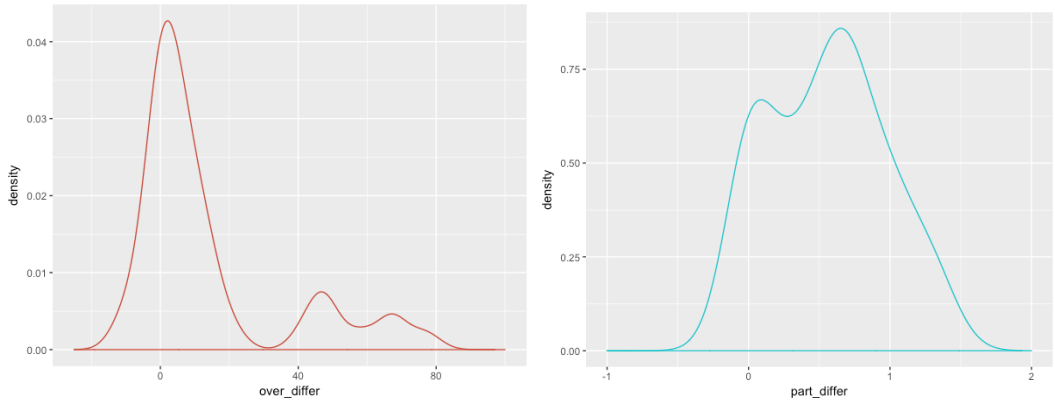

Figure_19. a) Density plot overlap difference between *M.tuberulosis* and *E.coli* values from *M.tuberculosis* and *E.coli* overexpression folding change multilayer networks. b) Density plot partition coefficient difference between *M.tuberculosis* and *E.coli* values from *M.tuberculosis* and *E.coli* overexpression folding change multilayer networks.

5.3. E.coli antigen analysis.

Considering the small number of antigens, 40 in the bibliography and only 14 in the *E.coli* overexpression folding change network (see selection process in 4. Methods), not so many statistical tests could be done with statistical power. However, here antigens are presented as some examples of how genes are behaving in each layer. We show in Table_6 the strength values for each antigen when subject to some stresses. Before analyzing these results, we note that in *Systems Biology of Mycobacterium tuberculosis. Immunogenicity, natural selection & gene expression in response to stress* by Miguel Báez Martín, the role of antigens in *M.tuberculosis* is studied and in this case they are really important, showing remarkable high ovelap and participation coefficient.

Table_6. Strenght values of *E.coli* antigen genes in *E.coli* overexpression folding change multilayer network for stresses (layers): acid, cell wall damage, hypoxia, ion deprivation, oxydative stress, starvation. Cell colour depends on strenght value: higher strenght value, more intensity blue scale.

STRENGTH	Acid	Cell wall damage	Hypoxia	Ion deprivation	Oxydative stress	Starvation
b3313	1,4405	3,8023	0	3,26717	0,709919	80,913947
b3314	4,0135	2,731059	0	5,1993	0,323483	163,924788
b1929	0	0	0	0	0	1,63319
b2697	0	0	0	0	0	4,6503
b2443	0	0	0	0	0	3,38055
b2395	0	0	0	0	0	0
b3732	0	0	0	3,976265	1,04424	4,960448
b1731	0	0	0	3,54264	0	6,38248
b1501	0	0	0	1,04793	0	1,00779
b3571	0	0	0	0	0	2,04925
b1048	0	0	0	0	0	0
b2319	0	0	0	0	0	15,15427
b1406	0	0	0	2,89066	0	0
b1606	0	0	0	2,71852	0	1,52553

Each gene in each dimension is evaluated independently. Each column in Table_6 is a layer representing a stress: acid, cell wall damage, hypoxia, ion deprivation, oxydative stress and starvation. As it can be seen, in the hypoxia situation, antigens are not overexpressed; in acid, cell wall damage, ion deprivation and oxydative stresses some of them are overexpressed, whereas starvation has most of the antigens overexpressed and strenght values stand out because they are higher. At this point, there are two important things to comment: i) as we have seen until now, all genes are not expressed in all stresses, and the same happens for antigens; and ii) the condition of stress by starvation has more samples that are used to build the network, thus, the higher overexpression leves for antigens in this experimental condition could be affected by the fact that there are more samples. This needs further investigation.

5.4. Gene ontology enrichement analysis comparing both E.coli and M.tuberculosis.

Enrichment of gene functions among the genes present in each layer are analysed by using ClueGo-cytoscape. However, we are still working on these results. Until now, both microorgnanisms have been analysed in the same conditions: using an input file of 141 genes as target (those with higher overlap), showing only pathways with pV ≤ 0.05 and choosing selected ontologies reference set. After running GlueGO Functional Analysis for studying each stress, both bacteria do not yield the same results: *E.coli* had functions enriched while *M.tuberculosis* have not. That fact may be because *E.coli* is one of the best characterized organisms and has an important role in genetic

engineering or synthetic biology (7 reference). Hence, a lot of data in relation to function, structure and gene information are collected in databases. On the other hand, the genome of *M.tuberculosis* is not as well-known as that of *E.coli* and many gene functions are not established. Regarding *E.coli* results, we see that several functions are enriched depending on the stress layer, see Figure_20 and Figure_21. Here, the results could be explained considering E.coli adaptability, which would allow the microorganism to vary and adapt its functional response according to the stress conditions and the environment. As for *M.tuberculosis*, we reanalyzed the data so that not only pathways with pV ≤ 0.05 were allowed. In this case, some common enriched functions do appear, and interestingly enough, their variability was much smaller, see Figure_22.

