TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2016

147

Vladimir Espinosa Angarica

A Bioinformatics Study of Protein Conformational Flexibility and Misfolding: a Sequence, Structure and Dynamics Approach

Departamento

Bioquímica y Biología Molecular y Celular

Director/es

Sancho Sanz, Javier

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Servicio de Publicaciones

ISSN 2254-7606

Tesis Doctoral

A BIOINFORMATICS STUDY OF PROTEIN CONFORMATIONAL FLEXIBILITY AND MISFOLDING: A SEQUENCE, STRUCTURE AND DYNAMICS APPROACH

Autor

Vladimir Espinosa Angarica

Director/es

Sancho Sanz, Javier

UNIVERSIDAD DE ZARAGOZA
Bioquímica y Biología Molecular y Celular

2016

Universidad Zaragoza

Thesis

Submitted to the *Universidad de Zaragoza*

in candidature for the degree of

Doctor of Philosophy

A Bioinformatics Study of Protein Conformational Flexibility and Misfolding: a sequence, structure and dynamics approach

Author
Vladimir Espinosa Angarica

Supervisor
Prof. Javier Sancho Sanz

Departamento de
Bioquímica y Biología
Molecular y Celular
Universidad Zaragoza
1542

Declaration of Authorship

I, VLADIMIR ESPINOSA ANGARICA, declare that this *PhD Thesis* titled, "**A Bioinformatics Study of Protein Conformational Flexibility and Misfolding: a sequence, structure and dynamics approach**" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University of Zaragoza.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this *PhD Thesis* is entirely my own work.
- Where the *PhD Thesis* is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I have acknowledged all main sources of help and funding.

Place and Date: ____

Signature: ____

Supervisor Statement

I, Prof. JAVIER SANCHO SANZ, Lecturer of the Department of Biochemistry and Molecular and Cellular Biology, University of Zaragoza.

DECLARE:

That this *PhD Thesis* titled, **“A Bioinformatics Study of Protein Conformational Flexibility and Misfolding: a sequence, structure and dynamics approach”** presented by the *B.Sc.* VLADIMIR ESPINOSA ANGARICA has been done under my supervision at the Department of Biochemistry and Molecular and Cellular Biology, University of Zaragoza. I would also like to state that this *PhD Thesis*, and the work included in it correspond to the Thesis Project approved by this institution and that this project satisfies all the requisites to be presented to obtain the scientific degree of Doctor of Philosophy by the University of Zaragoza.

And as evidence hereby I sing this copy.

Place and Date: ____

Signature: ____

“Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world . . . ”

Albert Einstein

Agradecimientos

Independientemente de que es incuestionable que el Inglés es el lenguaje de la ciencia, no es menos cierto que nuestra lengua es mucho más rica y apropiada para expresar sentimientos y emociones. Por esto, y porque no hubiese sido justo privar de algunas palabras en español a este documento, he decidido incluir en mi lengua materna mis agradecimientos a todos los que de una forma u otra han contribuido directa o indirectamente a mi formación personal y científica. Considero que llegar hasta aquí es el resultado del apoyo, el trabajo y el cariño de muchas personas durante años, e intentaré de alguna manera recordarlos a todos, aunque muy probablemente no lo conseguiré, en cuyo caso pido perdón por adelantado.

Primero que todo, y no por cumplir simplemente con un formalismo, quisiera agradecer a mi Tutor, el Profesor Javier Sancho Sanz, por ser el principal artífice de los proyectos que han constituido el cuerpo central de esta Tesis. Muchísimas gracias no solo por guiarme académicamente y contribuir significativamente a mi formación como investigador, sino también por proporcionarme un excelente entorno de trabajo en el cual desenvolverme, por financiarme durante estos años y por espolearme siempre que ha sido necesario, e intentar en cada momento sacar lo mejor de mí. Muchísimas gracias también por tantas discusiones científicas de las que tanto hemos aprendido ambos –por mi parte seguro, por la suya espero– y por darme la posibilidad de ser independiente y las facilidades para colaborar con otros grupos sin el corsé rígido de un doctorado al uso, y permitirme disfrutar así de lo realmente entretenido y fascinante de la ciencia: explorar y aprender cosas nuevas. También un agradecimiento muy especial a la Dra. María Ángeles Jiménez López, quien ha sido mi Supervisora institucional por parte del Instituto de Química Física "Rocasolano", del CSIC, por la gentileza de actuar como tal, lo cual me ha permitido beneficiarme del financiamiento del CSIC durante mi doctorado.

Siguiendo con contribuciones académicas, quisiera recordar a excelentes profesores y colaboradores que me han enseñado y marcado durante mi vida como estudiante primero y aprendiz de investigador después. Muy especialmente, a Ciro Mora (*R.I.P.*), Isidro Martínez Elejalde, Frank Coro y Joaquín Díaz Brito, excepcionales profesores que estoy seguro no solo me han marcado a mí, sino a una gran cantidad de profesionales cubanos ahora esparcidos por todo el mundo. También sería imposible dejar de reconocer a Abel David González Pérez, Ana Tereza Ribeiro de Vasconçelos, Julio Collado Vides y Bruno Contreras Moreira, a los que ha sido un placer tener como jefes o colaboradores y con los cuales di mis primeros pasos en el mundo de la ciencia, y de los que he aprendido mucho durante estos años de trabajo. También agradecer a los investigadores con los que he colaborado en múltiples proyectos científicos en España como Modesto Orozco, Salvador Ventura y María Fillat, quienes me han ofrecido la oportunidad de participar en sus investigaciones, me han permitido experimentar e investigar en nuevos temas y me han abierto la puerta de sus grupos de investigación.

Un apartado especial merecen las diferentes Instituciones, Fundaciones e Institutos que han financiado mis estudios en España que han concluido con la presentación de esta Tesis. Mil gracias al Banco Santander, Fundación Carolina y Universidad de Zaragoza, de cuyo programa de 'Becas para Lationamericanos' me beneficié para venir a España a iniciar mis estudios de Máster. En estos momentos tan negros para la financiación científica en España y cuando el Consejo Superior de Investigaciones Científicas pasa por tan mal momento económico, también quisiera agradecer al programa de 'Becas Doctorales de la Junta de Ampliación de Estudios' de esta institución, la cual me ha financiado completamente durante los 4 años de duración de este doctorado, incluyendo financiación directa y un excelente programa de intercambio científico y viajes. Ojalá esta institución de referencia siga existiendo y desempeñando este imprescindible papel de formación en investigación en el futuro. Una especial mención para la Universidad de Zaragoza, que además de financiarme se ha convertido en mi segunda *Alma Mater*, así como al Instituto de Biocomputación y Física de Sistemas Complejos, donde también he desarrollado gran parte de mis investigaciones.

Ahora quizás venga la parte más emotiva, correspondiente a los agradecimientos personales a tanta gente increíble que me ha asignado la vida por una parte, y que me he ido encontrando durante mi paso por diferentes países por otra. Primero, y como no podía ser de otra manera, el agradecimiento supremo a mi gran familia, a mi Mamá que me ha hecho quien soy y que tanto ha pasado conmigo en momentos de enfermedad y felicidad, a mi Papá quien siempre ha sido para mí un modelo de integridad y honradez, a mi hermano Pablo Andrés quien siempre fue un referente para mí profesional y personalmente y quien tanto ha tenido que ver en que yo haya podido venir a España, y a mi hermano Pedro Pablo quien siempre me enseñó a ver lo mejor de las cosas y a ser feliz no obstante la adversidad. Con esta gran familia me formé como persona en el mejor ambiente posible para aprender los mejores valores que me han guiado y espero me guíen toda la vida: la honradez, el amor a la familia por encima de todo y nunca desfallecer en la búsqueda de la excelencia. Estas personas tan importantes en mi vida han tenido que hacer un gran esfuerzo en los últimos años en los que me han tenido tan lejos, sin embargo, nunca han dejado de estar a mi lado y por eso les doy mil gracias.

También agradecer a los nuevos integrantes de mi familia, a todos mis sobrinos, algunos niños y otros ya casi mujeres y hombres, que tanta alegría me han dado y me dan, a mis cuñadas, pero muy especialmente a la Charo y a su gran familia, a Quini (*R.I.P.*) y Consuelo, quienes me han acogido en España como a un hijo y con los que he pasado tantos momentos inolvidables, y a mis abuelos Félix Angarica y Basilia Dueñas, los únicos que conocí y que ya no están. Y por supuesto todas las gracias y el amor del mundo para Francesca Spagnuolo, por darme tanto como me das cada día, por hacerme mejor persona y por hacerme pasar los momentos más increíbles de mi vida. Gracias a Dios has venido acompañada de una gran familia 'Napoletana' que me ha acogido como a un hijo y que tanto cariño me demuestran cada día, por eso quisiera recordar muy especialmente a Giuliana de Mattia, Fernando Spagnuolo, Giulio y tantos tíos y otros familiares, pero con especial cariño al Gran Adriano de Mattia.

No me permitiré olvidar a muy buenas amigas y excelentes personas que he encontrado aquí, con las cuales he pasado tan buenos y felices momentos y que me alegro formen parte importante de mi vida como Adrianita y Lourditas. Y a mi hermano Andrés González, al que me encontré aquí en Zaragoza de casualidad, y que ahora está en una posición muy especial entre mi gente más querida y a los que considero mi familia postiza.

Por último y no por esto menos importante, quisiera cerrar estos agradecimientos con un especial recuerdo para tantos y tan buenos amigos de allá y de acá, de ahora y de siempre. Mis grandes recuerdos para el Dersu, Juan Carlos, Lesly y otros tantos de "La Lenin" que vienen conmigo casi desde toda la vida y ahora están dispersos por todas partes del mundo, y algunos pocos en Cuba. A gran gente que conocí en la Universidad de la Habana, mi primera *Alma Mater*, durante la carrera de Bioquímica como Abelito, el Karel, Madelyn, Emilio y otros muchos. Recuerdo también a muy buena gente que conocí en Brasil y México, y especialmente a Lucía Morales y Patricia Oliver. Y ahora, los más recientes y con los que tan buenos momentos he pasado en España. A las personas del Departamento de Bioquímica de la Universidad de Zaragoza, empezando por la gente de secretaría, especialmente a Marta Fajés por su gran eficiencia en todo lo que hace y a la que quiero agradecer por toda la ayuda que me ha dado a lo largo de estos años, así como a todos los profesores y profesoras que tan agradables han sido conmigo durante mi estancia. Y finalmente, recordar a todos con los que más he interactuado durante este tiempo, a la gente de mi grupo, pasados y presentes, al Xabi, Sara, Nunilo, Laura, Jorge, Renzo, Juan José, María, Juanola, Reyes, Raquel y muchos otros, así como a los de otros grupos pero que forman parte del círculo íntimo, Ana Sánchez, Isaías, Sonia y muchos más que no por no ser incluidos explícitamente son menos queridos o recordados.

A todos muchas gracias!!!

Table of Contents

List of Figures

List of Tables

Abbreviations

PrD	**Pr**ion **D**omain
IDP	**I**ntrinsically **D**isordered **P**rotein
ROC	**R**eceiver **O**perating **C**haracteristics
AUC	**A**rea **U**nder the **C**urve
HMM	**H**idden **M**arkov **M**odel
NMR	**N**uclear **M**agnetic **R**esonance
LIP	**L**ight **I**nterface of high **P**olarity
PDB	**P**rotein **D**ata **B**ank
SCOP	**S**tructural **C**lassification **o**f **P**roteins
SNP	**S**ingle **N**ucleotide **P**olymorphism
FH	**F**amilial **H**yphercholesterolemia
LDL	**L**ow **D**ensity **L**ipoprotein
SASA	**S**olvent **A**ccessible **S**urface **A**rea
OMIM	**O**nline **M**endelian **I**nheritance in **M**an
HGMD	**H**uman **G**ene **M**utation **D**atabase
PCA	**P**rincipal **C**omponent **A**nalysis
MD	**M**olecular **D**ynamics
HTML	**H**yper **T**ext **M**arkup **L**anguage
Uniprot	**U**niversal **P**rotein **R**esource
TrEMBL	**T**ranscription **EMBL**
GO	**G**ene **O**ntology database
PHP	**H**ypertext **P**reprocessor
SQL	**S**tructured **Q**uery **L**anguage

A Mama, Papa, Pablo Andrés, Pedro Pablo, los niños y Francesca

Thesis Structure

The *Front Matter* of this Thesis contains the **Declaration of Authorship**, **Supervisor Statement**, **Agradecimientos**, **Table of Contents**, **List of Figures**, **List of Tables** and the **Abbreviations**. The **Agradecimientos** section had to be written in Spanish on a matter of principles, while the main part of the document is written in English.

The *Main Text* is distributed in a **Resumen** section and five chapters all of which, with the exception of the **General Introduction** and **Chapter 5**, are structured similarly to standard research articles, as most of the content in this document has been extracted from recent papers published by us during the duration of this Thesis. In some cases, the structure of the original articles has been modified to be adapted to a formal Doctoral Thesis format, and also to seamlessly connect and combine the information coming from different papers on the same subject. The contents in each chapter correspond fairly well to the results that can be found in the corresponding journal articles, though in some cases new text has been added, some new figures have been included and some references have been added or updated. The chapters are structured to be self-contained, each one having its own *Bibliography* section, and from **Chapter 2** to **Chapter 4** an *Introduction*, *Results*, *Discussion*, *Conclusions* and *Methodology* sections.

The **Resumen** section is written in Spanish to satisfy the formatting requisites from the University of Zaragoza for theses written in English. It includes a succinct summary of the previous knowledge in the fields that will be covered in this Thesis, as well as a brief description of each objective, the methodological approaches followed to try to prove our scientific hypotheses, and our contributions as a result of the work done during this Thesis.

In **Chapter 1** we include a **General Introduction** in which we provide a detailed outline of the state of the art in the fields of study that served as precedents for this Thesis. First, we treat the concepts of protein conformational instability and how it is related to protein evolvability, function and multitasking, as well as the group of mechanisms by means of which cells generate and take advantage of this phenomena and how they protect against the noxious effects that might be caused by protein local or partial unfolding. Then, we review protein aggregation as a general property of polypeptides, the structural and sequential characteristics of aggregation-prone domains and the beneficial roles played by amyloids in organisms, and also the diseases arising from the deposition of protein fibrils. Finally, we get into greater detail into the plethora of experimental techniques for the study of protein structural motions, unfolded intermediates and amyloid formation, and how this wealth of information has been used to devise computational approaches that can be useful for complementing experiments, and also for performing wide-scale predictive studies. After **Chapter 1** we present the **Objectives** section, outlining the principal aims of this Thesis which are formulated taking as inspiration the precedents described earlier in **Chapter 1**.

In **Chapter 2** we cover the work done by us to fulfill the first specific objective of this Thesis, in which we developed a methodology for predicting prionogenic domains in proteins based on primary sequence information. We include an *Introduction* describing some general characteristics of prions, which are mediators of important functions in cells, and also main factors of transmissible and inherited diseases. Then, we present the problem, which is that the number of prions known nowadays is limited to a few examples in microorganisms and mammals, and there does not exist a complete view of prion biology from a genomic or cross-species perspective, hence the necessity for the generation of computational strategies to predict prions. Later, we present our approach, which is based on the hypothesis that it would be possible to learn the key rules to determining the prionogenicity from the analysis of the composition of a sufficient number of experimentally tested prions. We also describe the particularities of our model when compared with others recently developed, and the details of the statistical calibration done before performing genome wide scanning in order to discover all the putative prion proteins in the genomes of all organisms. Finally, we present our predictions and discuss the possible functional implications of these findings to cell biology.

In **Chapter 3** we address the second specific objective of this Thesis, that is the development of a structured-based methodology for predicting flexible or conformationally unstable regions in proteins, trying to infer rules from the tridimensional structure of proteins. In the *Introduction* we outline the importance of protein conformational instability, how it is related to protein function and disease, and how it is possible to obtain atomic detail information of structural fluctuations that could lead us to a model based on structural properties. Then, we present our methodology, which comes from the rationalization of atomic-resolution studies of protein folding intermediates, and relies on the hypothesis that protein cores could contain regions with peculiar physicochemical properties specifically suited to ease the reorganization of the contacting segments, therefore allowing functionally relevant intradomain motions. We describe in detail our method, which systematically decomposes the structure of a given protein from end-to-end reckoning some properties of the buried contact interfaces between a short probe-segment and the rest of the protein. At the end of the chapter we describe the results obtained for the prediction of locally unstable regions in a significative group of protein families, and the good agreement resulting when we have contrasted our predictions with experimental data from protein conformational instability, folding intermediates and transition states.

In **Chapter 4** we undertake the last objective of the Thesis, for the study of the fate of mutations in protein structure and how it might be related with conformational diseases, directly assessing this phenomenon at the structure-temporal level using Molecular Dynamics. In the *Introduction* we describe the specific conformational disease that will be our case of study –*i.e.* **F**amilial **H**ypercholesterolemia (FH)– and how mutations in one of the proteins from the cholesterol metabolism could cause the disease by impeding its correct folding or function. We supply a description of FH, which is a very important genetic disease in human populations, and is mainly caused by mutations in the gene encoding the **L**ow **D**ensity **L**ipoprotein **r**eceptor (LDL-r), a modular transmembrane protein that plays an essential role in the mechanism of cholesterol uptake into cells. Then, we present the problem, which is that so far, the number of mutations linked to the disease in specific populations is scarce, and there are important experimental limitations to study how all the possible mutations affect the stability of the protein given its high size. Thus, our methodology intends to study the interaction domain (LA5) of the LDL-r computationally, and generates all the possible mutants arising from **S**ingle **N**ucleotide **P**olymorphisms (SNPs), to try to understand how mutations affect the conformational dynamics of this module, which although operationally complex, is attainable computationally. We present our results arising from the application of combined Data Mining methodologies to identify singularities in the conformational behavior of different types of mutants that could cause the destabilization of the LA5 domain, thus impairing recognition of LDLs, and discuss the possible applications of these kinds of methodologies for helping experimentalist to study other conformational diseases.

The **Chapter 5**, which is also written in Spanish to satisfy the formatting requisites from the University of Zaragoza for theses written in English, is a brief and summarizing section to include the main general *Conclusions* that can be drawn from our results, and the main *Perspectives* in these research lines for the future.

The *Back Matter* includes some **Appendices** with supplemental information to the results presented in each chapter, each one coded and ordered with consecutive capital letters. These appendices might also contain some figures and tables that, notwithstanding their relevance for the conducting thread of the Thesis and the results presented, are too big to be included in a common page, with the predefined portrait formatting. The final appendix contains a short *Curriculum Vitae* of the candidate. The last section is a **Keyword Index** with an alphabetical list of words and expressions and links to the pages of the document on which they can be found.

Resumen

Esta Tesis Doctoral se centra en varios estudios computacionales sobre la flexibilidad conformacional y problemas de plegamiento en proteínas realizados a diferentes niveles de complejidad: a nivel de secuencia, a nivel de estructura terciaria y a nivel de dinámica temporal. La adopción de la estructura nativa de las proteínas es uno de los procesos más importantes en la célula, siendo fundamental para posibilitar la correcta función de las mismas. En los últimos años, las evidencias experimentales y computacionales obtenidas han cambiado radicalmente la visión que se tenía de las proteínas, consideradas como entidades más bien estáticas en las cuales la función estaba mediada por la correcta adopción de una estructura nativa bien definida. Actualmente está prácticamente aceptado, en base a estudios de numerosas familias de proteínas diferentes, que en muchos casos la función de las proteínas está mediada por conjuntos de estructuras alternativas parcialmente desplegadas, que se encuentran en equilibrio con la estructura nativa. También se ha demostrado que la inestabilidad y flexibilidad conformacional de las proteínas juega un papel fundamental en su versatilidad funcional, permitiendo que la misma especie proteica esté relacionada con varias funciones en la célula. Esta inestabilidad estructural característica de las proteínas también está muy relacionada con su susceptibilidad a agregar formando fibras amiloideas o priones, los cuales están relacionados con una gran cantidad de enfermedades graves y casi siempre irreversibles. Ésta es una de las razones por las cuales el estudio de la estabilidad estructural, el plegamiento de proteínas y las enfermedades conformacionales son campos de estudio de gran actualidad y actividad, tanto desde el punto de vista experimental como computacional. El principal objetivo de estos estudios es intentar comprender a fondo los determinantes estructurales que provocan esta inestabilidad estructural, y como ésta se relaciona con la función y la posible agregación de las mismas, así como también poder desarrollar estrategias y terapias para tratar estas patologías. Nuestro trabajo incluye estudios computacionales de predicción y caracterización de motivos secuenciales y estructurales determinantes de agregación y desplegamiento en diferentes proteínas. Una descripción detallada de cada una de estas líneas de trabajo se incluye a continuación.

Inicialmente estábamos interesados en generar una metodología que nos permitiera hacer una predicción de proteínas prionogénicas basada en representaciones probabilísticas de dominios ricos en Glutamina/Asparagina. Los priones son generalmente proteínas propias codificadas en el genoma, con funciones específicas cuando se encuentran correctamente plegadas, que también pueden mediar otras funciones importantes en las células luego de su conversión amiloidea, por ejemplo como elementos epigenéticos, capacitores evolutivos y en procesos de adaptación a las fluctuaciones medioambientales en microorganismos. Además de estas funciones beneficiosas para los organismos, también pueden participar como factores principales en múltiples enfermedades hereditarias y transmisibles, así como en algunas enfermedades neurodegenerativas como Alzheimer y Parkinson, o en encefalopatías causadas por priones infecciosos en algunos mamíferos y el hombre. Sin embargo, hasta la fecha, el número de priones conocidos y caracterizados experimentalmente es muy escaso. De aquí el intenso trabajo que se está realizando para desarrollar estrategias que permitan predecir priones desde el punto de vista computacional y validarlos experimentalmente a gran escala. Nuestro principal objetivo fue intentar generar una metodología computacional capaz de diferenciar dominios sequenciales similares a los priónicos en búsquedas en los proteomas completos de los organismos. Nuestra estrategia se basó en la hipótesis de que sería posible aprender las reglas que determinan la prionogenicidad de una secuencia a partir del análisis de un conjunto lo suficientemente representativo de priones validados experimentalmente. Siguiendo esta idea, generamos un modelo probabilístico de estas regiones basado en su composición aminoacídica y luego lo validamos exhaustivamente para evaluar la capacidad predictiva de nuestro modelo para satisfacer nuestro objetivo principal: identificar todas las proteínas con dominios prionogénicos en el proteoma completo de un organismo. Esto diferenciaría significativamente nuestro método de otros disponibles, que son capaces de evaluar la prionogenicidad de una secuencia determinada pero que no son suficientemente robustos como para hacer búsquedas en grandes bases de datos genómicas. Al satisfacer estos objetivos, realizamos un estudio detallado de todos los proteomas completos disponibles en las bases de datos de secuencias, para predecir posibles proteínas priónicas, y estudiamos las posibles implicaciones de estos resultados para la biología celular desde el punto de vista comparativo a diferentes niveles taxonómicos. Como resultado final de este proyecto, desarrollamos una Base de Datos disponible en Internet en la cual distribuimos libremente nuestras predicciones en todos los proteomas completos, en un formato de fácil interrogación para facilitar el estudio de los procesos mediados por priones a nivel genómico y multigenómico.

Luego, al nivel de estructura tridimensional de proteínas, nos marcamos como objetivo generar un método de predicción de regiones con inestabilidad conformacional basado en las características físico-químicas y geométricas de interfases enterradas en proteínas. La flexibilidad local y global, que median la movilidad cooperativa de las diferentes regiones en las proteínas, son fundamentales en una gran cantidad de procesos celulares como procesos de reconocimiento entre biomoléculas, catálisis enzimática y alosterismo, translocación a través de biomembranas, entre otros. En este segundo objetivo, estábamos interesados en generar un modelo estructural que permitiera identificar regiones conformacionalmente inestables a partir del estudio de la información estructural que es posible extraer de la estructura tridimensional de las proteínas. Nuestra hipótesis era que el interior de las proteínas debería contener regiones con características físico-químicas y geométricas particulares, que facilitaran desde el punto de vista energético la reorganización y la movilidad entre diferentes dominios estructurales. En concordancia, nuestro método descompone sistemáticamente la estructura de una proteína calculando en cada caso una serie de propiedades (razón de polaridad, densidad de empaquetamiento) de la interfaz formada entre un pequeño segmento y el resto de la proteína. La idea de esta aproximación es que las características geométricas y físico-químicas de estas interfases de interacción entre diferentes regiones de la proteína, son las que determinan la fortaleza de dicha interacción y en consecuencia que, en ciertas condiciones, una región determinada adopte conformaciones no nativas con mucha mayor probabilidad. Una vez desarrollado y calibrado el método, estudiamos una gran variedad de proteínas, de diferentes familias estructurales y funcionales, con la intención de identificar este tipo de regiones flexibles. Luego, comparamos nuestros resultados computacionales con una gran cantidad de información estructural obtenida con diversos métodos experimentales, y pudimos comprobar que existía una excelente correspondencia entre nuestras predicciones y las observaciones experimentales. También hicimos un análisis de las posibles implicaciones evolutivas y funcionales de estas características interfases de las proteínas, que podrían estar relacionadas con la evolución de nuevas funciones biológicas manteniendo la misma dinámica conformacional propia del tipo de plegamiento inicial.

Finalmente, también al nivel estructural pero incluyendo la componente temporal, intentamos desarrollar una formulación para predecir fenotipos patológicos causados por Mutaciones de Nucleótido Simple (SNPs) en Enfermedades Conformacionales. Existen una gran cantidad de enfermedades conformacionales causadas por problemas de plegamiento de un grupo muy heterogéneo de proteínas que desempeñan importantes funciones en las células. En este último objetivo estudiamos cómo las mutaciones en una de las proteínas causantes de una enfermedad conformacional, en concreto la Hipercolesterolemia Familiar (FH), pueden estar relacionadas con los diferentes fenotipos patológicos. FH es una enfermedad genética muy importante, la cual afecta aproximadamente al 0.2 % de la población mundial, y que puede en muchos casos provocar la muerte. Esta patología está parcialmente asociada a mutaciones en el gen que codifica el

Receptor de Lipoproteínas de Baja Densidad (LDL-r), una proteína modular transmembranal que juega un papel esencial en la internalización del colesterol en las células. Es conocido que el reconocimiento por parte del receptor de las LDLs está mediado principalmente por un pequeño dominio rico en cisteínas (LA5). Hasta ahora, el número de mutaciones relacionadas con la enfermedad en poblaciones específicas es muy bajo, debido sobre todo a las grandes limitaciones experimentales que implica estudiar todas las posibles mutaciones en el gen codificante de esta proteína, y como afectan su plegamiento, debido su gran tamaño. Por lo tanto, nosotros centramos nuestro estudio en este dominio de interacción y generamos *in silico* todos los posibles mutantes generados por SNPs (227 mutantes diferentes), y hemos analizado en detalle las perturbaciones estructurales provocadas por las mutaciones utilizando Dinámica Molecular. Gracias a esto hemos podido identificar interesantes tendencias en el comportamiento conformacional de los mutantes, a partir de lo cual hemos identificado diferentes grupos de mutaciones más o menos desestabilizantes de la estructura del dominio extracelular del receptor de LDL, lo cual puede estar muy relacionado con el desarrollo de la patología. Esperamos que los resultados obtenidos en este estudio computacional puedan guiar a los experimentalistas para identificar posibles mutaciones patológicas en diferentes enfermedades conformacionales, que puedan ser estudiadas en detalle experimentalmente. Además, estos resultados podrían también contribuir a entender las perturbaciones estructurales causadas por mutaciones específicas, lo que permitiría desarrollar nuevas estrategias para estabilizar proteínas, y para aumentar el conocimiento existente sobre las enfermedades conformacionales.

CHAPTER 1

General Introduction

Contents

1.1 Protein Structure, Conformational Instability and Function

1.1.1 Protein Building Blocks and Structural Elements

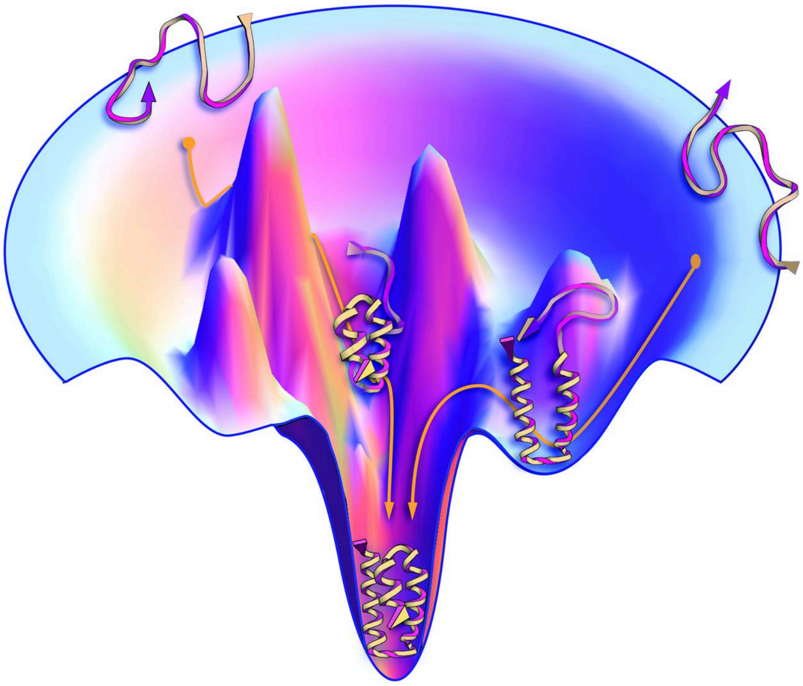
Proteins are the master players of cell biology, acting as facilitators of most of the biological processes that determine cellular homeostasis and responsiveness to environmental stimuli. These biomolecules are built up from a combination of approximately 20 different amino acids which are connected via peptide bonds forming continuous biopolymers, that constitute the building blocks of all the diversity existing in the protein universe. The known proteins coded in the genomes of all the organisms ($\approx 10^{13}$ variants) cover just a minute fraction of the permutational space defined by the amino acid sequences[1], with an upper limit of approximately 10^{469} possible sequence arrangements. However, the great number of possible tridimensional structures that can be generated from the limited sequence variants commonly used by the organisms in nature, can exponentially increase the repertoire of active variants needed to exert all the functions on which life is based upon. This great variability in the structural organization level is generated by the tridimensional array of different local segments –*i.e.* secondary structure– and the relative positioning of secondary structure elements to form the tertiary structure. There are mainly two kinds of secondary structure elements, the helical –*e.g.* 3_{10}-, π-, polyproline II- and α-helices– and the β-strands that form the so-called β-sheets. These structures are generated by distortions in the bond geometry along the polypeptide chain and are stabilized by short-range hydrogen bonding interactions among residues forming the secondary structure element. There are other regions that remain in an extended conformation with few or no internal contacts, called loops. From an organizational point of view, proteins can be decomposed as being formed by secondary structure elements that position to generate supersecondary motifs –*e.g.* $\beta - \alpha - \beta$, Greek key, β-hairpin, etc–, supersecondary structures that array to form domains and domains that form the tertiary structure. This holds for most of the mid-size and large proteins, while other small proteins only contain some of those structural elements. On the other hand, structural organization complexity can be significantly increased because there also exist supramolecular complexes formed by the interactions of proteins determining what is known as the quaternary structure.

The collapse of an unstructured protein sequence in aqueous solution into an organized tridimensional structure is a multifactorial process that depends on a wide group of physicochemical factors. These factors include the hydrophobic, electrostatic and van der Waals interactions, hydrogen bonds, packing, solvation and chain entropy loss. The study of protein stability has proven that the energetic difference between the folded and unfolded states is in average less than 10 $kcal/mol$, thus the cumulative effect of all the factors that contribute to folding must be carefully taken into account. It is well accepted that burying hydrophobic groups in the protein interior occluded from water, and the high number of hydrophobic interactions stablished in consequence in the protein cores, play a major role in protein folding[2–5]. The contribution of van der Waals interactions is also significative, primarily arising and closely related to the hydrophobic effect, in which the tightly packed side chains buried in the protein interior favor the formation of strong interactions[2,3,6]. Hydrogen bond formation is also very important, as it has been demonstrated that side chain-main chain hydrogen bonds among buried polar groups are responsible for a considerable stabilization of the protein architecture, by connecting secondary structural motifs distant in the sequence, and also favoring the increase of the packing density in the protein interior[7–9]. Electrostatic interactions, on the other hand, act more as auxiliary factors[2,3], and solvation contributes to the enthalpic factor of the folding reaction[3]. The packing density at protein cores is also an important factor. In general, the interior of proteins is highly heterogeneous with tightly-packed regions co-existing with others containing packing defects and cavities[10,11], and it is expected that the higher the level of packing resulting after the formation of the native structure, the more stable it will be with respect to the unfolded state. It is also at play an important opposing entropic effect, related to the loss of chain entropy due to the restrictions that the structure scaffold imposes, limiting the number of probable conformations adopted by the main chain and side chains[6,12,13]. The combined contribution of all these factors finally determines the folding reaction of a protein, a complex process not yet fully understood, in which a decreasing number of conformations are explored from among the huge number of possibilities while the protein transits towards the 'native state', as depicted in Figure 1.1. Thus, at the end, the information encoded in the primary sequence determines the specific folding path followed by a protein, as well as the characteristics of the energy landscape –*i.e.* the number of intermediates and the energetic barriers for the transitions between intermediates and/or the native state– which is central on determining the function or functions of the protein.

1.1.2 The 'Structure Determines Function' Dogma

The adoption of protein native structure is one of the most important processes in cell, being the key to determine protein function. At least, this has been the *established belief* accepted for decades among the scientific community. For a long time, as proposed in the 'Central Dogma of Genomics' –*i.e.* structure determines function[14,15]– proteins were regarded as a kind of static species in which each protein adopts a specific tridimensional structure linked to a defined function. Nowadays, however, there has

FIGURE 1.1: Protein Folding Energy Landscape

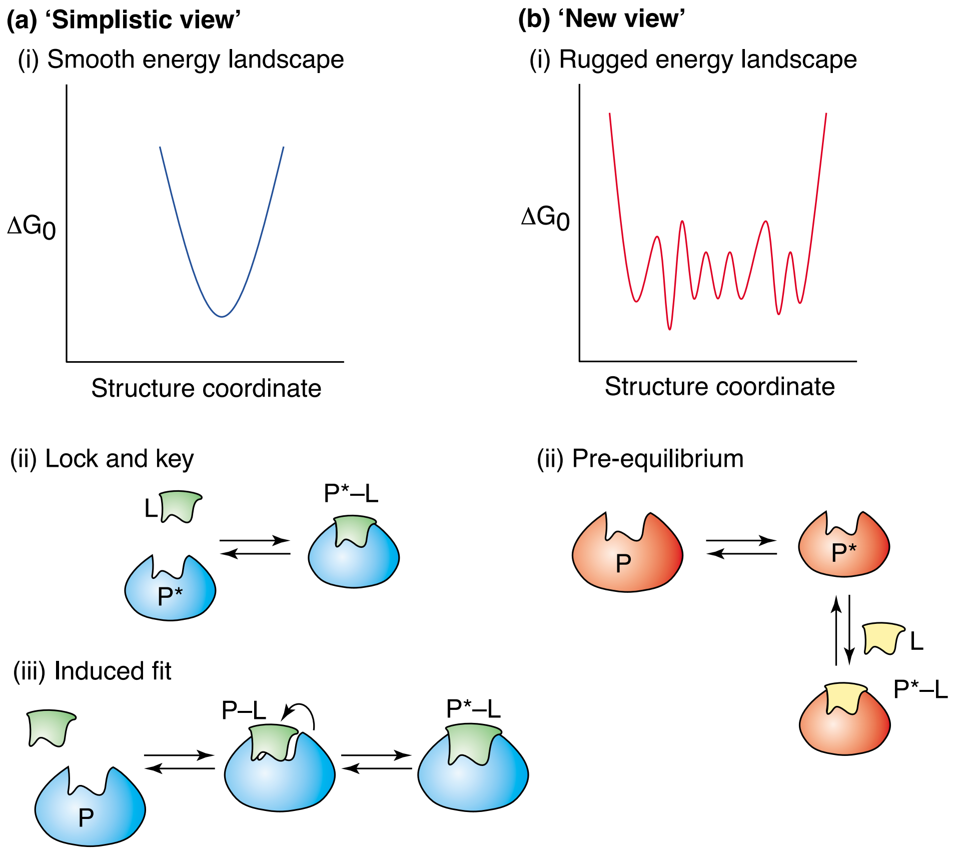


Proteins have a funnel-shaped energy landscape with many high-energy, unfolded structures and only a few low-energy, folded structures. Folding occurs via alternative microscopic trajectories. [Figure taken from: Dill, K.A. and MacCallum, J.L. (2012). *Science*, **338(6110):** 1042–1046[6]]

been accumulating mounting computational and experimental evidence that somehow invalidate this dogma. This evidence includes findings corroborating that proteins exhibit considerable conformational instability, ranging from local fluctuations of specific regions[16–18], to large-scale rearrangements involving partial or global unfolding of the native state[19–21], or proteins that do not adopt any defined structure in isolation –*i.e.* coined as **I**ntrinsically **D**isordered **P**roteins (IDPs)– but rather conform an ensemble of disordered conformations[15,22–25]. In this light, the traditional view that the biological functions of proteins are carried out by single, well-defined conformations is changing to a new scenario in which function would be mediated by ensembles of alternative structures in equilibrium with the 'native state'[26], as proposed in Figure 1.2.