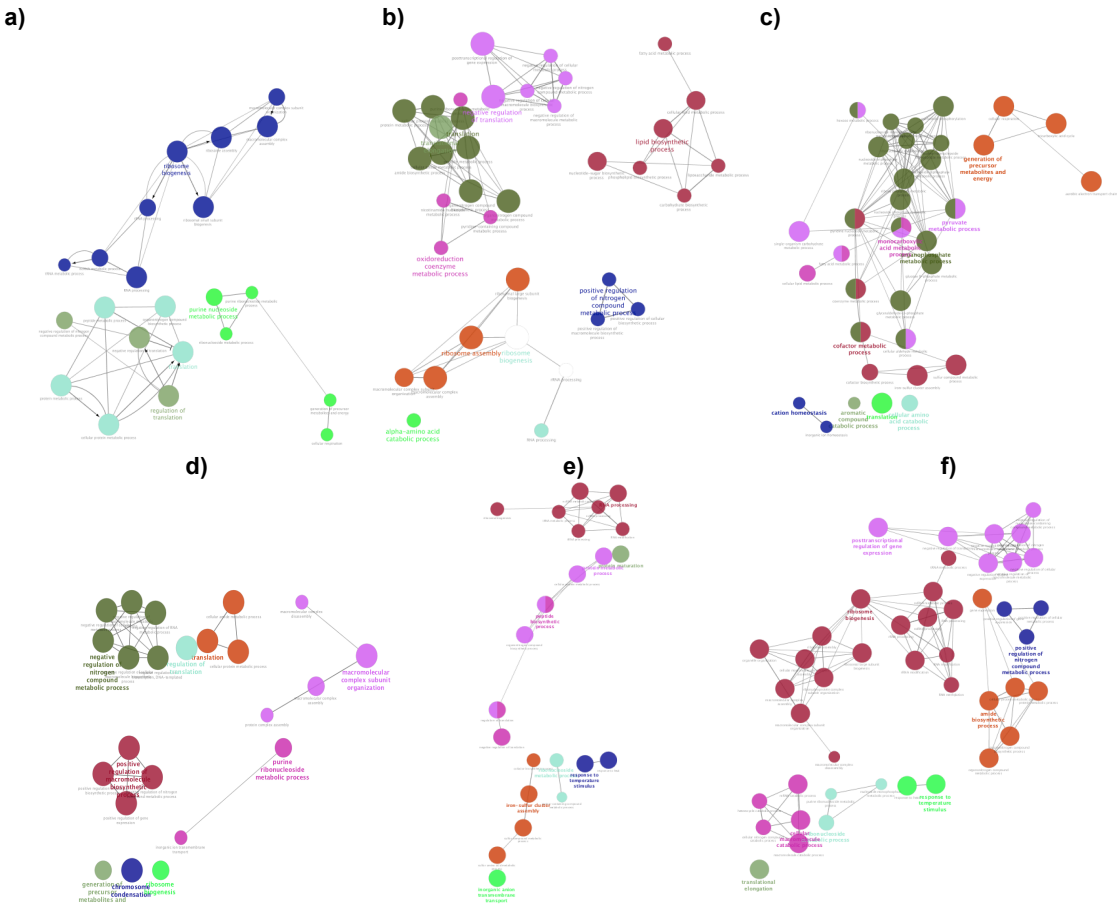

Figure_20. *E.coli* Enrichment of gene functions depending on the layer: a) Acid, b) Cell Wall Damage, c) Hypoxia, d) Ion deprivation and e) Starvation. Software Cluego-cytoscape, studying 141 genes, showing exclusively Pathways with pV ≤ 0.05 and choosing selected ontologies reference set.