This fact has important functional implications because, as can be conjectured from the traditional view of 'one sequence→one structure→one function' (Figure 1.2, panel a) there would exist a limit for the number of functions depending on the number of proteins encoded in the genome of an organism. Nonetheless, molecular biology is fairly more intricate and embodies far more complexity than can be puzzled out using the limited computational and experimental techniques devised by the human mind.

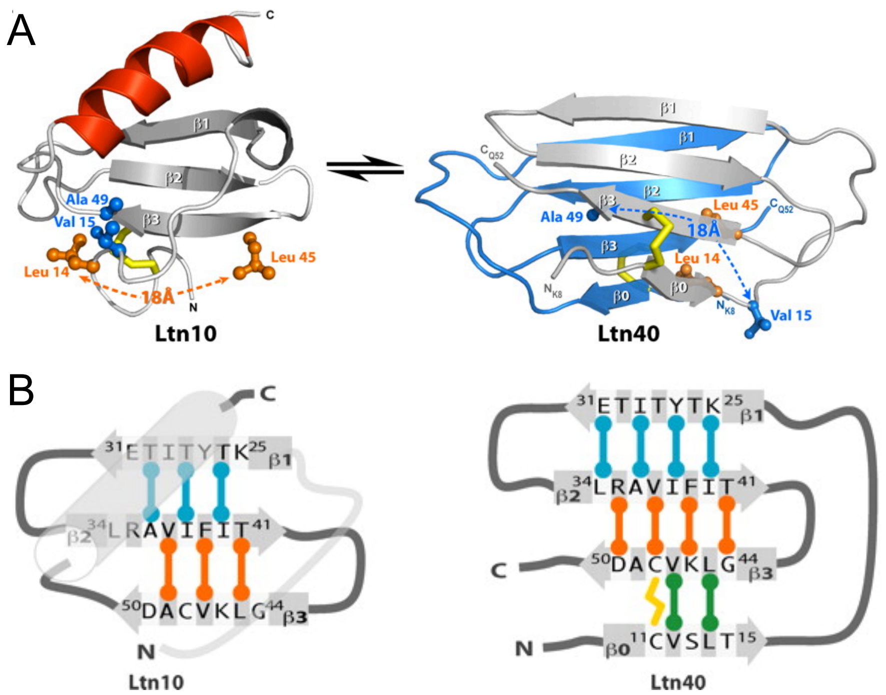
FIGURE 1.2: The 'Simplistic' *vs* 'New View' of Proteins



Schematic energy landscapes and modes of function that represent the 'simplistic' *vs* the 'new view' of proteins. **a**) The 'simplistic' model of proteins describes an energy landscape of a single stable conformer (i) and a function mode of either lock and key (ii) or induced fit (iii). **b**) The 'new view' assumes an ensemble of conformers of similar free energy (i), and a mode of function based on an equilibrium between two (or more) pre-existing isomers, only one of which exerts function (ii). [Figure taken from: James, L.C. and Tawfik, D.S. (2003). *Trends Biochem. Sci.*, **28(7):** 361–8[26]]

The real picture is that proteins have evolved and perfected ways to link the characteristic conformational instability of polypeptides, an expression of which are the rough, intermediate-populated energy landscapes of folding (Figures 1.1 and 1.2, panel b), to increase the complexness of the range of functions and processes that can be mediated by these biomolecules. As a corollary of this more realistic 'new view' of proteins, it is evident that in this scenario, depending on the necessity, the location and the environmental conditions, a given sequence can successfully perform different functions, or interact with different partners, increasing the flexibility and adaptability of the genome.

FIGURE 1.3: Native Lymphoactin (Ltn) Exchanges Between two Unrelated Structures

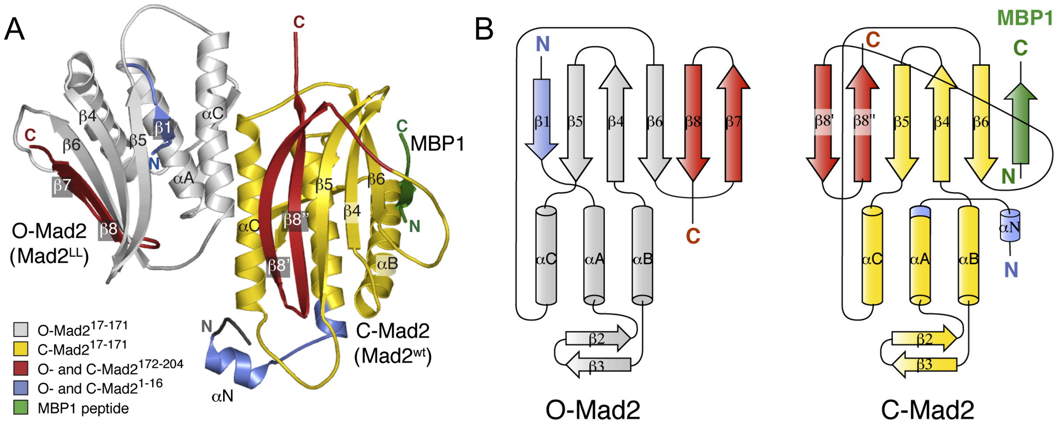


A) Ltn10 ↔ Ltn40 interconversion alters all tertiary contacts. Val^{15} and Ala^{49} pack together in the Ltn10 hydrophobic core (left) but are separated by 18 Å in Ltn40 (Right), whereas the converse is true for Leu^{14} and Leu^{45}. **B**) Rearrangement of hydrogen bonds defining the Ltn secondary structure. Each bar denotes a pair of backbone $N - H \cdots O = C$ hydrogen bonds connecting $\beta 1 - \beta 2$ (cyan), $\beta 2 - \beta 3$ (orange), and $\beta 0 - \beta 3$ (green). Ltn10 ↔ Ltn40 interconversion shifts $\beta 2$ by one residue relative to $\beta 1$ and $\beta 3$, which rotate $180°$ and establish a new hydrogen bond pattern with residues of $\beta 0$ and $\beta 2$. [Figure taken from: Tuinstra, R.L. *et al.* (2008). *Proc. Natl. Acad. Sci. USA*, **105(13):** 5057–62[19]]

1.1.3 Conformational Instability of Polypeptides

The 'one sequence adopts one structure' precept has been revisited. Conformational instability covers a wide range of structural variations, from mild distortions associated with the movement of side chains or loop regions, to more significative cooperative movements of protein subdomains, and constitutes an essential feature of protein evolvability[27,28]. There are multiple examples for some enzymes in which the existence of alternative energetically similar conformers plays a determinant role in catalysis and in the establishment of specific interactions with different partners[16,18,29]. More dramatic global changes are otherwise observed for prions as a result of the conversion between soluble and aggregated forms[18,30–36]. But probably the most eye-catching examples are those involving reversible fold transitions between two well defined conformations with completely different folding architecture, which have been coined as 'metamorphic proteins'[37]. For these proteins, among which the more remarkable examples are Lymphoactin[19] (Figure 1.3) and Mad2[20] (Figure 1.4), in physiologic conditions there exist at least two species with a stability high enough so as to allow their detailed characterization, exhibiting considerable structural differences. Though rather rare and seemingly against all the preconceived ideas of protein structure and folding, these kind of proteins might be much more common that thought, as apparently most structural biology efforts have inadvertently selected against their detection[19]. Indeed, the existence of multiple folded conformations is not prohibited by principles of physics and chemistry[37] and in some cases the selective pressure imposed by function has made it feasible for proteins to evolve to exist in these stable and dissimilar folds. For a special group of proteins that have gained the attention of the scientific community in the last years (IDPs), the case is even more stunning, as they do not adopt any definite structure when unbound in solution[38–40]. This subgroup of proteins, as the one shown in Figure 1.5, are fairly abundant in the genomes of organisms[41] providing crucial functional advantages, and their sequences have been tuned during evolution for circumventing important thermodynamical setbacks –*e.g.* having solvent exposed hydrophobic residues– which imposes an extra effort to the cell for maintaining disorder under control[23].

FIGURE 1.4: Structure of the O-Mad2–C-Mad2 Dimer

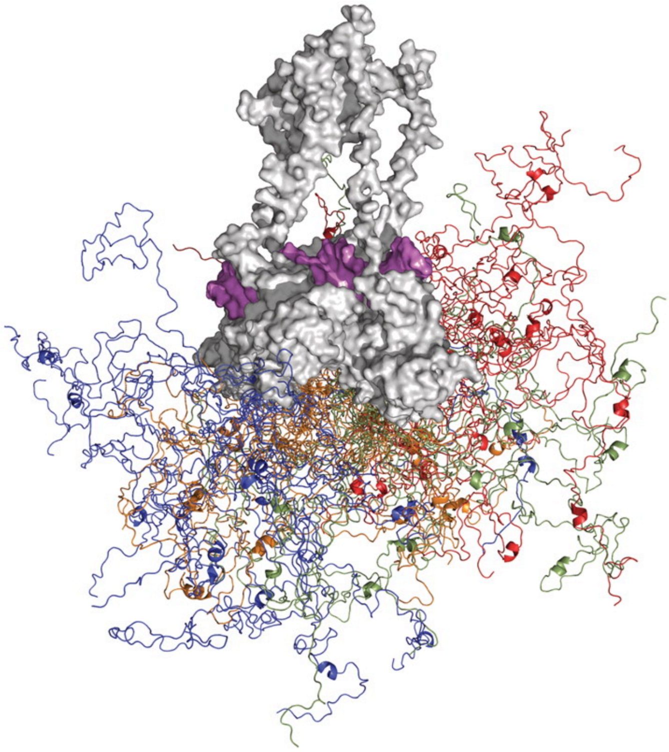


A) Ribbon models of the Mad2 conformational dimer. **B**) Topology diagram of O-Mad2 and C-Mad2. In C-Mad2, the two strands $\beta8' - \beta8''$ are extensions in opposite directions of the $\beta8$ strand of O-Mad2, which justifies the nomenclature $\beta8' - \beta8''$ for these strands [Figure taken from: Mapelli, M. *et al.* (2007). *Cell*, **131(4):** 730–743[20]]

1.1.4 Function Versatility and Multitasking in Proteins

In addition to structural instability, and in some way closely related to it, there are substantial grounds from the 'function' point of view that somehow challenge the 'structure determines function' dogma. There are many examples of proteins that perform

FIGURE 1.5: Model of the N-terminal Domain in the p53-DNA Complex

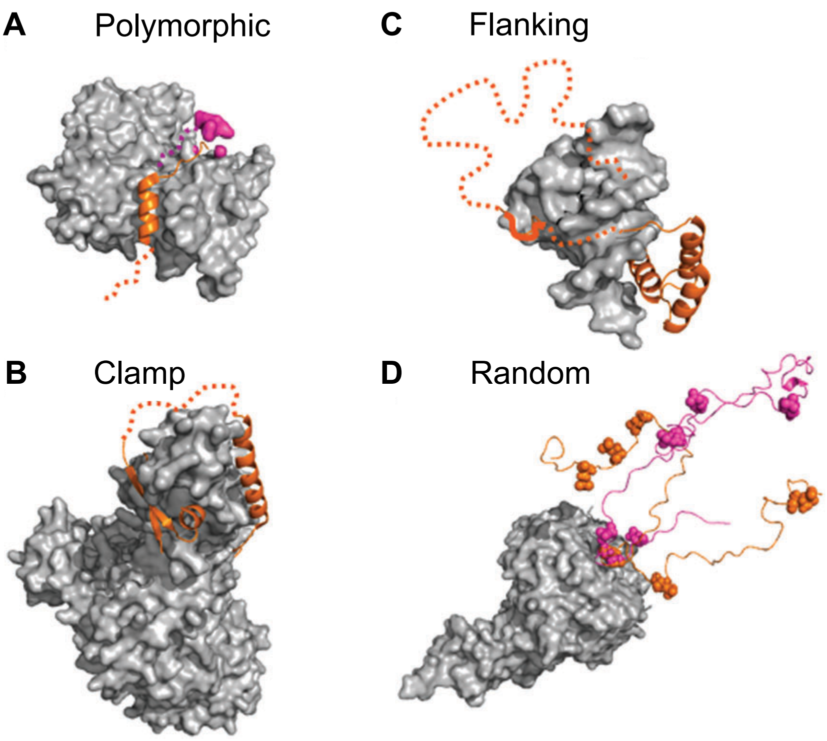


Shown is the N-terminal domain ensemble of one representative full-length p53 molecule included for illustration. p53CTetD (gray) and DNA (magenta) are shown in space fill mode. The flexible C-terminal domain is not shown for reasons of clarity. N-terminal domains forming the four different monomers are shown in different colors for clarity. Twenty copies are shown for each monomer [Figure taken from: Wells, M. *et al.* (2008). *Proc. Natl. Acad. Sci. USA*, **105(15):** 5762–7[42]]

'secondary functions' for which they have not evolved in the first place –*i.e.* coined as moonlighting proteins[43–49]– as is the case of crystallins that are structural constituents of the eye-lens but also have dissimilar enzymatic activities[50,51]. Other cases comprise enzymes of the tRNA synthetases family that additionally to their main function of synthesizing aminoacyl-tRNAs for protein synthesis, also exhibit a wide group of secondary functions, including transcriptional regulation[52–54], regulation of translation and tumor suppression[55–57], angiogenesis[58–60], as secretion proteins mediating proinflammatory responses[61], splicing factors[62,63] and an even larger group of other functions in different organisms groups and taxonomic classifications[64]. In the LDL receptor family, whose members are mainly related to the endocytosis and uptake of lipoproteins, they also participate in an ample group of other processes by interacting with an heterogeneous set of ligands, thanks to structural rearrangements of the recognition domain region[65], and determining their important role in, among others, signal transduction, protein processing and synaptic plasticity[65–67]. Another interesting example is that of **G**lycer**a**ldehyde 3-**p**hosphate **d**ehydrogenase (GAPDH), a glycolytic enzyme with a myriad of diverse functions in cells. In microorganisms, it has been found as structural constituent of the cell walls[68,69]. In mammals, its corresponding orthologs participate as important cellular sensors[70–72], suffering post-translational modifications inducing significant quaternary structural changes leading the trafficking of the protein towards nucleus, mitochondria and other cellular compartments, where it undertakes non-glycolytic functions in DNA repair, RNA binding, phosphorylating activity and interacting with other proteins[70,72–76]. Besides, it has been recently discovered that this enzyme plays an important role in a novel cell death cascade, and is also implicated in some neurodegenerative diseases in man[70,71,73,75,77]. Once again, as explained above for conformational instability, these cases of protein multitasking tough fascinating, are more frequent than commonly expected, and the continuous discovery of examples of moonlighting in the last years has prompted the proposition that this is possibly the mainstream in the protein universe[48].

The connection between conformational instability and multitasking in proteins is a complex problem that we are only beginning to understand. More and more examples are being characterized and apparently there does not appear to be any common structural features among moonlighting proteins[43,45], as it seems that different proteins have evolved dissimilar mechanisms to link instability and function versatility. The first studies of these particular proteins, based on the established structural concepts of ‘one structure-one function’, resulted in the characterization of an ample group of globular proteins, for which the different functions were attributed to specific, well-defined structural *loci*[43,45]. In other cases, more striking structural changes are observed, and different subdomains rearrange to provide the exact topology necessary for binding different chemical compounds or to interact with different partners, as is the case of Lymphoactin[19] (Figure 1.3) and Mad2[20] (Figure 1.4). The extreme case would be IDPs, in which significant disorder provides great potential for forming many interactions via structurally dissimilar complexes –*e.g.* as for one of the most famous and well characterized proteins belonging to this class: p53[42] (Figure 1.5)– since residues that are key for one interaction can often be decoupled from those important for others[78]. Thus, in this type of proteins, the same architecture can adapt and efficiently interact with different or alternative partners, forming the so-called *fuzzy* complexes[38,39], see Figure 1.6. The lack of a defined structure could be seen as the most advanced or even extravagant strategy for linking conformational instability with function versatility. These proteins supply the cells with a set of adaptable machines with promiscuous basal activity, enhanced specificity, surface burial areas and affinity, the possibility to form complex interactions, with an enhanced disposition for regulation via post-translational modifications and proteolysis, higher capture radius for complex formation, among others[38,39].

FIGURE 1.6: Examples of Different Structural Categories of Fuzzy Complexes



The IDPs are shown in orange and magenta, fuzzy regions are represented by dotted lines. The binding partners are displayed as gray surfaces. **A**) Polymorphic complex: the WH2 domain of Wiskott-Aldrich syndrome protein interacts with actin in alternative locations: via an 18 residue segment (orange; PDB id: 2A3Z) or via only 3 residues (magenta; PDB id: 2FF3). **B**) Clamp complex: the nonsense mediated decay factor UPF2 binds to UPF1 via two structured regions (PDB id: 2WJV) and the connecting linker remains ambiguous in the complex (dotted line). **C**) Flanking complex: DNA-binding by the transcription factor Ultrabithorax is strongly influenced by various disordered regions that flank the structured homeodomain. Interactions with another, Extradenticle homeodomain are mediated by a short motif (shown by bold line) located in a clamp-like region (PDB id: 1B8I). **D**) Random complex: the cyclin-dependent kinase inhibitor Sic1 has nine phosphorylation sites that interchange upon contacting Cdc4. Contacts with two of them, T45 and S76, are shown by orange and magenta respectively [Figure taken from: Fuxreiter, M. (2012). *Mol. BioSyst.*, **8(1):** 168–77[39]]

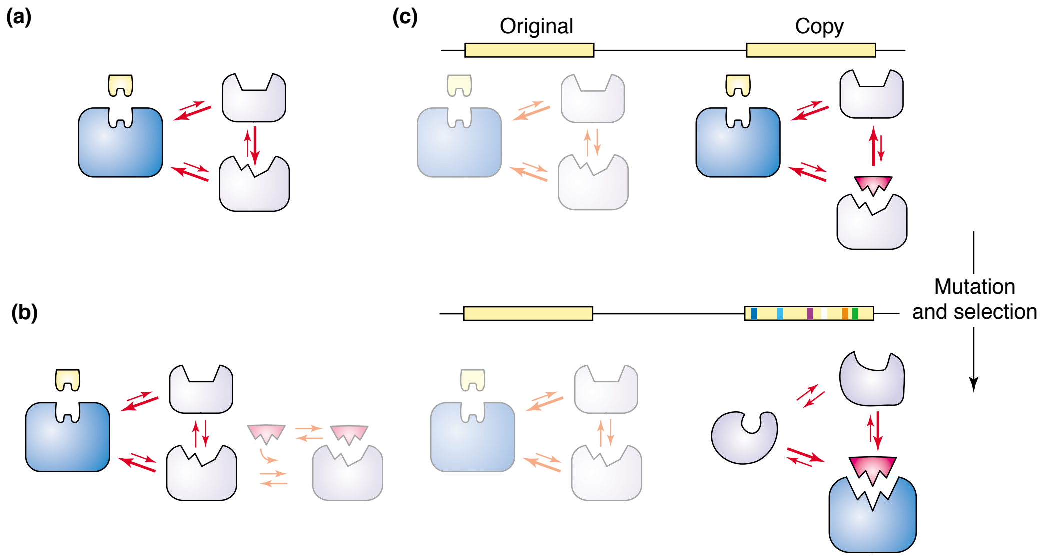
Not surprisingly, IDPs are frequently linked to important regulatory processes, involving the formation of composite supramolecular complexes, related to the most specialized functions in cells, such as gene expression, protein synthesis and cell and tissue differentiation[38–40,43,44,79,80].

1.1.5 Cell Mechanisms for Generating Protein Functional Versatility

Dynamism and conformational variability are intrinsic to polypeptides[29,81], and there exists a complex network of interrelated mechanisms in the cell to take advantage of these physicochemical peculiarities not only to generate proteins that can accomplish a given function with minor structural changes, but also to express proteins with more or less structural instability, which will be better suited for achieving more complex or multiple functions. These mechanisms for generating functional multiplicity and promiscuity span through almost all the processes controlling genome instability, gene and protein expression. New protein alternatives can be generated at the genomic level by mutations –*e.g.* small and/or large-scale mutations[27,82–84]. Gene duplications contribute to the generation of new templates for experimenting new structural scaffolds with alternative affinities and/or functions by means of point mutations. Recent reports argue on the importance of gene duplication on the generation of new protein alternatives, as a significative number of genes in the genomes coexist with their nearly identical duplicates[85,86]. This mechanism allows the exploration of new functionalities and/or specificities taking advantage of the relaxed selection pressure arising from the fact that the other copy can continue fulfilling the original function[27,83,87,88]. This has been observed for the diversification of functions in some protein families[89–91]. In this context, the existence of a conformational ensemble of structures could be determinant in the generation of functional promiscuity, as proposed in the model in Figure 1.7. Mutations in the duplicated gene could improve catalytic efficiency towards the new substrate by optimizing active site chemistry and also by stabilizing the promiscuous conformation, and it is highly likely that the new enzyme will have completely lost the original activity and conformation[26]. In fact, mutation and selection may lead to the new enzyme acquiring new conformations and consequently new promiscuous activities[26].

Although very important for many key biological processes and functions, cells have to deal with important risks of producing a huge amount of these conformationally unstable proteins[23,92]. In accordance, there is a complex and multilevel set of processes for the generation and for the regulation of the expression of these proteins post-transcriptionally by alternative splicing events[93–96], at the translation phase by the processes regulating the initiation, elongation and termination stages[97–101] and by the broad spectrum of post-translational modifications[102–112]. These mechanisms ultimately contribute to generate proteins which can exhibit more or less structural instability, that can simultaneously evolve to perfect their principal functions, while also being able to be involved in other ‘moonlighting’ functions. These secondary functions are organism and/or tissue specific[27,39] and once established, are subject to selection and can be fixed in evolution.

FIGURE 1.7: Proposed Model of Enzyme Evolution Mediated by Conformational Diversity and Functional Promiscuity



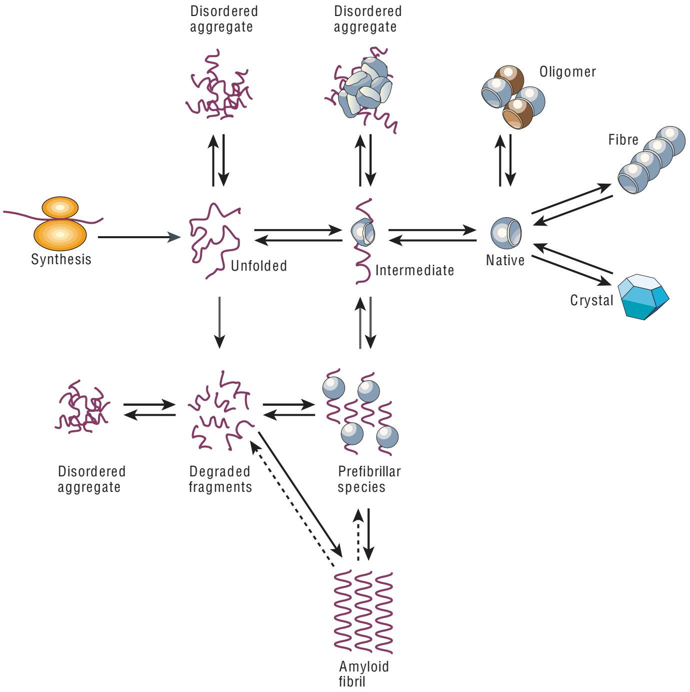
a) The enzyme is in equilibrium between different conformations. The native substrate (yellow) selects the dominant conformer (dark blue) and, thus, enzyme activity confers selective advantage. **b**) An alternative conformation potentiates the binding of a second substrate (pink). The secondary activity confers a limited selective advantage under changing environmental conditions. **c**) Gene duplication enables one copy to evolve improved activity with the promiscuous substrate while the original gene maintains its original function [Figure taken from: James, L.C. and Tawfik, D.S. (2003). *Trends Biochem Sci.*, **28(7):** 361–8[26]]

1.2 Protein Aggregation and Diseases

1.2.1 Amyloidogenesis is a General Property of Polypeptide Chains

The expression of proteins able to experience important structural changes, adopting partially unstructured intermediates in the transition among structural states, or being most of their lifetime partially or totally unfolded in the crowded protein interior entail important downsides, as these species have a high susceptibility to aggregate. Accordingly, although very important for function, proteins' dynamism and structural instability is also related to their susceptibility to aggregate forming amyloid fibrils. Initially, it was thought that amyloidogenesis was a property of a special group of partially unstructured domains, but recent findings suggest that this might be an inherent characteristic of polypeptides, as almost any globular protein can form amyloid fibrils under certain specific conditions[113–119], with independence of its folding architecture, amino acid sequence and molecular weight[32]. This is depicted in Figure 1.8, in a graphical summary of the possible states than can be visited by a protein in the cell[114]. Moreover, it has been proposed that fibril formation appears to be a generically stable structural state of polypeptide chains, competing thermodynamically and kinetically with globular and unfolded monomeric states[32], though the propensity to form such structures can vary dramatically with sequence[120]. Cells have then developed molecular mechanisms to deal with the problem of intracellular protein aggregation, such as the facilitator effect carried out by chaperones[121,122], compartmentalization of aggregates[123] and targeted protein degradation[124].

FIGURE 1.8: A Unified View of some Types of Structures that can be Formed by Polypeptide Chains

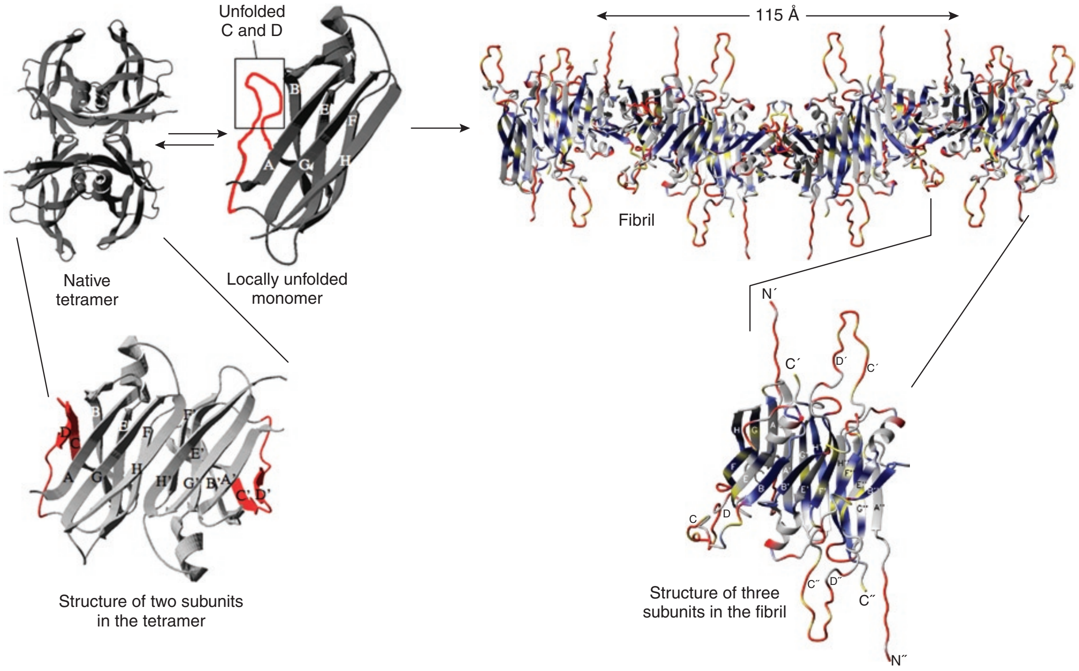


An unstructured chain, for example, newly synthesized on a ribosome, can fold to a monomeric native structure, often through one or more partly folded intermediates. It can, however, experience other fates such as degradation or aggregation. An amyloid fibril is just one form of aggregate, but it is unique in having a highly organized 'misfolded' structure. Other assemblies, including functional oligomers, macromolecular complexes and natural protein fibers, contain natively folded molecules, as do the protein crystals produced *in vitro* for X-ray diffraction studies of their structures. The populations and interconversions of the various states are determined by their relative thermodynamic and kinetic stabilities under any given condition. In living systems, however, transitions between the different states are highly regulated by the environment and by the presence of molecular chaperones, proteolytic enzymes and other factors. Failure of such regulatory mechanisms is likely to be a major factor in the onset and development of misfolding diseases [Figure taken from: Dobson, C.M. (2003). *Nature*, **426(6968):** 884–90[114]]

The formation of amyloid fibrils from short protein stretches has been widely studied for a variety of different models of protein aggregation. From these studies it has been proposed that the process starts with a 'nucleated growth' stage in which the monomers assemble to form which will be the nucleus of the future fibril[125–127]. Then, the subsequent association of monomers and/or oligomers contributes to the growing of the fibril, the velocity of the reaction depending on a wide number of factors, such as solution conditions and the sequence of monomers. This initial stage in reality coincides in time with other advanced stages of fibril formation, as is the case of structured protofibrils, which have a metastable structure and can form on- or off-pathway to the formation of the fibril[128,129]. It has been widely argued that these prefibrillar species mediate the most important part of the cytotoxicity of protein aggregates in cells[130,131]. There are some missing links in the understanding of protein aggregation because the knowledge of these initial stages of amyloid formation has been hampered by the lack of atomistic studies of these prefibrillar species, as it would require techniques able to decipher molecular architectures within ensembles of interconverting and commonly heterogeneous structure[132]. In general, a great diversity of states coexist in time (Figure 1.8), and the formation of the mature fibrils proceeds along a path in which different amyloid species pack together, either by means of considerable structural refolding of the interacting zones or with limited conformational changes[31,133,134]. The sequence of the protein stretches forming the fibrils determines the efficiency of coaggregation, which decreases markedly with the decreasing of the sequence identity of the contacting domains[135]. Besides, the compositional characteristics of the stretches influence the aggregation propensity: **a**) there is a direct proportionality between the hydrophobicity of the aggregation nuclei[136–138], **b**) an indirect proportionality between the charge of the amino acids in the monomer[139,140], **c**) an indirect proportionality between the propensity to form α-helical structures and a direct proportionality between the propensity to form β-sheet structures[136,137,141], and the aggregation potential of the protein stretches.

Protein aggregates can also be formed by well-folded proteins *in vitro* and *in vivo*, a phenomenon different to the one described above for the formation of fibrils from short unfolded peptides. The experimental evidence suggests that aggregation of at least some globular proteins may well be initiated by fluctuations giving rise to the population of amyloidogenic native-like states, without the need to cross the major free energy barrier for unfolding[142]. As depicted in Figure 1.9 for the fibril formation of Transthyretin (TTR), the partial unfolding of the region comprising the peripheral β-strands C and D suffices to determine the region responsible for initiating the amyloid aggregation[143]. The resulting fibril is formed with repeats of TTR molecules positioned longitudinally with respect to the fibril axis connected by the unfolded patches formed in the regions of the C and D β-strands[134,144]. There are plenty of other examples of this behavior in which local conformational changes lead to amyloid formation[145–147]. In the process of fibril formation the constituent molecules of the resulting aggregates may maintain their native-like structure only in the initial oligomers and later undergo a global structural reorganization to form the amyloid structure[142].

FIGURE 1.9: Proposed Process of Fibril Formation for Human Transthyretin (TTR)



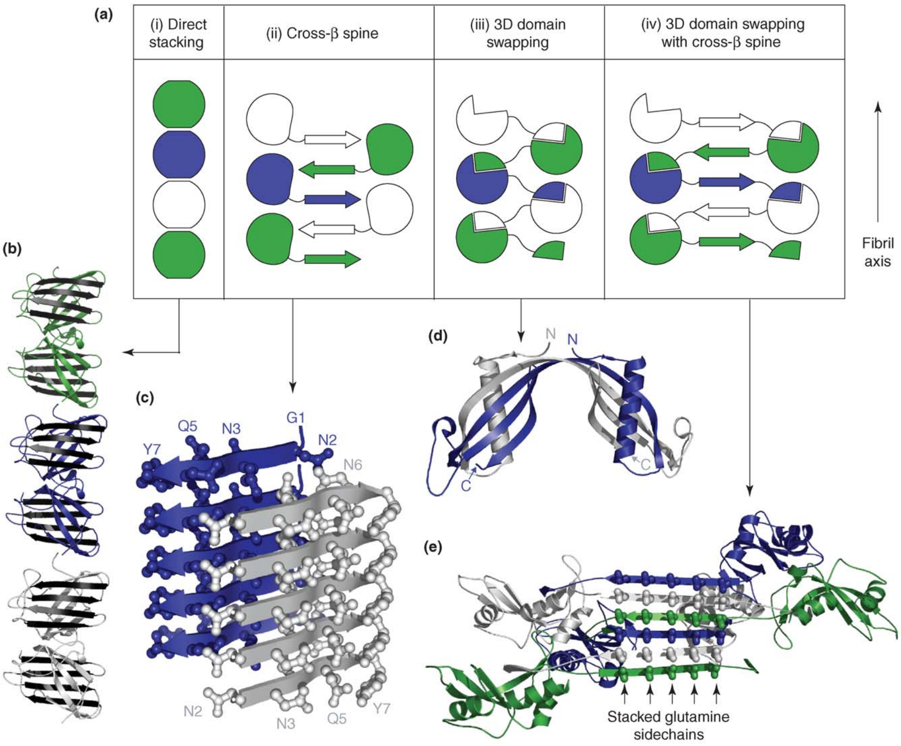
The protein is initially in its native tetrameric form (left, PDB id: 1F41). Two subunits of the tetramer are shown enlarged to illustrate the structural arrangement of the DAGH and CBEF sheets in one subunit and of the D'A'G'H' and C'B'E'F' sheets in the other subunit. The region encompassing β-strands C and D is shown in red in both subunits. Aggregation involves dissociation of the tetramer to form a monomer in which the C and D strands are unfolded. The transition from the native tetramer to the locally unfolded monomer is enhanced by the mutations associated with disease. The region that is unfolded in the monomer is also unfolded in the fibril (right). A section of the fibril is shown enlarged to illustrate the unfolded state of the C and D strands and the orientation of the remaining β-strands. The A and B strands of each subunit form a continuous hydrogen bonded network with the A and B strands of the preceding subunit. The F and H strands of each subunit also form a continuous hydrogen bonded network with the F and H strands of the next subunit (right) [Figure taken from: Chiti, F. and Dobson, C.M. (2009). *Nat. Chem. Biol.*, **5(1):** 15–22[142]]

1.2.2 Amyloid Fibrils Structural Characteristics and Properties

These amyloid aggregates are highly polymorphic, adopting fairly different structural arrangements, and apparently there are no universal molecular structural features in amyloid fibrils except for the cross-β motifs[32]. In Figure 1.10 there is a summary of the possible arrays adopted by proteins in fibrils. In Figure 1.10, panel c, it is included a structural description of the cross-β motif, which is stabilized by hydrophobic interactions at the buried surface between adjacent β-sheets, with the β-strands oriented almost perpendicular to the fibril axis. The polar and charged groups are positioned pointing outwards in the outer surface and perpendicular to the β-sheets, or are establishing pairing interactions with complementary groups in the other sheet when found buried between the paired β-sheets[34,148]. The orientation of the β-sheets can be either parallel or antiparallel and the hydrogen bond registry might also vary significantly between different amyloids[30–36,149–152], rendering structures with characteristic dyeing properties –*e.g.* tincture with Thioflavin T and Congo Red, SDS insolubility and protease resistance. In the case of fibrils formed from the aggregation of specific *loci* from well-folded proteins (Figure 1.10, panels ii and iv, and Figure 1.9), a specific region of the protein establish the interactions via cross-β motif after unfolding, while the rest of the protein retains its native conformation[31,142]. These folded domains could dangle at the sides of the growing spine or they could swap with complementary domains[31]. There is also the possibility of the formation of fibrils mediated by the complementary interactions of folded or partially unfolded regions, as those shown in Figure 1.10, panels i(b) and iii(d).