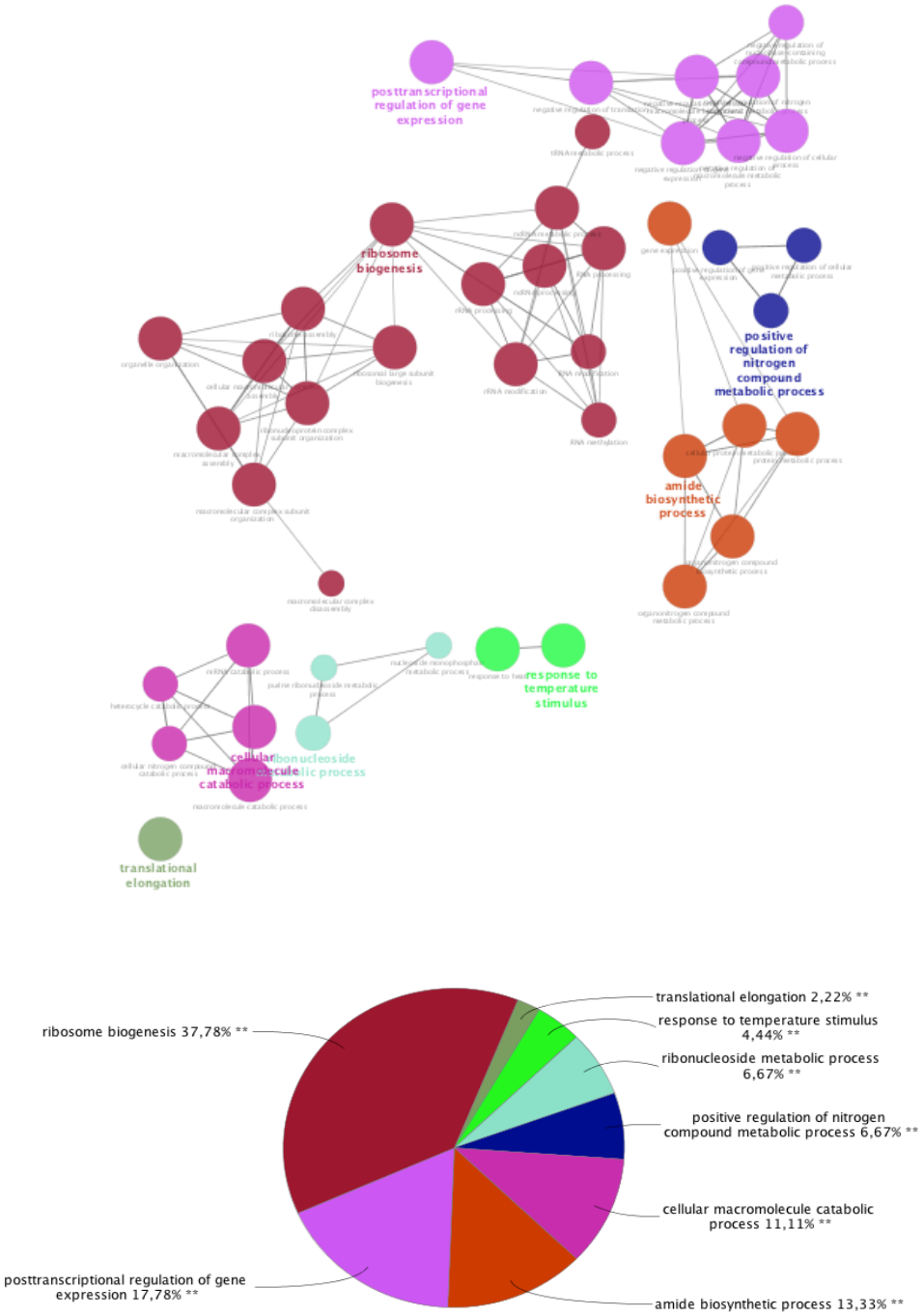

Figure_21. *E.coli* Enrichment of gene functions in the Acid layer. Software Cluego-cytoscape, studying 141 genes, showing exclusively Pathways with pV ≤ 0.05 and choosing selected ontologies reference set. Those enriched functions are ribosome biosynthesis (~38%), postranscriptional regulation of gene expression (~18%), amide biosynthetic process (~13%), cellular macromolecule catabolic process (~11%), positive regulation of nitrogen compound metabolic process (~7%), ribonucleoside metabolic process (~7%), response to temprature stimulus (~4%) and translational elongation (~2%).

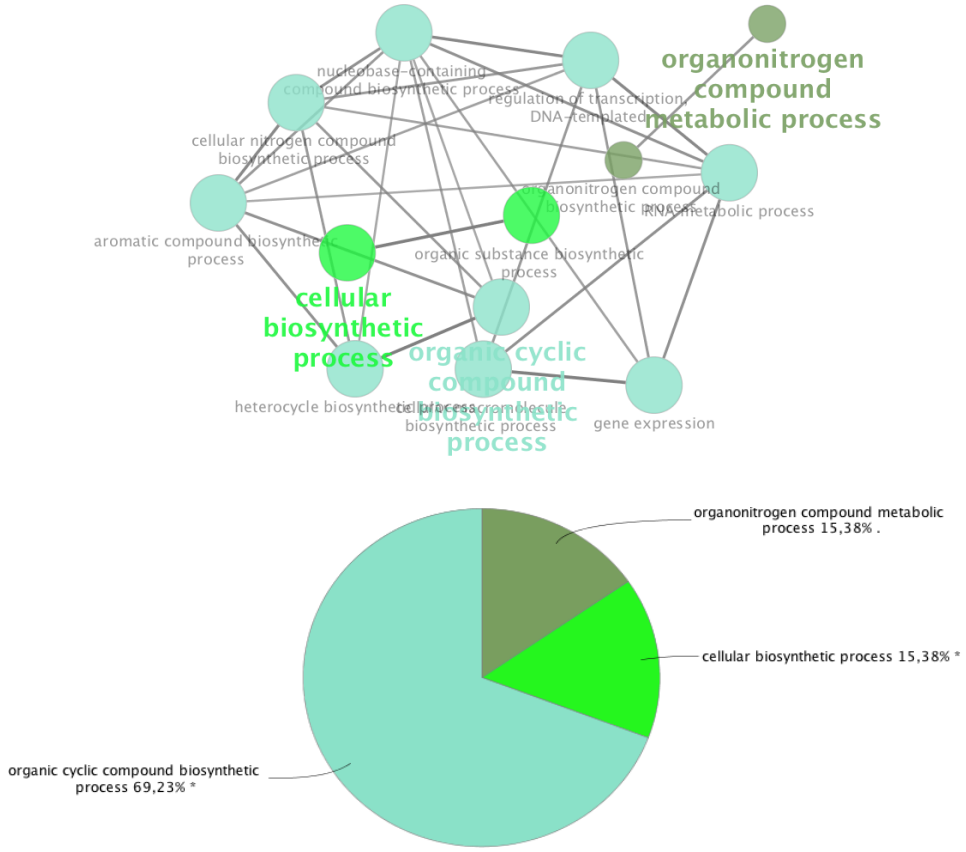

Figure_22. *Mycobacterium tuberculosis* Starvation stress gene ontology enrichment analysis from sofware ClueGo, being target-gene input file 141 genes. Without selecting p-value 0.05 (significance level) and choosing as reference set options custum reference set. Those enriched functions are organic cyclic compound biosynthetic process (~69%), organonitrogen compound metabolic process (~15%) and cellular biosynthetic process (~15%).