Prions, a term initially used to refer to infectious misfolded proteins, are a particular kind of amyloids in which the nucleation domains –*i.e.* prionogenic domains in this case– are enriched in glutamine and asparagine residues[141,153]. Some recent reports have underscored the compositional characteristics of these domains, concluding that prion conversion depends significantly on the amino acid composition and the length of such regions[154–156]. There are other factors that also play an important role on determining the prionogenicity of a sequence stretch, such as the number and distribution of prolines and charged residues along the sequence[154]. Interestingly, it has been found by means of exhaustive mutational studies, that prion formation is mainly determined by the amino acid composition of the domain independently of the primary sequence[155,156], a fact that differentiates prions from other amyloids, in which the formation of the cross-β structures during nucleation is highly dependent on hydrophobicity and sequential factors that determine the orientation and hydrogen bond registry of fibrils[32,135]. In general, the tridimensional structure of prions is fairly similar to that of other amyloids, basically stabilized by a cross-β spine. The X-ray structure of the nucleation domain of the yeast prion ***Sup35***[157] and other studies relying in multiple experimental techniques on this same protein[33,158], and the NMR studies of the ***HET*** prion from *Podospora anserina*[159] attest to this fact. Mammalian prions have been more difficult to study from the structural point of view, but some structures have been reported for the prion ***PrP***[160,161], which have resulted in models for the formation of these prions fibrils[162], revealing the great similitudes with that of other amyloids.

FIGURE 1.10: Models of Protein Fibrils



a) Cartoon depicting the four subtypes of gain-of-interaction models. In direct stacking models (panel i), the gained interaction is achieved via simple stacking of subunits. Alternatively, in the cross-β spine models (panel ii), a segment of the protein separates from the core domain to stack into a cross-β spine, with the core domain decorating the edges of the spine. In the somewhat more elaborate model shown in panel iv, the molecules at the edges of the spine domain swap with identical molecules. In the remaining subtype (panel iii), proteins first domain swap and then stack into the fibril. **b**) Ribbon diagram showing a crystalline filament of human superoxide dismutase mutant S134N (PDB id: 1OZU). **c**) Ribbon diagram showing the pair of sheets of the $GNNQQNY$ cross-β spine, with backbones represented by arrows and sidechains by ball-and-stick structures (PDB id: 1YJP). **d**) Ribbon diagram showing the crystal structure of a 3D domain-swapped dimer of human cystatin C (PDB id: 1G96). **e**) Ribbon diagram showing one sheet of the 3D domain-swapped cross-β spine model of fibrillar polyglutamine mutants of RNase A [Figure taken from: Nelson, R. and Eisenberg, D. (2006). *Curr. Opin. Struct. Biol.*, **16(2):** 260–5[31]]

1.2.3 Protein Aggregation Plays Important Roles in Many Cell Processes

The aggregation propensity of proteins has been used by living beings, from microorganisms to human, to carry out important physiological functions, giving them great adaptive advantages in some environmental conditions[163]. Some of the characteristics of amyloid fibrils, such as their resistance to change their properties by the exposure to chemicals and their resistance to protease digestion, as well as their astonishing strength and mechanical stiffness[164], make them a perfect choice to be used in a wide variety of cell functions. In pathogen bacteria, some amyloid proteins –*e.g.* Curlins and Tafi–

are implicated in adhesion to surfaces, cell aggregation and biofilm formation, but also mediate host invasion and pathogenesis through their activation of host extracellular matrix remodeling enzymes[165–167]. In fact, extracellular biofilm formation appears to be a general characteristic among bacteria[168–171], and commonly these complex scaffolds are built up by a diverse group of proteins assembled into amyloid structures[172–174] that are known to be important virulence factors for bacteria, favoring the attachment to eukaryotic cells[168,169,175–178]. These amyloid proteins are indispensable components in biofilms, allowing the formation of the structural framework and interacting with the bacterial cell wall[172–174]. In other free-living bacteria, amyloid forming Chaplins are essential for attachment to hydrophobic surfaces allowing hyphae to escape the aqueous environment and grow into the air, generating an amphipathic film that is an important step in spore formation[166,179,180]. A similar mechanism is used in some fungi, in which special amyloid-forming proteins called Hydrophobins, contribute to the formation of aerial structures for sporulation and adherence to hydrophobic surfaces, in a fairly similar way as Chaplins do[166,179,181].

In multicellular organisms there also exist multiple examples of the beneficial use of amyloid-like structures, as it is the case of one of the most renowned biomaterial: silk. It is now known that silk is formed by amyloidogenic proteins that associate at the spider's spinning duct to generate the nanofibrils that are the constituents of silk fibers[182–185]. As in microorganisms, amyloid structures also take part in protective functions in multicellular organisms[186]. Amyloid structures have been found forming the eggshells of insects, helping the protection of the oocyte and the embryo against the environmental hazards[183,185,186]. In some species of fishes, a similar mechanism is used to protect the eggs from dehydration and other perils of the water medium[187–189]. In some insects and fishes, these kinds of proteins are also responsible for protection against freezing, in a mechanism in which the so called 'antifreeze' or 'thermal hysteresis' proteins, can effectively lower the freezing point of bodily fluids, thereby preventing the formation of microscopic ice crystals[190–192]. In mammals and human, there are also examples of the usefulness of amyloid-like structures, an example of which is the coagulation cascade, in which Fibrin is a key player. Following a group of dissimilar signals the activation of Factor XII triggers a proteolytic cascade that resulted in the formation of Fibrin from the inactive Fibrinogen that polymerize to form a coagulum. It has been shown that Fibrin forms amyloid structures upon polymerization[193], which helps to prevent blood loss and the entrance of infections. The process of melanin biosynthesis in mammals is also mediated by the amyloid protein ***Pmel17***[194,195] that templates and accelerates the covalent polymerization of reactive small molecules into melanin, and also mitigates the toxicity associated with melanin formation by sequestering and minimizing the diffusion of highly reactive and toxic melanin precursors out of the melanosome[194]. It has also been recently proved that in mammals, peptide and protein hormones in secretory granules of the endocrine system are stored in an amyloid-like cross-β sheet-rich conformation, which may explain the processes of granule formation, including hormone selection, membrane surrounding and inert hormone storage, and subsequently the release of hormones from the granules[196].

A special type of amyloids are prions, which have the distinctive properties of acting as heritable elements when in their aggregated forms, constituting self replicating entities that can perpetuate and transmit over generations. Prions are generally ubiquitous proteins with specific functions when folded, that also perform important functions in cells following their amyloid conversion, as is the case of the yeast prion ***Sup35***, a protein that participates in mRNA translation termination. The prion conversion provokes the inactivation of the protein which acts as an epigenetic element, with the subsequent decreasing of the fidelity of translation termination, allowing yeast cells to exploit pre-existing genetic variation to thrive in fluctuating environments[197,198]. Other case is the yeast ***Ure2***, a nitrogen catabolite repressor. When in its soluble form in cells with a good nitrogen source, ***Ure2*** binds to the transcription factor ***Gln3p***, keeping it in the cytoplasm and thereby preventing expression of a set of genes for utilizing poor nitrogen sources[199], that become constitutively expressed if the cell inherited the [***URE3***] prion. Yeast is the organism in which prion biology has been more studied, and there are more examples of prions performing cell functions, like ***Mot3***[200], a transcription factor involved in controlling the cell wall composition and pheromone signaling, the chromatin remodeling factor ***Swi1***[201] and the transcriptional co-repressor ***Cyc8***[202]. In all these cases the prions acting as bet-hedging devices give the organism great reproductive fitness for living in fluctuating environments by creating variant subpopulations with distinct phenotypic states[203–205]. There are also some few examples in multicellular organisms, such as the prion-based generation of durable molecular memory for maintaining long-term physiological states, as it has been proven for prion ***CPEB*** in invertebrates[206–209].

1.2.4 Conformational Diseases and Aggregation Pathologies

Despite its beneficial roles in cell physiology, as described above, protein aggregation is more commonly associated with disease, thanks to the growing number of serious and in some cases incurable pathologies that are being discovered to be caused by the deposition of amyloid fibrils. The formation of intracellular aggregates can be harmful for cells as it promotes the deregulation of the cytosolic stress response because the aggregates, by establishing aberrant protein interactions, sequester a great variety of endogenous multifunctional proteins that occupy essential hub positions in cellular protein networks, with key roles in chromatin organization, transcription, translation, maintenance of cell architecture and protein quality control[210,211]. Besides, the extracellular accumulation of diffusible amyloid oligomers could lead to their non-specific binding to receptors and channel proteins on the synaptic plasma membrane, thus interfering with numerous signal-transduction cascades[212] and seriously affecting the morphology of the neural presynaptic terminals[213]. These diseases can be classified as sporadic ($\approx 80\%$), hereditary ($\approx 15\%$) and transmissible ($\approx 5\%$)[120]. In the group of amyloidoses are included a diverse number of neurodegenerative disorders such as Alzheimer and Parkinson's diseases and various ataxias and dementias, nonneuropathic amyloidoses, either systemic such as Lysozyme and Fibrinogen amyloidoses, or localized suchs as Type II diabetes and Pulmonary alveolar proteinosis[120,212]. In this group of diseases are also included disorders caused by infectious prions in human and mammals like the Creutzfeldt-Jakob disease and bovine spongiform encephalopathy[214,215].

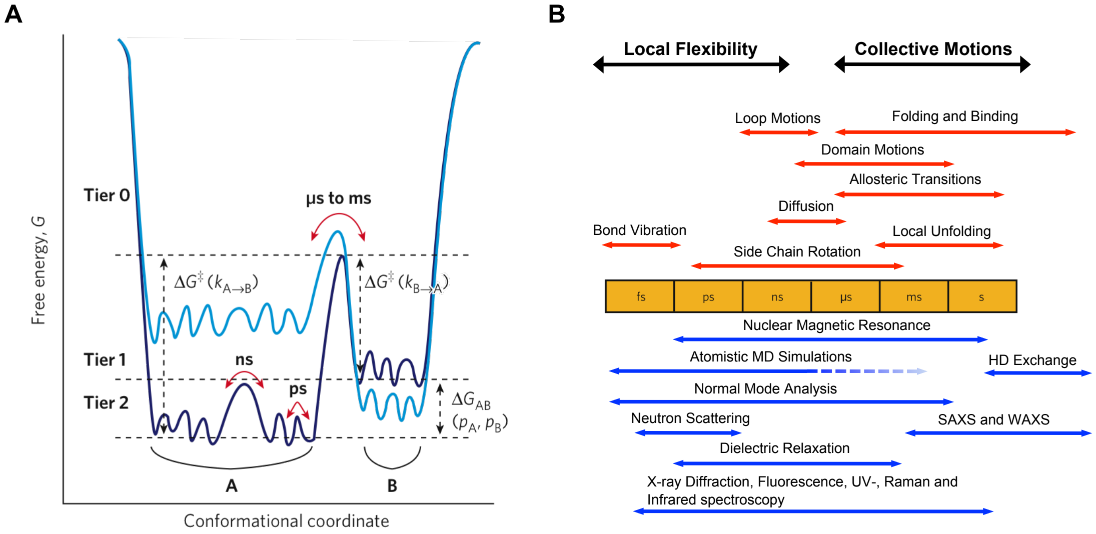
In the last years, the concept of *Conformational Diseases* is gaining great interest from the realization that the perturbation of the equilibrium among a folded protein and its partially unstructured conformers can cause considerable disturbances in cell physiology, and tissue dysfunction in higher organisms, thus leading to diseases. Initially, this concept was proposed to describe a diverse group of disorders in which the abnormal phenotype arises when a constituent protein suffers a transition to a conformation prone to aggregate, resulting in the intra and/or extra-cellular accumulation of amyloid depositions[216–218]. Nevertheless, the concept has evolved to also include pathological states in which an impairment in the folding efficiency results in a reduction in the quantity of the protein that is available to play its normal role[120], or in a reduction of the quality of the protein, when the defective protein even if expressed in sufficient quantities, is unable to correctly carry out its function[219]. These diseases can be inherited, when missense genetic mutations cause alterations in the 3D structure of the proteins[120,219,220], directly affecting their functional sites or indirectly affecting their thermodynamic stability, or acquired when caused by deregulations in the protein expression machinery and protein quality control systems[120,220]. It is known that most mutations compromise protein function, mainly indirectly due to their destabilizing effects[27,84,221], which somehow explains the ever-growing list of conformational diseases, and the importance of studying protein conformational flexibility, how it is related to function and misfolding, and how mutations, chemical compounds and environmental changes could promote or abolish such disorders.

1.3 Experimental Methods for Studying Protein Flexibility and Conformational Instability

1.3.1 Different Timescales of Protein Conformational Motions

In the preceding sections, we have supplied a representative outlook of the intricate problem of protein conformational instability, and its links with protein multitasking and evolvability. As discussed above, there are indisputable and detailed structural evidences of these phenomena, and to no surprise, the inclusion of a fourth dimension into this problem –*i.e.* the time– results in an additional increasing of the complexity. Understanding and modeling these processes to devise experimental and computational methodologies to study them is fairly difficult from an operational point of view. A major obstacle is that it is not possible to watch experimentally individual atoms moving within a protein, but instead, sophisticated biophysical methods are needed to measure the physical properties from which the dynamics can be inferred[222]. As can be seen in Figure 1.11, panel B, conformational motions in proteins occur at different timescales[222,223], depending on the extent of the structural changes, either involving local flexibility or collective motions, which is related to the thermodynamics of the transitions –*i.e.* the free energy (ΔG) of the interconversion, as described in Figure 1.11, panel A. The transitions involving local flexibility (Tier 1 and 2 in Figure 1.11, panel A) correspond to interconversions between structures with marginal energetic differences close to the native basin, and are common *in vivo* and *in vitro* at physiological temperatures. These transitions have been found important for protein function in the case of small solute diffusion in myoglobin[224], in processes of binding and molecular recognition[225–227] and in the ion selectivity of some membrane channels[228,229]. On the other hand, slow motions, such as those corresponding to Tier 0 in Figure 1.11, panel A, involve large collective movements that mediate the conversion between energetically and structurally different species. These more dramatic structural changes are essential in other biological processes such as enzymatic catalysis[230–232] and in fold transitions necessary to set up different structural frameworks for interacting with different partners or substrates[19,20], for details see Figures 1.3 and 1.4. All this wide temporal range can not be covered or addressed by using any single experimental or computational technique (Figure 1.11, pabel B), thus a diverse group of experimental and computational methodologies, based on different physical and chemical principles are used alone or in conjunction to unravel protein conformational motions.

FIGURE 1.11: Timescales of Protein Motions



A) One-dimensional cross-section through the high-dimensional energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. Each Tier is classified following the description introduced in[233]. Lower Tiers describe faster fluctuations between a large number of closely related substates while higher Tiers are related to slower transition between states with significant structural differences. A change in the system will alter the energy landscape (from dark blue to light blue, or vice versa). For example, ligand binding, protein mutation and changes in external conditions shift the equilibrium between states. **B)** Timescale of dynamic processes in proteins (red arows) and the experimental methods (blue arrows) that can detect fluctuations on each timescale. [Figure adapted from: Henzler-Wildman, K. and Kern, D. (2007). *Nature*, **450(7172):** 964–72[222] and Fenwick, R.B. *et al.* (2011). *Eur. Biophys. J.*, **40(12):** 1339–55[223]]

1.3.2 Studying Protein Motions at Low-resolution

Protein motions can be assessed using a group of techniques initially developed in the field of chemistry and latter adapted to the study of the biophysics of biomolecules, including spectroscopic and absorbance methods, fluorescence and vibrational spectroscopy and Electron Paramagnetic Resonance (EPR), among others. With these methods, no detail of all the atoms of the system can be obtained, but instead average information of the entire system or for some specific protein regions. However, these techniques allow the generation of very accurate kinetic information of the transitions and can perform in a wide range of timescales (Figure 1.11, panel B). Some techniques like neutron scattering[234–237] and dielectric spectroscopy[236,238–241], Mössbauer spectroscopy[242], vibrational spectroscopy[243,244] –*i.e.* Raman, resonance Raman, and infrared– and also classical spectroscopy[245], have been widely used to study protein function and its relationship with conformational fluctuations. Also, great insights have been obtained with the use of EPR[246,247]. The partially unfolded intermediates that appear during transitions between states can also be successfully detected by calorimetric and/or spectroscopic techniques[245,248] and the combination of spectroscopic techniques with protein engineering can give some important structural clues at low-resolution of these transient species[249–254], a method that has been coined as Φ-value analysis. This method is based in systematically perturbing the structure of a protein by making single point mutations and assessing the stability of the mutants by spectroscopic techniques. Then the nativeness of the structure in the intermediate around a given residue is associated to a numeric value –*i.e.* the Φ-value. Although laborious, this method is remarkable as it is one of the few techniques available for obtaining structural information of high energy and highly transient species such as transition states. Very recently, with the advent of the possibility of studying single protein molecules, the combination of fluorescence with these kinds of approaches, has opened a new door for the study of protein folding and conformational dynamics[255–257].