6. DISCUSSION

The comparison of both folding change networks *E.coli* and *M.tuberculosis* is the main objective of this work. We have shown that our results could provide new biological insights and correlate well with the biology of these two microorganims (see Section 5.1. Comparison E.coli versus M.tuberculosis). We observed that differences in life styles were translated into different distributions of genes in the plane overlap vs participation coefficient. Here, we would further discuss several aspects of the metaanalysis done. To better see whether both microorganisms folding change overexpression networks have different statistical power, we compare the strenghts (gene expression level for each stress). Figure_23 shows density plots for each layer (strenght) for each bacteria. As discussed previoulsy, the statistical tests (e.g., Mann-Whitney test) showed that most of the medians of *E.coli* when compared to those of *M.tuberculosis* were different. M.tuberculosis strength is greater than E.coli strenght for acid (p-value=3.759*10-5), cell wall damage (p-value=2.2*10-16), hypoxia (p-

value=2.2*10-16) and oxydative stress (p-value=2.2*10-16); whereas for starvation *E.coli* strength is greater than M.tuberculosis strenght. Finally, for ion deprivation, the overlap of both medians was not rejected as the p-value is 0.332. The previous results appear to be robust, both statistically and biologically, although we mention that *E.coli*, overall, could have lower statistical power than *M.tuberculosis*. Consequently, even thought it is not within the scope of this work, three data treatments to improve the statistics are planed for future research. In the first place, increasing the number of *E.coli* samples could be an obvious option if updated databases include a higher number of samples with correct nomenclature or a coherent sampling results. In the second place, instead of increasing *E.coli* sampling, a solution could be to reduce *M.tuberculosis* sampling. Lastly, another option would be to be more flexible in the *E.coli* network, that is, choosing a higher threshold or significance level: 0.1 instead of 0.05 (the limit had been used), although this solution is not recommended from a statistical point of view. The same could apply to the network analysis in relation to the orthologous.

Finally, the last analysis focused on functions' enrichment is still in process and it is not finished yet. However, the comparison of *E.coli* and *M.tuberculosis* results lead to the question of how much information about each of them is needed to make a thorough comparison. In this regard, our limitation is given by the relative high number of unknown functions and unclassified genes of *M.tuberculosis* with respect to what is known for *E.coli*.

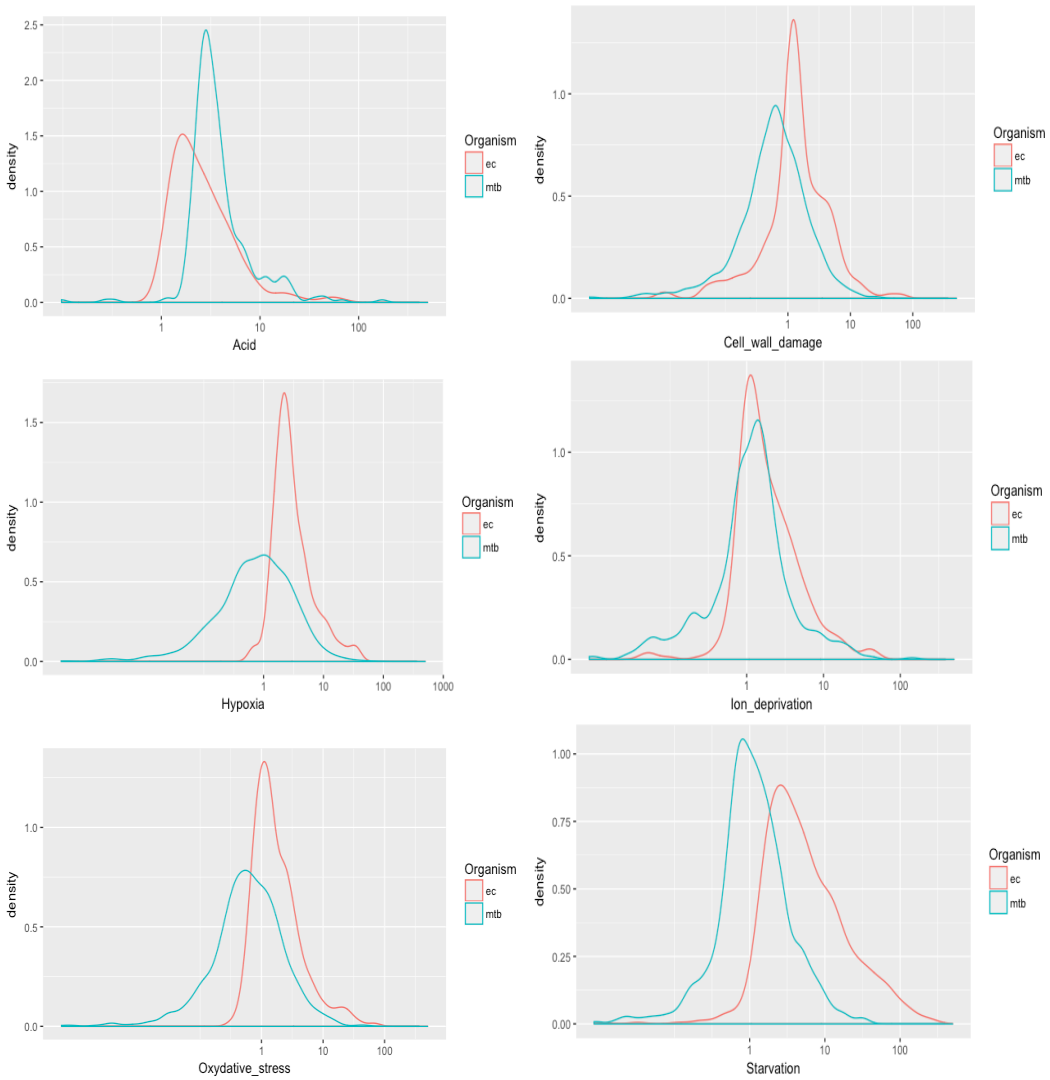

Figure_23 Density plot of *M.tuberculosis* versus *E.coli* gene strenght for each layer: a) Acid b) Cell Wall Damage c) Hypoxia d)Ion deprivation e) Oxydative stress f)Starvation. All strength data coming form *E.coli* and *M.tuberculosis* folding change overexpression multilayer network.