1.3.3 Atomic-detail Methodologies for High-resolution Study of Structural Fluctuations

The intense research in the interrelated fields of protein structural fluctuations, prions and amyloidoses, folding intermediates and conformational diseases has generated a wealth of experimental information regarding the structural and compositional determinants of protein motions, and the driving forces of amyloid and prion formation. With these methods it is possible to obtain a snapshot of the different substates visited by the protein during its conformational exploration with atomic or near atomic resolution. A multitude of different NMR techniques[222,258–263] have been used to study protein dynamical properties at different timescales and in different model wild-type or mutant proteins, being perfectly suited for studying the complete conformational ensemble of a protein or a reaction in solution. Small and wide angle X-ray scattering[264–271] can also be used to study all the possible conformations at the same time, but in this case at lower resolutions. Crystallographic methods are also used, although in this case, as it is imperative to obtain a crystal that of course is a homogeneous structure in which the molecules have no significative conformational differences, some biochemical artifices are required to successfully study the different states[232,272]. Time resolved X-ray diffraction crystallography[237,273–279] has constituted an advanced tool for following complex biological processes in real-time, but also to study complex problems inaccessible to most techniques in the femtosecond range, such as the formation and rupture of bonds, the transfer of atoms, ions and electrons among chemical groups during biological processes. Greater insights can be generated with the combined use of different techniques, such as NMR, Hydrogen-Deuterium exchange NMR, Mass Spectrometry Hydrogen-Deuterium exchange, Native State Hydrogen exchange, real-time NMR, pulse-labeling Hydrogen-Deuterium exchange, small angle X-ray scattering[147,280–291], among others in current development.

Amyloid and prion fibrils have been structurally studied using a combination of techniques such as X-ray diffraction[120,149,152,292], Atomic Force Microscopy[293], Transmission Electron Microscopy[35,152,292,294], cryo-Electron Microscopy[35,151], site-directed spin-labeling EPR spectroscopy[30] and solid-state NMR[32,36,120,159,295,296]. From these techniques one ultimately obtains a detailed description of the system at different stages during the conformational transition or aggregate formation, with a delineation of the regions that are important for aggregation, or that fold late or contribute the most to conformation instability, or that fluctuate during the function of a protein, or that are unfolded in the transition intermediates. Consequently, these methodologies are invaluable not only for describing these complex processes, but also to contribute to the generation of computational and theoretical models that can be used synergistically in conjunction with experiments, to get a deeper insight into the study of protein conformational instability.

1.3.4 Structural Characterization of Partially Unfolded Intermediates

Protein conformational instability and local or major structural rearrangements, that as has been explained in the paragraphs above, are key to protein function and versatility, is accomplished through a set of partially unfolded *intermediate species* that have special structural characteristics. During the last years the study of protein folding intermediates has shed some light into the structural characteristics of these transient species from the protein energy landscape, comprising equilibrium intermediates and molten globules[147,280–285,297,298], kinetics intermediates[286–289], and the transition state ensemble[299–304]. These intermediates have been found to be involved in a number of central cellular processes such as membrane translocation[15,305–307] or ligand recognition[308], among others covered in detail in the preceding sections. An important role of partially unfolded intermediates in diseased-related protein aggregation[21,309,310] has also been established. The research of the structural features of folding intermediates would provide important clues regarding the physical forces determining many natural and disease-related phenotypes, and the transitions between different conformations during the function of a protein, which would be of great importance in the development of therapeutic strategies to modify those phenotypes.

1.4 Computational Methods for Complementing Experimental Procedures

1.4.1 Using Primary Sequence and Compositional Characteristics of Amyloid, Intrinsically Disordered and Prion Proteins

In a research field mainly dominated by experimental techniques for many years, computational modeling and bioinformatics are areas in continuous development, since these theoretical strategies have proven to be of great help to describe complex systems with great detail and in a wide range of timescales. Indeed, the predictions and models generated with these methods have been found in many cases fairly consistent and have been subsequently validated experimentally. Thus, theoretical models and bioinformatics are important tools for conducting research in the field of protein biophysics, either in isolation or in combination with experimental procedures. In the specific case of the prediction of amyloid aggregation, the knowledge obtained during years of study of the biogenesis of fibrils *in vitro* and *in vivo* –*e.g.* mutational studies, assessment of the fibril formation rate and structural studies of fibrils– has set the grounds for the development of theoretical models for predicting the amyloid formation propensities of new species[311,312]. The first methodologies developed were based on the essential characteristics of amyloids described in the preceding sections of this chapter, such as hydrophobicity, secondary structure formation propensities and charge[313]. As these characteristics of a given protein stretch or peptide can be modeled at the sequence level, this above mentioned methodology and the group of alternatives and variations that followed, were used to, for example, predicting the effect of a given mutation on the aggregation rate, as a function of the changes in these factors upon mutation[136,313–321], and then validating these predictions experimentally. Structural information has also been used to try to predict amyloidogenesis, with approaches that rely on potential energy calculations for assessing secondary structure patterns[322], on energetic calculations based on structural motifs obtained from fibril crystals[323], inter sheet hydrogen-bonding register patterns[324–326], estimation of packing density after amyloid formation[327], or position-specific weight matrices representations of aggregation nuclei obtained from datasets of amyloidogenic peptides tested *in vitro*[328].

Intrinsic Disorder of proteins has also been studied from a theoretical perspective. Initial propositions for theoretically rationalizing the discrimination of disordered and ordered proteins based on primary sequence information[329], were followed by a group of more refined methods that assessed disorder propensities of consecutive protein stretches[330] using information of compositional bias, sequence complexity and physicochemical factors such as hydropathy and charge, combined into a unitary evaluation function using computer learning approaches. Other approximations were based in a reformulation of the initial propositions for the assessment of charge and hydrophobicity as main predictors of disorder[329], including some variations using the mean net charge and hydropathy combined into a two dimensional CH-plot binary predictor[331,332], which considerably increased the classification accuracy. Notwithstanding these initial steps and the insights obtained from their findings, characterizing IDPs was a great challenge, which was evident as more and more examples were discovered that could not be rationalized using the methods available at that time. Thus, new more refined methods were developed based on artificial intelligence algorithms, aiming to learn the general rules in protein sequences that define local disorder propensities[333–335] from datasets of disordered proteins. Other approaches used contact information and packing density to anticipate disorder[336,337], following the rationale that they would be indicators of the balance between conformational entropy –*i.e.* favoring disorder– and native state contacts –*i.e.* favoring folding. And there is a myriad of other methods, using a combination of a varied set of primary sequence information and amino acid physicochemical properties, processed with more or less elaborated learning methodologies[338], that have played a vital role in widening our understanding of protein intrinsic disorder at the genomic scale[41,339,340], and the regulation and possible functional implications of this kind of proteins in cells[79,341,342].

As explained in the preceding sections, prions are a different type of amyloids that, though sharing similitudes with amyloid aggregates such as the structural characteristics of fibrils, have different compositional features, such as an enrichment in glutamine and asparagine[141,153,343] and the specific length of the prionogenic domains[154–156]. In fact, the initial attempts to predict prions using the well-established methods for amyloid prediction described in the paragraphs above in this section[315,319,324,325], yielded rather discrete results, because they were unable to correctly identify putative aggregation domains in these compositionally diverse prion sequences. The primary sequence particularities of prions have been used for developing predictive methodologies that identify prion domains in protein sequences based on the compositional bias towards Q and N[344,345]. In the early XXIst century with the uprising of the first worldwide epidemics of prion-related diseases in mammals, these pioneering predictive works significantly contributed to present a picture of prion biology at a genomic scale, as their propositions suggested that the genomes of organisms encoded a high proportion of proteins that could behave as prions[344,345]. However, until recently, the number of prions experimentally tested were just a few, and prion biology was a challenging and unexplored area of research. Very recently, the compositional characteristics of known prions were used in a groundbreaking study to perform genome wide prediction of prions in the complete proteome of yeast, combined with high-throughput experimental validation of the predictions[346]. From this work, an ample set of new prions were discovered, alongside with another set of highly scoring putative predictions that were false positives and that did not behave as prions neither *in vitro* nor *in vivo*. Another recent study following this line developed a method to determine the prion propensities of different amino acids from the expression of prion ***Sup35p*** mutants libraries, whose prion behavior was tested *in vivo*[154]. Ultimately the model was also used to design *de novo* synthetic prions –*i.e.* sequences not found in nature but artificial combinations of amino acids for generating a protein stretch with a maximized score, as calculated with the amino acid prion propensities described before[154]. In this work they demonstrated that methodologies relying on compositional information performed better than other based on modeling the formation of parallel β-sheets –*e.g.* those used to predict amyloids– and interestingly also proved that all the synthetic prions designed formed curable prions when expressed *in vivo*, and stably propagated over many generations[347].

1.4.2 Methodologies Relying on Simplified Structural Models of Polypeptides

Among these computational approaches, there exists a group of methods based on simplified structural models of proteins, that can be studied in detail from a dynamical point of view using physical and mathematical formulations during the temporal evolution of the system. This group includes Elastic Networks Models (ENM), Normal Mode Analysis (NMA), Gaussian Network Models (GNM), Anisotropic Network Models (ANM), among others[348–354]. These coarse-grained models represent proteins as a network of connected nodes and rely on harmonic representations for modeling the interactions between nodes –*e.g.* physicochemical interaction between residues. Although the formulation of the model is fairly similar for this varied group of methodologies, there are theoretical differences in the mathematical and physical formulations used in each case for treating the network of contacts[29,352], which results in more or less different outcomes depending on the methodology used. Notwithstanding their simplicity, these methodologies can cover an ample range of motion timescales for the study of fairly different biological phenomena, see Figure 1.11, panel B. In accordance, they have been very successful on describing the dynamics of different protein systems for modeling conformational changes[348,350,355], processes of protein-substrate interactions[356,357], the structural transitions mediating the mechanisms of molecular motors[358], functional mechanisms of membrane proteins[349] or even complex systems such as the allosteric transitions in chaperones[359].

1.4.3 Estimation of the Free Energy of Conformational Transitions

There are other methods based on the estimation of the free energy of conformational transitions, that are also useful for addressing protein flexibility and estimate local instability. There is one well known method in this category, known as the COREX algorithm[360–363]. This method is intended to explore the conformational space of a protein starting from its native state, and from that structure an ensemble of locally unfolded regions is built using a window of a given length[362]. This procedure results in a high number of microstates, in which a given protein region is in an open (unfolded) or a closed (folded) conformation. From this ensemble of conformations, each microstate is analyzed to calculate the free energy of unfolding based on estimations of the Solvent Accessible Surface Area (SASA), which are then transformed into a probability map of the susceptibility of each region to be in an unfolded or folded conformation by using Boltzmann potentials[361,363]. Then, by partitioning the ensemble in unfolded/folded microstates, it is possible to estimate the free energy of folding at the residue level, rendering sequence profiles that represent the probability of each residue in the protein to adopt an open conformation. This method, though computationally expensive and relying on a set of assumptions that are difficult to model computationally –*e.g.* generating all the possible unfolded conformations of a given protein stretch– has been very used and with fairly good success to study allosterism[308,364,365] and to predict amide hydrogen exchange rates in proteins[366].

1.4.4 Computational Simulation Methods

One of the most active fields is that of Computational Simulation of biomolecules. A field with more than 50 years of existence and initially developed for modeling small chemical compounds and liquids[*], that however has been gaining great importance in biomolecules modeling. These strategies have proven very useful for studying biological systems, answering some important questions that can not be otherwise clarified with experimental approaches, such as why different parts of a molecule move and how fast, as well as finding the correlations between motions[222] at atomic resolution. Besides of being able to work consistently with the detailed positions of all the atoms in a protein, it is also possible to study these molecules in simplified but biologically relevant conditions, by including solvent, osmolytes, other proteins, substrates, membrane systems, etc[367]. Indeed, instead of being used as a complementary approach to experiments, these methods have been crucial in many occasions to predict and propose molecule motion behaviors that have been successfully used to design new experiments.

[*] The development of this formulation earned the Nobel Prize in Chemistry 2013: `Prize Announcement`

Molecular Dynamics Simulations are based on atomistic, coarse-grained or hybrid models of biomolecules, which are combined with a formalism that allows following the evolution of motion over time by using the laws of classical physics, which has been coined as 'Molecular Mechanics'. With the exception of systems involving quantum mechanics modeling –*i.e.* QM/MM– on which a part of the system is modeled using quantum physics formulations[368–370], and in coarse-grained models[371,372] in which groups of atoms are represented as abstract 'pseudo-atoms', in most cases the majority if not all atoms in the system are described using force fields embodying the Newtonian Physics Laws, as shown in the following expression:

$$\begin{aligned}\vartheta_{(r^N)} = \sum_{bonds}\frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles}\frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions}\frac{V_n}{2}(1+\cos(\eta\omega - \gamma)) \\ + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right)\end{aligned} \tag{1.1}$$

in which, $\vartheta_{(r^N)}$ is the potential energy as a function of the position of the atoms in the system, as taken from[373]. In this formula, all the contributions of the different types of atomic motions and interactions that determine the evolution of the system over time are contained, including bonded –*e.g.* bond stretching, angle bending and bond rotations– (first three addends in Equation (1.1)) and non-bonded interactions –*e.g.* electrostatic and van der Waals– (last addend in Equation (1.1)). This expression is the main ingredient of force fields, which are parameterized from experimental data, *ab initio* calculations or a combination, and exist different types each determining different dynamics behaviors and suited to model different kinds of molecules[374,375].

The use of molecular dynamics in the study of biomolecules was for a long time limited by the size of the system being studied –*i.e.* the number of atoms– and usually the higher limit of simulation time was below the microseconds as of 2007[222]. Nevertheless, in the last years, the emergence of new hardware and software improvements –*e.g.* increasing of parallelization efficiency of simulation packages[376–378] and improvements of force fields[379–382], continuous growing of size and efficiency of computer clusters[†] and generation of special-purpose architectures[383,384] (and groundbreaking innovative world-wide projects such as Folding@Home[‡]), the advent of Graphical Processing Unit (GPU) simulation strategies[385–387]– are significantly contributing to widen the range of processes that can be studied with these methods. Thanks to these hardware and software advances, at present it is fairly feasible to work in the millisecond and tens of

[†] Rankings of Supercomputers: `http://www.top500.org/lists/`

[‡] Available at: `http://folding.stanford.edu/home/`

milliseconds range[388–391] (Figure 1.11, panel B), and it is very likely that in the near future we will be able to surpass this limit[392].

MD approaches have been used to study protein motions and structural instability in multiple contexts. The relationship between conformational transitions and function has been studied for different proteins such as G-proteins[393–395] and kinases[396–399], and in processes of ligand binding[400,401] and the determination of ion selectivity in membrane channels[402–404], among others. The great technical advances in this field, as described in the preceding paragraph, have permitted the study of fairly more complex processes and systems, such as folding and unfolding of some small proteins[388–391,405–409], the exploration of a protein's energy landscape and folding intermediates[410–412], protein synthesis at the ribosome[413,414], the study of crowded cytosol-like systems[415,416] and the function of molecular chaperones[417–419]. One of the most important features of these methods is that they can be efficiently and synergistically combined with experimental data –*e.g.* NMR parameters, Φ-values– to obtain a better description of the systems under study, contributing to explore new points of view and proposing more accurate strategies to be explored experimentally[223,302,367,420–424].

1.4.5 Estimating the Fate of Mutations in Conformational Diseases

Since the arrival of the 'Genomic Era', with tangible technical possibilities of sequencing large quantities of DNA variants and genomes in relatively short periods of time, great effort has been devoted to rationalize the huge amount of data generated from those projects, and try to relate, for example, genetic mutations to impairments in protein function and disease. Even before that the Human Genome Sequencing Project was completed, initial analysis of the data generated led to the surprising discovery that slightest genetic variations –*e.g.* variations in a single DNA position in which different bases are found in a population, coined as Single Nucleotide Polymorphisms (SNPs)– were very frequent in human genes. These polymorphisms account for most of human genetic variations and are the main source of phenotypic differences among individuals[425,426], and also comprise most of the mutations known to be involved in human inherited diseases[427]. This trend is maintained today, as can be inferred from the statistics in the most extensive databases of human variations, OMIM and HGMD[428,429]. Despite the emergence of the Next Generation Sequencing techniques and the possibility to analyze gene variants at the population scale[430,431], these kinds of studies are still inaccessible from an operational point of view. Thus, a group of computational methodologies have been developed to try to predict if a given non-synonymous SNPs (nsSNPs) could affect the function of the protein and be related with a disease phenotype. Even when these theoretical approaches could fail to accurately predict a significative number of cases, they could be of great help by significantly reducing the number of possible variants to be tested experimentally.

Computational methodologies for predicting the effects of nsSNPs on protein function use sequence, structural or evolutionary information alone or in combination. Sequence-based approaches[432–437] work under the assumption that, by studying the sequences of closely related proteins, it is possible to infer deleterious mutations by analyzing the frequency of the substitute amino acid in a given position in the sequence. Thus, if a given amino acid is to be replaced in a position in the sequence, and in the multiple alignment it is rather common at this position, or there is a high frequency of amino acids with similar physicochemical properties, then it is highly likely that this would be a neutral mutation. If this precept does not hold, then it is expected that the substitution will be deleterious under the principle that during evolution the introduction of this amino acid at this position has been negatively selected. In general, these methods start from the sequence of the protein to be studied and build a multiple sequence alignment with homologous proteins extracted from sequence databases, and analyze the relative relevance of each position in the alignment by calculating scores that represent the amino acid variability.

Structural-based methods[84,438–443] work by taking as input a sequence or a structure of a protein and start by modeling/matching the sequence/structure against a database of protein structures. Then, they account for a group of structural factors that might determine the potential effects of the mutation in the structure, such as solvent accessibility, $C\beta$ density, crystallographic B-factors, estimations of the energetic differences upon mutation, turn or helix breaking, among others. Complementary information is also included to increase the classification efficiency, if for example exists previous knowledge of the important residues of the active site or related to substrate or ligand recognition, disulfide bridge forming cysteines, or residues in protein-protein interaction patches[435,437,440,443,444]. Other methods use a combination of sequence and structural information, combined in some cases with machine learning methodologies for selecting the best arrange of structural and sequence factors for maximizing the predictive accuracy[445–447].

Though these methods have significantly contributed to make extensive studies of the effects of mutations in protein function and disease, they have some important drawbacks. Sequence-based methods are completely reliant on the quality of the multiple alignment generated and the homology criteria followed, which are case dependent and arbitrary. Besides, the information in multiple sequence alignments is biased by the homology criteria and the number of sequences known in a given protein family.

This also applies for structured-based methods because the databases for finding homologs based on structure are less informative, and have lower coverages than protein sequence databases. And even when a structure is available, prediction based solely on protein structure can be misleading because the protein's structure is often determined in the isolated context of a crystal, and cannot take into consideration supramolecular interactions[448]. Recently, with the developments in atomistic molecular dynamics, it has begun to be possible to study the effect of mutations directly at the structural level, without previous assumptions or prior evolutionary knowledge, just by exploring the effects of mutations on the structure of a protein running molecular dynamic simulations, as has been recently reported for the $A\beta$ Alzheimer peptide[449,450] and the Low Density Lipoprotein receptor LA5 domain[451].

1.5 Bibliography

[1] Gustavo Caetano-Anollés et al. The origin, evolution and structure of the protein world. *Biochem J*, **417**: 621–37, 2009. (see p. 34)

[2] Ken A Dill et al. The protein folding problem. *Annual review of biophysics*, **37**: 289–316, 2008. (see p. 35)

[3] Robert L Baldwin. Energetics of protein folding. *Journal of Molecular Biology*, **371**: 283–301, 2007. (see p. 35)

[4] C N Pace et al. Forces contributing to the conformational stability of proteins. *FASEB J*, **10**: 75–83, 1996. (see p. 35)

[5] K A Dill. Dominant forces in protein folding. *Biochemistry*, **29**: 7133–55, 1990. (see p. 35)

[6] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, **338**: 1042–6, 2012. (see pp. 35, 36)

[7] Catherine L Worth and Tom L Blundell. On the evolutionary conservation of hydrogen bonds made by buried polar amino acids: the hidden joists, braces and trusses of protein architecture. *BMC Evol Biol*, **10**: 161, 2010. (see p. 35)

[8] Catherine L Worth and Tom L Blundell. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. *Proteins*, **75**: 413–29, 2009. (see p. 35)

[9] David Schell et al. Hydrogen bonding increases packing density in the protein interior. *Proteins*, **63**: 278–82, 2006. (see p. 35)

[10] J Liang and K A Dill. Are proteins well-packed? *Biophys J*, **81**: 751–66, 2001. (see p. 35)

[11] J Tsai et al. The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology*, **290**: 253–66, 1999. (see p. 35)

[12] J A D'Aquino et al. The magnitude of the backbone conformational entropy change in protein folding. *Proteins*, **25**: 143–56, 1996. (see p. 35)

[13] A J Doig and M J Sternberg. Side-chain conformational entropy in protein folding. *Protein Sci*, **4**: 2247–51, 1995. (see p. 35)

[14] G A Petsko. Dog eat dogma. *Genome Biol*, **1**: comment1002.1–1002.2, 2000. (see p. 35)

[15] P E Wright and H J Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, **293**: 321–31, 1999. (see pp. 35, 36, 56)

[16] Oliver F Lange et al. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**: 1471–5, 2008. (see pp. 36, 38)

[17] B K Muralidhara et al. Thermodynamic fidelity of the mammalian cytochrome P450 2B4 active site in binding substrates and inhibitors. *Journal of Molecular Biology*, **377**: 232–45, 2008. (see p. 36)

[18] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci*, **33**: 2–8, 2008. (see pp. 36, 38)

[19] Robbyn L Tuinstra et al. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, **105**: 5057–62, 2008. (see pp. 36, 38, 39, 41, 53)

[20] Marina Mapelli et al. The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint. *Cell*, **131**: 730–43, 2007. (see pp. 36, 39, 41, 53)

[21] Christopher M Dobson. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol*, **15**: 3–16, 2004. (see pp. 36, 56)

[22] Joseph A Marsh, Sarah A Teichmann, and Julie D Forman-Kay. Probing the diverse landscape of protein flexibility and binding. *Current opinion in structural biology*, **22**: 643–50, 2012. (see p. 36)

[23] A Schlessinger et al. Protein disorder–a breakthrough invention of evolution? *Curr Opin Struct Biol*, **21**: 412–8, 2011. (see pp. 36, 39, 43)

[24] M Gerstein, A M Lesk, and C Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**: 6739–49, 1994. (see p. 36)

[25] H Frauenfelder, S G Sligar, and P G Wolynes. The energy landscapes and motions of proteins. *Science*, **254**: 1598–603, 1991. (see p. 36)

[26] Leo C James and Dan S Tawfik. Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends Biochem Sci*, **28**: 361–8, 2003. (see pp. 36, 37, 43, 44)

[27] Misha Soskine and Dan S Tawfik. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*, **11**: 572–82, 2010. (see pp. 38, 43, 52)

[28] Nobuhiko Tokuriki and Dan S Tawfik. Protein dynamism and evolvability. *Science*, **324**: 203–7, 2009. (see p. 38)

[29] Ivet Bahar et al. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys*, **39**: 23–42, 2010. (see pp. 38, 42, 59)

[30] Nathan J Cobb et al. Molecular architecture of human prion protein amyloid: a parallel, in-register beta-structure. *Proc Natl Acad Sci USA*, **104**: 18946–51, 2007. (see pp. 38, 48, 56)

[31] Rebecca Nelson and David Eisenberg. Recent atomic models of amyloid fibril structure. *Current opinion in structural biology*, **16**: 260–5, 2006. (see pp. 38, 46, 48, 49)

[32] Robert Tycko. Molecular structure of amyloid fibrils: insights from solid-state NMR. *Q Rev Biophys*, **39**: 1–55, 2006. (see pp. 38, 44, 45, 47, 48, 56)

[33] Rajaraman Krishnan and Susan L Lindquist. Structural insights into a yeast prion illuminate nucleation and strain diversity. *Nature*, **435**: 765–72, 2005. (see pp. 38, 48)

[34] Andrey V Kajava et al. A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure. *Proc Natl Acad Sci USA*, **101**: 7885–90, 2004. (see pp. 38, 48)

[35] Ulrich Baxa et al. In Sup35p filaments (the [PSI+] prion), the globular C-terminal domains are widely offset from the amyloid fibril backbone. *Mol Microbiol*, **79**: 523–32, 2011. (see pp. 38, 48, 56)

[36] Frank Shewmaker, Reed B Wickner, and Robert Tycko. Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure. *Proc Natl Acad Sci USA*, **103**: 19754–9, 2006. (see pp. 38, 48, 56)

[37] Alexey G Murzin. Biochemistry. Metamorphic proteins. *Science*, **320**: 1725–6, 2008. (see p. 39)

[38] H Jane Dyson. Expanding the proteome: disordered and alternatively folded proteins. *Q Rev Biophys*, **44**: 467–518, 2011. (see pp. 39, 41, 42)

[39] Monika Fuxreiter. Fuzziness: linking regulation to protein dynamics. *Mol Biosyst*, **8**: 168–77, 2012. (see pp. 39, 41–43)

[40] Jianhan Chen. Towards the physical basis of how intrinsic disorder mediates protein function. *Arch Biochem Biophys*, **524**: 123–31, 2012. (see pp. 39, 42)

[41] A K Dunker et al. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*, **11**: 161–71, 2000. (see pp. 39, 58)

[42] Mark Wells et al. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci USA*, **105**: 5762–7, 2008. (see pp. 40, 41)

[43] Constance J Jeffery. Moonlighting proteins–an update. *Mol Biosyst*, **5**: 345–50, 2009. (see pp. 40–42)

[44] Peter Tompa, Csilla Szász, and László Buday. Structural disorder throws new light on moonlighting. *Trends Biochem Sci*, **30**: 484–9, 2005. (see pp. 40, 42)

[45] Constance J Jeffery. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current opinion in structural biology*, **14**: 663–8, 2004. (see pp. 40, 41)

[46] Constance J Jeffery. Multifunctional proteins: examples of gene sharing. *Ann Med*, **35**: 28–35, 2003. (see p. 40)

[47] C J Jeffery. Moonlighting proteins. *Trends Biochem Sci*, **24**: 8–11, 1999. (see p. 40)

[48] Shelley D Copley. Moonlighting is mainstream: paradigm adjustment required. *Bioessays*, **34**: 578–88, 2012. (see pp. 40, 41)

[49] Daphne H E W Huberts and Ida J van der Klei. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta*, **1803**: 520–5, 2010. (see p. 40)

[50] J Horwitz. Alpha-crystallin can function as a molecular chaperone. *Proc Natl Acad Sci USA*, **89**: 10449–53, 1992. (see p. 40)

[51] J Piatigorsky et al. Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc Natl Acad Sci USA*, **85**: 3479–83, 1988. (see p. 40)

[52] Rey-Ting Guo et al. Crystal structures and biochemical analyses suggest a unique mechanism and role for human glycyl-tRNA synthetase in Ap4A homeostasis. *J Biol Chem*, **284**: 28968–76, 2009. (see p. 40)

[53] Yu-Nee Lee et al. The function of lysyl-tRNA synthetase and Ap4A as signaling regulators of MITF activity in FcepsilonRI-activated mast cells. *Immunity*, **20**: 145–51, 2004. (see p. 40)

[54] A Brevet et al. Zinc-dependent synthesis of 5',5'-diadenosine tetraphosphate by sheep liver lysyl- and phenylalanyl-tRNA synthetases. *J Biol Chem*, **257**: 14613–5, 1982. (see p. 40)

[55] Nam Hoon Kwon et al. Dual role of methionyl-tRNA synthetase in the regulation of translation and tumor suppressor activity of aminoacyl-tRNA synthetase-interacting multifunctional protein-3. *Proc Natl Acad Sci USA*, **108**: 19635–40, 2011. (see p. 40)

[56] Jie Jia et al. WHEP domains direct noncanonical function of glutamyl-Prolyl tRNA synthetase in translational control of gene expression. *Mol Cell*, **29**: 679–90, 2008. (see p. 40)

[57] Prabha Sampath et al. Noncanonical function of glutamyl-prolyl-tRNA synthetase: gene-specific silencing of translation. *Cell*, **119**: 195–208, 2004. (see p. 40)

[58] Quansheng Zhou et al. Orthogonal use of a human tRNA synthetase active site to achieve multifunctionality. *Nat Struct Mol Biol*, **17**: 57–61, 2010. (see p. 40)

[59] Yoshiaki Kise et al. A short peptide insertion crucial for angiostatic activity of human tryptophanyl-tRNA synthetase. *Nat Struct Mol Biol*, **11**: 149–56, 2004. (see p. 40)

[60] K Wakasugi and P Schimmel. Two distinct cytokines released from a human aminoacyl-tRNA synthetase. *Science*, **284**: 147–51, 1999. (see p. 40)

[61] Sang Gyu Park et al. Human lysyl-tRNA synthetase is secreted to trigger proinflammatory response. *Proc Natl Acad Sci USA*, **102**: 6356–61, 2005. (see p. 40)

[62] Paul J Paukstelis et al. Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. *Nature*, **451**: 94–7, 2008. (see p. 40)

[63] Seung Bae Rho, Tommie L Lincecum, and Susan A Martinis. An inserted region of leucyl-tRNA synthetase plays a critical role in group I intron splicing. *EMBO J*, **21**: 6874–81, 2002. (see p. 40)

[64] Corinne D Hausmann and Michael Ibba. Aminoacyl-tRNA synthetase complexes: molecular multitasking revealed. *FEMS Microbiol Rev*, **32**: 705–21, 2008. (see p. 40)

[65] J Herz and D K Strickland. LRP: a multifunctional scavenger and signaling receptor. *J Clin Invest*, **108**: 779–84, 2001. (see p. 40)

[66] Anders Nykjaer and Thomas E Willnow. The low-density lipoprotein receptor gene family: a cellular Swiss army knife? *Trends Cell Biol*, **12**: 273–80, 2002. (see p. 40)

[67] B W Howell and J Herz. The LDL receptor gene family: signaling functions during development. *Curr Opin Neurobiol*, **11**: 74–81, 2001. (see p. 40)

[68] Mônica Santiago Barbosa et al. Glyceraldehyde-3-phosphate dehydrogenase of Paracoccidioides brasiliensis is a cell surface protein involved in fungal adhesion to extracellular matrix proteins and interaction with cells. *Infect Immun*, **74**: 382–9, 2006. (see p. 41)

[69] M L Delgado et al. The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins. *Microbiology (Reading, Engl)*, **147**: 411–7, 2001. (see p. 41)

[70] Makoto R Hara, Matthew B Cascio, and Akira Sawa. GAPDH as a sensor of NO stress. *Biochim Biophys Acta*, **1762**: 502–9, 2006. (see p. 41)

[71] Thomas W Sedlak and Solomon H Snyder. Messenger molecules and cell death: therapeutic implications. *JAMA*, **295**: 81–9, 2006. (see p. 41)

[72] M A Sirover. New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochim Biophys Acta*, **1432**: 159–84, 1999. (see p. 41)

[73] Carlos Tristan et al. The diverse functions of GAPDH: views from different subcellular compartments. *Cell Signal*, **23**: 317–23, 2011. (see p. 41)

[74] Elena I Arutyunova et al. Oxidation of glyceraldehyde-3-phosphate dehydrogenase enhances its binding to nucleic acids. *Biochem Biophys Res Commun*, **307**: 547–52, 2003. (see p. 41)

[75] A Sawa et al. Glyceraldehyde-3-phosphate dehydrogenase: nuclear translocation participates in neuronal and nonneuronal cell death. *Proc Natl Acad Sci USA*, **94**: 11669–74, 1997. (see p. 41)

[76] K Meyer-Siegler et al. A human nuclear uracil DNA glycosylase is the 37-kDa subunit of glyceraldehyde-3-phosphate dehydrogenase. *Proc Natl Acad Sci USA*, **88**: 8460–4, 1991. (see p. 41)

[77] De-Maw Chuang, Christopher Hough, and Vladimir V Senatorov. Glyceraldehyde-3-phosphate dehydrogenase, apoptosis, and neurodegenerative diseases. *Annu Rev Pharmacol Toxicol*, **45**: 269–90, 2005. (see p. 41)

[78] Gideon Schreiber and Amy E Keating. Protein binding specificity versus promiscuity. *Current opinion in structural biology*, **21**: 50–61, 2011. (see p. 41)

[79] Recep Colak et al. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput Biol*, **9**: e1003030, 2013. (see pp. 42, 58)

[80] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**: 197–208, 2005. (see p. 42)

[81] Harold A Scheraga, Mey Khalili, and Adam Liwo. Protein-folding dynamics: overview of molecular simulation techniques. *Annu Rev Phys Chem*, **58**: 57–83, 2007. (see p. 42)

[82] Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*, **19**: 596–604, 2009. (see p. 43)

[83] Shimon Bershtein, Korina Goldin, and Dan S Tawfik. Intense neutral drifts yield robust and evolvable consensus proteins. *Journal of Molecular Biology*, **379**: 1029–44, 2008. (see p. 43)

[84] Nobuhiko Tokuriki et al. How protein stability and new functions trade off. *PLoS Comput Biol*, **4**: e1000002, 2008. (see pp. 43, 52, 63)

[85] Michael Lynch. Genomics. Gene duplication and evolution. *Science*, **297**: 945–7, 2002. (see p. 43)

[86] J Zhang. Evolution by gene duplication: an update. *Trends in ecology & evolution*, 2003. (see p. 43)

[87] Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, **9**: 938–50, 2008. (see p. 43)

[88] Amir Aharoni et al. The 'evolvability' of promiscuous protein functions. *Nat Genet*, **37**: 73–6, 2005. (see p. 43)

[89] Xinjiang Cai and David E Clapham. Evolutionary genomics reveals lineage-specific gene loss and rapid evolution of a sperm-specific ion channel complex: CatSpers and CatSperbeta. *PLoS ONE*, **3**: e3569, 2008. (see p. 43)

[90] Xinjiang Cai. Molecular evolution and structural analysis of the Ca(2+) release-activated Ca(2+) channel subunit, Orai. *Journal of Molecular Biology*, **368**: 1284–91, 2007. (see p. 43)

[91] David McNally and Mario A Fares. In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae. *BMC Evol Biol*, **7**: 81, 2007. (see p. 43)

[92] Jörg Gsponer et al. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, **322**: 1365–8, 2008. (see p. 43)

[93] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual review of biophysics*, **37**: 215–46, 2008. (see p. 43)

[94] Javier F Cáceres and Alberto R Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*, **18**: 186–93, 2002. (see p. 43)

[95] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nat Genet*, **30**: 13–9, 2002. (see p. 43)

[96] B R Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, **17**: 100–7, 2001. (see p. 43)

[97] Fátima Gebauer and Matthias W Hentze. Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol*, **5**: 827–35, 2004. (see p. 43)

[98] Olivier Namy et al. Reprogrammed genetic decoding in cellular gene expression. *Mol Cell*, **13**: 157–68, 2004. (see p. 43)

[99] Benjamin P Lewis, Richard E Green, and Steven E Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA*, **100**: 189–92, 2003. (see p. 43)

[100] Christian Touriol et al. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol Cell*, **95**: 169–78, 2003. (see p. 43)

[101] Marilyn Kozak. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**: 1–34, 2002. (see p. 43)

[102] Sudhakaran Prabakaran et al. Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip Rev Syst Biol Med*, **4**: 565–83, 2012. (see p. 43)

[103] George A Khoury, Richard C Baliban, and Christodoulos A Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*, **1**: 2011. (see p. 43)

[104] Chunaram Choudhary et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**: 834–40, 2009. (see p. 43)