7. CONCLUSIONS

In summary, the conclusions of this work are:

Differences in life styles translate into different distributions of genes in the plane overlap versus the participation coefficient. *E.coli*, as a “generalist” organism, presents high overlaps and not so high partipation coefficients, meaning acute and specific gene overexpression in response to each environmental stress and high adaptation capacity while *M.tuberculosis* presents high participation coefficients as a “pathogen-specialist” organism, which implies the same kind of response to all stresses and only one objective: human infection.

The overlap-partition coefficient *E.coli* and *M.tuberculosis* orthologous genes pattern behaves in the same was as the rest of genes in both bacteria, despite of representing the same conserved function. In *E.coli* they are expressed in an acute specific way and in *M.tuberculosis* they are overexpressed in all stresses.

Antigens in *E.coli*, unlike in *M.tuberculosis*, do not have a noticeable role in *E.coli* genome and neither show a universal response to all sort of stresses.

In the future, gene function enrichment analyses should be completed. Meanwhile, we have obtained that *E.coli* has different groups of enriched proteins overexpressed per layer and that there is variability between them.

Finally, as per the academic and formative objectives of this work, we mention that we have learnt about network building, GEO, Biobase, Limma, measures as strength, overlap and partition coefficient, Mann-Whitney test, Peacock test, Cluego-cytoscape and applied these different stats/computational methods. Improving meta-analisis and solving possible statistical inconsistencies are our future goals.

Bibliography

[1] Oxford living Dictionaries
[2] Cambridge Dictionary
[3] Online etymology dictionary
[4] Wilson and Walker's. Principles and Techniques of Biochemistry and Molecular Biology (book).
[5] Fernanda Morillo, María Bordons, Isabel Gómez, Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and Technology Volume 54, Issue 13
[6] Carl R.Woese, Otto Kandler, Mark L. Wheelis. Proc. Nati. Acad.Towards a natural system of organisms: Proposal for the domains
Archaea, Bacteria, and Eucarya Sci. USA. Vol. 87, pp. 4576-4579, June 1990. Evolution
[7] Mohammad Arifuzzaman, Maki Maeda, Aya Itoh, Kensaku Nishikata, Chiharu Takita, Rintaro Saito, Takeshi Ara, Kenji Nakahigashi, Hsuan-Cheng Huang, Aki Hirai, Kohei Tsuzuki, Seira Nakamura,Mohammad Altaf-Ul-Amin, Taku Oshima, Tomoya Baba, Natsuko Yamamoto, Tomoyo Kawamura, Tomoko Ioka-Nakamichi, Masanari Kitagawa,Masaru Tomita, Shigehiko Kanaya, Chieko Wada, and Hirotada Mori. Large-scale identification of protein–protein interaction of Escherichia coli K-12. Published by Cold Spring Harbor Laboratory Press
[8] Zachary D Blount. The unexhausted potential of *E.coli*DOI:10.7554/eLife.05826.001
[9] David A. Hufnagel, William H. DePas, and Matthew R. Chapman. The Biology of the *Eschericha coli* Extracellular Matrix.Microbiol Spectr. 2015 June ; 3(3): . doi:10.1128/microbiolspec.MB-0014-2014.

[10] Matthew A. Croxen,* Robyn J. Law, Roland Scholz, Kristie M. Keeney, Marta Wlodarska, B. Brett Finlay. Recent Advances in Understanding Enteric Pathogenic Escherichia coli.
[11] David A. Hufnagel, William H. DePas, and Matthew R. Chapman. The Biology of the Eschericha coli Extracellular Matrix.Microbiol Spectr. 2015 June ; 3(3): . doi:10.1128/microbiolspec.MB-0014-2014.
[12] Leimbach A, Hacker J, Dobrindt U. 2013. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Current Topics in Microbiology and Immunology 358:3–32. doi: 10.1007/82_2012_303.
[13] Nerino Allocati, Michele Masulli, Mikhail F. Alexeyev, and Carmine Di Ilio. *Escherichia coli* in Europe: An Overview. Int J Environ Res Public Health. 2013 Dec; 10(12): 6235–6254.Published online 2013 Nov 25. doi: 10.3390/ijerph10126235
[14] MeSH (Medical Subject Heading), ncbi tool.
[15] Jan Dirk van Elsas, Alexander V Semenov, Rodrigo Costa and Jack T Trevors. Survival of Escherichia coli in the environment: fundamental and public health aspects.The ISME Journal (2011) 5, 173–183.
[16] Madhukar Pai, Marcel A. Behr, David Dowdy, Keertan Dheda, Maziar Divangahi, Catharina C. Boehme, Ann Ginsberg, Soumya Swaminathan, Melvin Spigelman, Haileyesus Getahun, Dick Menzies and Mario Raviglione. Tuberculosis. Natura Reviews, Disease Primers. Vol.2. 2016.
[17] James E. Galagan. Genomic insight into tuberculosis.
[18] Digby F. Warner. Mycobacterium tuberculosis Metabolism.doi: 10.1101/cshperspect.a021121. Cold Spring Harb Perspect Med 2015;5:a021121
[19] S.Boccaletti, V.Latora, Y.Moreno, M.Chavez, D.-U. Hwang. Complex networks: Structure and dynamics. Physics Reports 424 (2006) 175 – 308
[20] M. E. J. Newman. The Structure and Function of Complex Networks. SIAM REVIEW Vol.45,No.2,pp.167–256
[21] Albert-László Barabási. Network science. The scale-free property. Book is licensed under aCreative Commons: CC BY-NC-SA 2.0.
[22] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. arXiv:1308.3182v3. March 17, 2014
[23] Barabasi, A.L. and Z.N. Oltvai, Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004. 5(2): p. 101-13.
[24] EMBL-BEI, network analysis protein interaction data introduction
[25] Fold change rank ordering statistics: a new method for detecting differentially expressed
genes. Doulaye Dembélé and Philippe Kastner.
[26] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski and Jérôme Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics Vol. 25 no. 8 2009, pages 1091–1093
[27] Harry Caufield. A very brief guide to converting E. coli gene IDs. June 17, 2015
[28]PubMed Health Glossary. Source: NIH - National Institute of Arthritis and Musculoskeletal and Skin Diseases

[29]Julien I.E. Hoffman. Meta-analysis. In Biostatistics for Medical and Biomedical Practitioners, 2015.
[30] Sean Davis. Using the GEOquery package. April 7,2011.
[31] Buhm Han and Eleazar Eskin, “Interpreting Meta-Analysis of Genome-wide Association Studies”, In Press. PLoS Genetics (2012).
[32] Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith a, Thomas E. Nichols. Permutation inference for the general linear model. NeuroImage 92 (2014) 381–397.
[33] J.A.Peacock. Two-dimensional goodness-of-fit testing in astronomy. Mon. Not. R.astr. Soc (1983) 202, 615-627.
[34] Yuanhui Xiao.Package “Peacock.test”
[35] Hirokazu Chiba, Hiroyo Nishide, and Ikuo Uchiyama. Construction of an Ortholog Database Using the Semantic Web Technology for Integrative Analysis of Genomic Data . PLoS One. 2015; 10(4): e0122802. Published online 2015 Apr 13. doi: 10.1371/journal.pone.0122802]
[36] Anna Hart. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ. 2001 Aug 18; 323(7309): 391–393.
[37] Roger Y. Stanier, Michael Doudoroff, Edward A. Adelberg. The microbial World. 1966 Wiley-VCH

Annex

Table_1. *E.coli* and *M.tuberculosis* orthologous pairs are rows, columns are overlap and parition coefficients measures for both microorganism.

M.tuberculosis	Overlap	Partition Coefficient	*E.coli*	Overlap	Partition Coefficient
Rv0001	0,136806	0	b3702	8,489278	0
Rv0002	0,161188	0	b3701	67,6818573	0,295642614
Rv0054	1,45838	0	b4059	43,398092	0,39040296
Rv0189c	0,346829	0	b3091	7,72419	0
Rv0573c	0,831172	0	b0931	0	0
Rv0670	5,101826	0,634069995	b2159	7,24013	0,329278699
Rv0820	1,41939	0,556336961	b3725	78,3541714	0,743890553
Rv0884c	0,588186	0,671047538	b0907	2,28351	0
Rv0951	0,0503199	0	b0728	46,867817	0,701423925
Rv1017c	6,70078	0,575981228	b1207	54,660934	0,217980276
Rv1079	0,9345479	0,158411762	b3008	2,90733	0
Rv1098c	0,112152	0	b1611	7,380955	0,621227049
Rv1213	4,14774	0,676375723	b3430	0	0
Rv1286	16,611308	0,87749637	b2751	6,21715	0,342388151
Rv1293	3,8122355	0,477537546	b2838	2,70204	0
Rv1317c	0,925167	0	b2068	0	0
Rv1437	1,2147442	0,016801188	b2926	48,878036	0,173849638
Rv1451	1,016292	0,753161186	b0428	7,421869	0,596059071
Rv1522c	0,122564	0	b0462	2,6251	0
Rv1552	5,98064	0,585375631	b4154	2,66893	0
Rv1613	0,50608	0	b1260	1,72853	0
Rv1657	2,221192	0,461169199	b3237	8,06224	0,274641468
Rv1781c	7,717156	0,660430759	b3416	22,78961	0,361155664
Rv1931c	4,115873	0,761958886	b2916	1,5088751	0,084861294
Rv2139	0,303774	0	b0945	4,105	0
Rv2150c	0,8702489	0,314981592	b0095	21,645583	0,415141508
Rv2215	0,453712	0	b0727	8,19979	0,318431532
Rv2343c	2,426593	0,489251023	b3066	4,76571	0,440544623
Rv2428	7,082561	0,596103206	b0605	64,8917033	0,683179742
Rv2504c	12,493056	0,97500901	b2221	2,98506	0
Rv2785c	2,0844625	0,373690857	b3165	8,300067	0
Rv2841c	0,359162	0	b3169	15,381274	0,112556751
Rv2919c	0,9585003	0,643886833	b2553	0	0
Rv2986c	3,9233781	0,655101096	b4000	70,881193	0,404114553
Rv2987c	4,215527	0,684625383	b0071	16,51011	0
Rv3153	0,554706	0	b2281	0	0
Rv3215	0,3580571	0,426816161	b0593	12,42492	0,213687569
Rv3302c	5,0805866	0,633963912	b3426	7,33101	0
Rv3314c	1,15121	0	b4382	14,324039	0,415564629
Rv3411c	3,166208	0,584653105	b2508	1,78967	0
Rv3436c	2,35279	0,780063461	b3729	12,73453	0,240158705
Rv3534c	0,835702	0	b0352	3,58281	0
Rv3535c	4,7922834	0,421222871	b0351	6,22645	0

Table_2. Final set of antigens and some basic measures: overlap and partition coefficient.

Antigens	overlap	part_coeff
b3314	81691604	0.981508162214461
b3313	450768754	0.972440106766251
b1929	0.000000	0.0000000
b2697	1292441	0.597448295500777
b2443	0.904974	0.000000
b2395	0.000000	0.000000
b3732	117496984	0.92380902439915
b1731	0.000000	0.000000
b1501	0.000000	0.000000
b3571	0.000000	0.000000
b1048	0.000000	0.000000
b2319	145102886	0.893842588931651
b1406	372848	0.566580978090698
b1606	197414	0.000000

Table_3. *E.coli* individual series processing in network construction.

Individual series processing
GSE365.R
GSE1642.R
GSE5977.R
GSE6209.R
GSE6750.R
GSE8664.R
GSE8689.R
GSE8732.R
GSE8786.R
GSE8827.R
GSE8829.R
GSE8839.R
GSE9331.R
GSE10391.R
GSE13978.R
GSE14005.R
GSE14840.R
GSE15976.R
GSE16146.R
GSE21112.R
GSE21113.R
GSE35362.R
GSE50159.R

Code_1. Code to get measures: strength, overlap and partition coefficient of *E.coli* fonding change multilplex network. (R-studio).

```
#################### Folding_Change_Ecoli_Multilayer_Network ##################

# DATA PREPARATION -----------------------------------------------------------
#Load data

setwd("~/Desktop/TFM_analisis")
network <- read.table("Inputs/ppi_interaction_pairs_b.txt", header = FALSE,
 stringsAsFactors=FALSE)
genes <- unique(c(network[[1]],network[[2]]))

coefficients <-
 read.table("multiplexes_coli/multiplex_stress_wise/fcs/coefficients.txt", header =
 TRUE, stringsAsFactors=FALSE)
fdrs <- read.table("multiplexes_coli/multiplex_stress_wise/fcs/fdrs.txt", header = TRUE,
 stringsAsFactors=FALSE)

#Filter coefficients
fdrs_thres <- 0.05
coefficients_filtered = (coefficients*(coefficients < 0))*(fdrs<fdrs_thres)
#coefficients_filtered_positive = (coefficients*(coefficients > 0))*(fdrs<fdrs_thres)
#coefficients_filtered_negative = (coefficients*(coefficients < 0))*(fdrs<fdrs_thres)

stress_set <- c(1, 3, 6, 7, 9, 10)  #Ecoli_1 Acid,3 Cell wall damage, 6 hypoxia, 7 Ion
 deprivation,
#, 8 Oxydative stress, 9 Starvation

# FUNCTIONS & STRENGTH MATRIX ------------------------------------------------------------
 ------
strength <- function(gene, stress)
{intervent <- unique(c(which(network[[1]] %in% gene), which(network[[2]] %in% gene)))
  strength <- sum(coefficients_filtered[intervent, stress], na.rm=TRUE)
  return(strength)
}

#Matrix for strength_values, rows = genes, columns = stress
strength_matrix <- data.frame(matrix(nrow=length(genes), ncol=length(stress_set)))
#To name correctly the columns in strength_matrix
stress_names <- colnames(coefficients)[c(1, 3, 6, 7, 9, 10)]
colnames(strength_matrix) <- stress_names
rownames(strength_matrix) <- genes
for (stress in stress_names)
{
  for (gene in genes)
  {strength_matrix[gene, stress] <- strength(gene, stress)}
}

overlap <- function(gene)
{overlap <- sum(strength_matrix[gene, ], na.rm=TRUE)
  return(overlap)
}
overlap_values <- data.frame(sapply(genes, overlap))
colnames(overlap_values) <- c("overlap")

part_coeff <- function(gene)
{ M <- length(stress_set)
  square_values <- c()
  for (stress in stress_names)
  {
    square_values <- append(square_values, (strength_matrix[gene,
 stress]/overlap_values[gene,])^2)
```

```
  }
  squares <- sum(square_values)
  part_coeff <- (M/(M-1)) * (1-squares)
  return(part_coeff)
}

# DATA OUTPUT ---------------------------------------------------------
#Some genes NaN as result for part_coeff, reason --> overlap = 0 (and strength = 0 in
 every stress)
part_coeff_values <- data.frame(sapply(genes, part_coeff))
colnames(part_coeff_values) <- c("part_coeff")

metrics_ec_fc <- cbind.data.frame(overlap_values, part_coeff_values)
write.table(metrics_ec_fc, "o_p_ec_fc_neg.txt", sep ="\t")

library(ggplot2)
plot_o_p_ec_fc <- ggplot(metrics_ec_fc, aes(x = part_coeff, y = overlap)) + geom_point()
 +
  scale_y_log10(breaks = c(1, 10, 100, 1000, 10000), limits = c(1, 10000))
```

Code_2. Bootstrap algorithm for *E.coli* and *M.tuberculosis* overlap and partition coefficient data to get p-value from Peacok test (D). (R-studio)

```
fc_ec <- read.table("overlap_partcoef_ec.txt")
fc_mtb <- read.table("overlap_partcoef_mtb.txt")

## remove NAs of E.coli

set_NAs_overlap=which(is.na(fc_ec$overlap))
set_NAs_part_coeff=which(is.na(fc_ec$part_coeff))
set_NAs=unique(c(set_NAs_overlap,set_NAs_part_coeff))
ec=exp[-set_NAs,]

## remove NAs of M.tuberculosis

set_NAs_overlap=which(is.na(fc_mtb$overlap))
set_NAs_part_coeff=which(is.na(fc_mtb$part_coeff))
set_NAs=unique(c(set_NAs_overlap,set_NAs_part_coeff))
mtb=fc[-set_NAs,]

## Peacock original:
library(Peacock.test)
ks2 <- peacock2(ec, mtb)

## 1. declare a loop

iterations=100
random_peacocks=rep(NA,iterations)
genes_fc_ec=nrow(ec)
genes_fc_mtb=nrow(mtb)

for(i in 1:iterations)
{
    print(i)
    tab_tot <- rbind(ec, mtb)
    tab_tot=rbind(ec,mtb)
    tab_tot=tab_tot[sample(c(1:nrow(tab_tot))),]
    random_fc_ec=tab_tot[1:genes_fc_ec,]
    random_fc_mtb=tab_tot[(1+genes_fc_ec):nrow(tab_tot),]
    random_peacocks[i]=peacock2(random_fc_ec, random_fc_mtb)
}
```

Code_3. Code to get metadata and matrix of FCs per sample in serie name GSE15976, *E.coli.* (R-studio).

```
library(GEOquery)
library(Biobase)

dcols=function(x){data.frame(colnames(x))}

#############################################################
#### PART 1: Get metadata and matrix of FCs per sample ####
#############################################################

network=read.table("Inputs/ppi_interaction_pairs.txt",header=FALSE)

genes=c(as.character(network$V1),as.character(network$V2))
genes=unique(genes)
genes_up=toupper(genes)
net_up=network
net_up$V1=toupper(net_up$V1)
net_up$V2=toupper(net_up$V2)

### 1. Declare series name
name="GSE15976"

### 2. Browse series and declare number of experiments. This command often fails for no
 logical reason
gse <-
 getGEO(GEO=name,GSEMatrix=TRUE,destdir=paste0(getwd(),"/outputs/GEO_files/series"))
experiments=length(gse)
experiments
# 3
### ### For each experiment, repeat steps 3-4 to declare metadata
i=1

### 3. select samples useful for the analyses & columns carrying: title source_name_ch1
 source_name_ch2 description platform_id
metadata=pData(gse[[i]])
dcols(metadata)
intcols=c(1,8,18,31,33)
metadata=metadata[,intcols]
metadata[,-4]
samples=rownames(metadata)[c(1,2,3,8,13,14,15,19,21,31,34,36)]

### 4. select columns (metadata) and rows (samples), declare medium strain stress
 experiment and dyeswap setup.
###    Declare also metadata of discarded samples.

#Interesting columns: title, source_name_ch1, source_name_ch2, description, platform_id,
 then add: strain, CTL_medium, dye_swap.

metadata_discarded=metadata[which(!(rownames(metadata) %in% samples)),]
metadata=metadata[which(rownames(metadata) %in% samples),]
metadata$strain="H37RV"
metadata$dye_swap=0
metadata$comment="OK"
metadata$medium="7H9"
metadata$stress="Oxydative_stress"
metadata$stress[c(2,4,5,9:11)]="Cell_wall_damage"
metadata$experiment=c(
"5mM_Diamide_60min",
"0_05pc_SDS_60min",
"5mM_Diamide_60min",
"0_05pc_SDS_60min",
```

```
"0_05pc_SDS_60min",
"5mM_Diamide_60min",
"5mM_Diamide_60min",
"5mM_Diamide_60min",
"0_05pc_SDS_60min",
"0_05pc_SDS_60min",
"0_05pc_SDS_60min",
"5mM_Diamide_60min"
)
metadata$experiment=paste0(metadata$stress,"_",name,"_",metadata$strain,"_",metadata$exp
 eriment)
metadata_1=metadata
metadata_discarded_1=metadata_discarded

### 6. Get expression matrixes
exp_1=exprs(gse[[1]])
exp_1=exp_1[,which(colnames(exp_1) %in% rownames(metadata))]

### 6. Get feature-data matrixes and check extent of lost genes and reiterative
 expression estimates
fdata=fData(gse[[1]])
head(fdata)
fdata=fdata[,c(1,7)]
colnames(fdata)=c("ID","RV")
fdata=fdata[which(toupper(fdata$RV) %in% toupper(genes)),]
dim(fdata)
# 2841    2
length(unique(fdata$RV))
# 2841, to 2907, we need 66 genes more; and there is 2841-2841=0 reiterative entries
 (calculate medians in each of those).
fdata_1=fdata

### 7. Transform the expression matrixes so the rows are RV IDs present in the network
(and nothing else)

field_function=function(i){
    gene=genes_up[i]
    set=which(fdata$RV==gene)
    IDs=fdata$ID[set]
    set_values=which(rownames(exp) %in% IDs)
    if(mdata$dye_swap[j]==0){
        values=exp[set_values,j]}else{
            values=-(exp[set_values,j])}
        if(length(values)==1){
            return(values)
        }else{
            values=values[which(!is.na(values))]
            if(length(values)==1){
                return(values)
            }else{
                return(median(values))
            }}
}
# Declare fdata,exp,j before calling sapply
fdata=fdata_1
exp=exp_1
mdata=metadata_1
fdata$RV=toupper(fdata$RV)

genes_exp=data.frame(matrix(NA, ncol = ncol(exp), nrow = length(genes)))
colnames(genes_exp)=colnames(exp)
rownames(genes_exp)=genes
for(j in 1:ncol(exp))
{    print(j)
    genes_exp[,j]=sapply(c(1:length(genes)),field_function)}
```

```
genes_exp_1=genes_exp

##### 9. Transform the matrixes so rows are links.
link_field_function=function(i){
    index_a=which(genes_up==net_up$V1[i])
    index_b=which(genes_up==net_up$V2[i])
    value=genes_exp[index_a,j]+genes_exp[index_b,j]
    return(value)}

links_exp=data.frame(matrix(NA, ncol = ncol(genes_exp), nrow = nrow(network)))
colnames(links_exp)=colnames(genes_exp)
rownames(links_exp)=paste0(network$V1,"_",network$V2)

for(j in 1:ncol(links_exp))
{   print(j)
    links_exp[,j]=sapply(c(1:nrow(links_exp)),link_field_function)
}
```