[105] Damian F Brennan and David Barford. Eliminylation: a post-translational modification catalyzed by phosphothreonine lyases. *Trends Biochem Sci*, **34**: 108–14, 2009. (see p. 43)

[106] Xiang-Jiao Yang and Edward Seto. Lysine acetylation: codified crosstalk with other posttranslational modifications. *Mol Cell*, **31**: 449–61, 2008. (see p. 43)

[107] Jeffrey A Ubersax and James E Ferrell. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, **8**: 530–41, 2007. (see p. 43)

[108] Michele A Glozak et al. Acetylation and deacetylation of non-histone proteins. *Gene*, **363**: 15–23, 2005. (see p. 43)

[109] Stefan Westermann and Klaus Weber. Post-translational modifications regulate microtubule function. *Nat Rev Mol Cell Biol*, **4**: 938–47, 2003. (see p. 43)

[110] G Manning et al. The protein kinase complement of the human genome. *Science*, **298**: 1912–34, 2002. (see p. 43)

[111] Gerard Manning et al. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci*, **27**: 514–20, 2002. (see p. 43)

[112] C S Walker et al. On a potential global role for vitamin K-dependent gamma-carboxylation in animal systems. Evidence for a gamma-glutamyl carboxylase in Drosophila. *J Biol Chem*, **276**: 7769–74, 2001. (see p. 43)

[113] Gaetano Invernizzi et al. Protein aggregation: mechanisms and functional consequences. *Int J Biochem Cell Biol*, **44**: 1541–54, 2012. (see p. 44)

[114] Christopher M Dobson. Protein folding and misfolding. *Nature*, **426**: 884–90, 2003. (see pp. 44, 45)

[115] T A Pertinhez et al. Amyloid fibril formation by a helical cytochrome. *FEBS Lett*, **495**: 184–6, 2001. (see p. 44)

[116] M Fändrich, M A Fletcher, and C M Dobson. Amyloid fibrils from muscle myoglobin. *Nature*, **410**: 165–6, 2001. (see p. 44)

[117] J C Rochet and P T Lansbury. Amyloid fibrillogenesis: themes and variations. *Current opinion in structural biology*, **10**: 60–8, 2000. (see p. 44)

[118] C M Dobson. Protein misfolding, evolution and disease. *Trends Biochem Sci*, **24**: 329–32, 1999. (see p. 44)

[119] P T Lansbury. Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease. *Proc Natl Acad Sci USA*, **96**: 3342–4, 1999. (see p. 44)

[120] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, **75**: 333–66, 2006. (see pp. 45, 52, 56)

[121] Jason C Young et al. Pathways of chaperone-mediated protein folding in the cytosol. *Nat Rev Mol Cell Biol*, **5**: 781–91, 2004. (see p. 45)

[122] F Ulrich Hartl and Manajit Hayer-Hartl. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, **295**: 1852–8, 2002. (see p. 45)

[123] Jens Tyedmers, Axel Mogk, and Bernd Bukau. Cellular strategies for controlling protein aggregation. *Nat Rev Mol Cell Biol*, **11**: 777–88, 2010. (see p. 45)

[124] Alfred L Goldberg. Protein degradation and protection against misfolded or damaged proteins. *Nature*, **426**: 895–9, 2003. (see p. 45)

[125] Vladimir N Uversky et al. Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma-synucleins. *J Biol Chem*, **277**: 11970–8, 2002. (see p. 46)

[126] Youcef Fezoui and David B Teplow. Kinetic studies of amyloid beta-protein fibril assembly. Differential effects of alpha-helix stabilization. *J Biol Chem*, **277**: 36948–54, 2002. (see p. 46)

[127] T R Serio et al. Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science*, **289**: 1317–21, 2000. (see p. 46)

[128] Ravindra Kodali and Ronald Wetzel. Polymorphism in the intermediates and products of amyloid assembly. *Curr Opin Struct Biol*, **17**: 48–57, 2007. (see p. 46)

[129] Byron Caughey and Peter T Lansbury. Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu Rev Neurosci*, **26**: 267–98, 2003. (see p. 46)

[130] Charles G Glabe and Rakez Kayed. Common structure and toxic function of amyloid oligomers implies a common mechanism of pathogenesis. *Neurology*, **66**: S74–8, 2006. (see p. 46)

[131] Dominic M Walsh and Dennis J Selkoe. Oligomers on the brain: the emerging role of soluble protein aggregates in neurodegeneration. *Protein Pept Lett*, **11**: 213–28, 2004. (see p. 46)

[132] Thomas R Jahn and Sheena E Radford. Folding versus aggregation: polypeptide conformations on competing pathways. *Arch Biochem Biophys*, **469**: 100–17, 2008. (see p. 46)

[133] N Cerdà-Costa et al. Early kinetics of amyloid fibril formation reveals conformational reorganisation of initial aggregates. *Journal of Molecular Biology*, **366**: 1351–63, 2007. (see p. 46)

[134] Ahmed A Serag et al. Arrangement of subunits and ordering of beta-strands in an amyloid sheet. *Nat Struct Biol*, **9**: 734–9, 2002. (see p. 46)

[135] Caroline F Wright et al. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, **438**: 878–81, 2005. (see pp. 46, 48)

[136] Natalia Sánchez de Groot et al. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J*, **273**: 658–68, 2006. (see pp. 46, 57)

[137] Fabrizio Chiti et al. Kinetic partitioning of protein folding and aggregation. *Nat Struct Biol*, **9**: 137–43, 2002. (see p. 46)

[138] D E Otzen, O Kristensen, and M Oliveberg. Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci USA*, **97**: 9907–12, 2000. (see p. 46)

[139] Jason P Schmittschmitt and J Martin Scholtz. The role of protein stability, solubility, and net charge in amyloid fibril formation. *Protein Sci*, **12**: 2374–8, 2003. (see p. 46)

[140] Fabrizio Chiti et al. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc Natl Acad Sci USA*, **99 Suppl 4**: 16419–26, 2002. (see p. 46)

[141] Marcus Fändrich and Christopher M Dobson. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J*, **21**: 5682–90, 2002. (see pp. 46, 48, 58)

[142] Fabrizio Chiti and Christopher M Dobson. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*, **5**: 15–22, 2009. (see pp. 46–48)

[143] Z Lai, W Colón, and J W Kelly. The acid-mediated denaturation pathway of transthyretin yields a conformational intermediate that can self-assemble into amyloid. *Biochemistry*, **35**: 6470–82, 1996. (see p. 46)

[144] Anders Olofsson et al. Probing solvent accessibility of transthyretin amyloid by solution NMR spectroscopy. *J Biol Chem*, **279**: 5699–707, 2004. (see p. 46)

[145] Georgia Plakoutsi et al. Aggregation of the Acylphosphatase from Sulfolobus solfataricus: the folded and partially unfolded states can both be precursors for amyloid formation. *J Biol Chem*, **279**: 14111–9, 2004. (see p. 46)

[146] Jennifer Stine Elam et al. Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS. *Nat Struct Biol*, **10**: 461–7, 2003. (see p. 46)

[147] Denis Canet et al. Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nat Struct Biol*, **9**: 308–15, 2002. (see pp. 46, 56)

[148] Aneta T Petkova, Wai-Ming Yau, and Robert Tycko. Experimental constraints on quaternary structure in Alzheimer's beta-amyloid fibrils. *Biochemistry*, **45**: 498–512, 2006. (see p. 48)

[149] Ulrich Baxa et al. Filaments of the Ure2p prion protein have a cross-beta core structure. *J Struct Biol*, **150**: 170–9, 2005. (see pp. 48, 56)

[150] Nathan A Oyler and Robert Tycko. Absolute structural constraints on amyloid fibrils from solid-state NMR spectroscopy of partially oriented samples. *J Am Chem Soc*, **126**: 4478–9, 2004. (see p. 48)

[151] L C Serpell and J M Smith. Direct visualisation of the beta-sheet structure of synthetic Alzheimer's amyloid. *Journal of Molecular Biology*, **299**: 225–31, 2000. (see pp. 48, 56)

[152] L C Serpell et al. Fiber diffraction of synthetic alpha-synuclein filaments shows amyloid-like cross-beta conformation. *Proc Natl Acad Sci USA*, **97**: 4897–902, 2000. (see pp. 48, 56)

[153] Randal Halfmann et al. Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. *Mol Cell*, **43**: 72–84, 2011. (see pp. 48, 58)

[154] James A Toombs, Blake R McCarty, and Eric D Ross. Compositional determinants of prion formation in yeast. *Molecular and Cellular Biology*, **30**: 319–32, 2010. (see pp. 48, 58, 59)

[155] Eric D Ross et al. Primary sequence independence for prion formation. *Proc Natl Acad Sci USA*, **102**: 12825–30, 2005. (see pp. 48, 58)

[156] Eric D Ross, Ulrich Baxa, and Reed B Wickner. Scrambled prion domains form prions and amyloid. *Molecular and Cellular Biology*, **24**: 7206–13, 2004. (see pp. 48, 58)

[157] Rebecca Nelson et al. Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, **435**: 773–8, 2005. (see p. 48)

[158] Motomasa Tanaka et al. Mechanism of cross-species prion transmission: an infectious conformation compatible with two highly divergent yeast prion proteins. *Cell*, **121**: 49–62, 2005. (see p. 48)

[159] Christiane Ritter et al. Correlation of structural elements and infectivity of the HET-s prion. *Nature*, **435**: 844–8, 2005. (see pp. 48, 56)

[160] Cédric Govaerts et al. Evidence for assembly of prions with left-handed beta-helices into trimers. *Proc Natl Acad Sci USA*, **101**: 8342–7, 2004. (see p. 48)

[161] R Zahn et al. NMR solution structure of the human prion protein. *Proc Natl Acad Sci USA*, **97**: 145–50, 2000. (see p. 48)

[162] Sangho Lee and David Eisenberg. Seeded conversion of recombinant prion protein to a disulfide-bonded oligomer by a reduction-oxidation process. *Nat Struct Biol*, **10**: 725–30, 2003. (see p. 48)

[163] Douglas M Fowler et al. Functional amyloid–from bacteria to humans. *Trends Biochem Sci*, **32**: 217–24, 2007. (see p. 49)

[164] Jeffrey F Smith et al. Characterization of the nanoscale properties of individual amyloid fibrils. *Proc Natl Acad Sci USA*, **103**: 15806–11, 2006. (see p. 49)

[165] Michelle M Barnhart and Matthew R Chapman. Curli biogenesis and function. *Annu Rev Microbiol*, **60**: 131–47, 2006. (see p. 50)

[166] Martijn F B G Gebbink et al. Amyloids–a functional coat for microorganisms. *Nat Rev Microbiol*, **3**: 333–41, 2005. (see p. 50)

[167] Matthew R Chapman et al. Role of Escherichia coli curli operons in directing amyloid fiber formation. *Science*, **295**: 851–5, 2002. (see p. 50)

[168] Luanne Hall-Stoodley and Paul Stoodley. Evolving concepts in biofilm infections. *Cell Microbiol*, **11**: 1034–43, 2009. (see p. 50)

[169] Luanne Hall-Stoodley, J William Costerton, and Paul Stoodley. Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol*, **2**: 95–108, 2004. (see p. 50)

[170] G O'Toole, H B Kaplan, and R Kolter. Biofilm formation as microbial development. *Annu Rev Microbiol*, **54**: 49–79, 2000. (see p. 50)

[171] J W Costerton, P S Stewart, and E P Greenberg. Bacterial biofilms: a common cause of persistent infections. *Science*, **284**: 1318–22, 1999. (see p. 50)

[172] Diego Romero et al. Amyloid fibers provide structural integrity to Bacillus subtilis biofilms. *Proc Natl Acad Sci USA*, **107**: 2230–4, 2010. (see p. 50)

[173] Poul Larsen et al. Amyloid-like adhesins produced by floc-forming and filamentous bacteria in activated sludge. *Appl Environ Microbiol*, **74**: 1517–26, 2008. (see p. 50)

[174] Poul Larsen et al. Amyloid adhesins are abundant in natural biofilms. *Environ Microbiol*, **9**: 3077–90, 2007. (see p. 50)

[175] Jeremy M Yarwood et al. Generation of virulence factor variants in Staphylococcus aureus biofilms. *J Bacteriol*, **189**: 7961–7, 2007. (see p. 50)

[176] Karen E Beenken et al. Global gene expression in Staphylococcus aureus biofilms. *J Bacteriol*, **186**: 4665–84, 2004. (see p. 50)

[177] Jeremy M Yarwood et al. Quorum sensing in Staphylococcus aureus biofilms. *J Bacteriol*, **186**: 1838–50, 2004. (see p. 50)

[178] D G Davies et al. The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science*, **280**: 295–8, 1998. (see p. 50)

[179] Marie A Elliot and Nicholas J Talbot. Building filaments in the air: aerial morphogenesis in bacteria and fungi. *Curr Opin Microbiol*, **7**: 594–601, 2004. (see p. 50)

[180] Dennis Claessen et al. A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils. *Genes Dev*, **17**: 1714–26, 2003. (see p. 50)

[181] J P Mackay et al. The hydrophobin EAS is largely unstructured in solution and functions by forming amyloid-like structures. *Structure*, **9**: 83–91, 2001. (see p. 50)

[182] Ute Slotta et al. Spider silk and amyloid fibrils: a structural comparison. *Macromol Biosci*, **7**: 183–8, 2007. (see p. 50)

[183] S J Hamodrakas, A Hoenger, and V A Iconomidou. Amyloid fibrillogenesis of silkmoth chorion protein peptide-analogues via a liquid-crystalline intermediate phase. *J Struct Biol*, **145**: 226–35, 2004. (see p. 50)

[184] John M Kenney et al. Amyloidogenic nature of spider silk. *Eur J Biochem*, **269**: 4159–63, 2002. (see p. 50)

[185] V A Iconomidou, G Vriend, and S J Hamodrakas. Amyloids protect the silkmoth oocyte and embryo. *FEBS Lett*, **479**: 141–5, 2000. (see p. 50)

[186] Vassiliki A Iconomidou and Stavros J Hamodrakas. Natural protective amyloids. *Curr Protein Pept Sci*, **9**: 291–309, 2008. (see p. 50)

[187] J E Podrabsky, J F Carpenter, and S C Hand. Survival of water stress in annual fish embryos: dehydration avoidance and egg envelope amyloid fibers. *Am J Physiol Regul Integr Comp Physiol*, **280**: R123–31, 2001. (see p. 50)

[188] V A Iconomidou et al. Secondary structure of chorion proteins of the teleostean fish Dentex dentex by ATR FT-IR and FT-Raman spectroscopy. *J Struct Biol*, **132**: 112–22, 2000. (see p. 50)

[189] P Papadopoulou, V Galanopoulos, and S Hamodrakas. Molecular and Supramolecular Architecture of the Salmo gairdneri Proteinaceous Eggshell during Development. *J Struct Biol*, **116**: 399–412, 1996. (see p. 50)

[190] Steffen P Graether and Brian D Sykes. Structural characterization of amyloidotic antifreeze protein fibrils and intermediates. *J Toxicol Environ Health Part A*, **72**: 1030–3, 2009. (see p. 50)

[191] Steffen P Graether and Brian D Sykes. Cold survival in freeze-intolerant insects: the structure and function of beta-helical antifreeze proteins. *Eur J Biochem*, **271**: 3285–96, 2004. (see p. 50)

[192] Steffen P Graether, Carolyn M Slupsky, and Brian D Sykes. Freezing of a fish antifreeze protein results in amyloid fibril formation. *Biophys J*, **84**: 552–7, 2003. (see p. 50)

[193] Onno Kranenburg et al. Tissue-type plasminogen activator is a multiligand cross-beta structure receptor. *Curr Biol*, **12**: 1833–9, 2002. (see p. 50)

[194] Douglas M Fowler et al. Functional amyloid formation within mammalian tissue. *PLoS Biol*, **4**: e6, 2006. (see p. 50)

[195] Joanne F Berson et al. Proprotein convertase cleavage liberates a fibrillogenic fragment of a resident glycoprotein to initiate melanosome biogenesis. *J Cell Biol*, **161**: 521–33, 2003. (see p. 50)

[196] Samir K Maji et al. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science*, **325**: 328–32, 2009. (see p. 51)

[197] Heather L True, Ilana Berlin, and Susan L Lindquist. Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. *Nature*, **431**: 184–7, 2004. (see p. 51)

[198] H L True and S L Lindquist. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, **407**: 477–83, 2000. (see p. 51)

[199] Andreas Brachmann, Ulrich Baxa, and Reed Brendon Wickner. Prion generation in vitro: amyloid of Ure2p is infectious. *EMBO J*, **24**: 3082–92, 2005. (see p. 51)

[200] Daniel L Holmes et al. Heritable remodeling of yeast multicellularity by an environmentally responsive prion. *Cell*, **153**: 153–65, 2013. (see p. 51)

[201] Zhiqiang Du et al. Newly identified prion linked to the chromatin-remodeling factor Swi1 in Saccharomyces cerevisiae. *Nat Genet*, **40**: 460–5, 2008. (see p. 51)

[202] Basant K Patel, Jackie Gavin-Smyth, and Susan W Liebman. The yeast global transcriptional co-repressor protein Cyc8 can propagate as a prion. *Nat Cell Biol*, **11**: 344–9, 2009. (see p. 51)

[203] Randal Halfmann, Simon Alberti, and Susan Lindquist. Prions, protein homeostasis, and phenotypic diversity. *Trends Cell Biol*, **20**: 125–33, 2010. (see p. 51)

[204] Olivier Namy et al. Epigenetic control of polyamines by the prion [PSI+]. *Nat Cell Biol*, **10**: 1069–75, 2008. (see p. 51)

[205] M M Patino et al. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science*, **273**: 622–6, 1996. (see p. 51)

[206] Sven U Heinrich and Susan Lindquist. Protein-only mechanism induces self-perpetuating changes in the activity of neuronal Aplysia cytoplasmic polyadenylation element binding protein (CPEB). *Proc Natl Acad Sci USA*, **108**: 2999–3004, 2011. (see p. 51)

[207] Paromita Banerjee et al. Short- and long-term memory are modulated by multiple isoforms of the fragile X mental retardation protein. *J Neurosci*, **30**: 6782–92, 2010. (see p. 51)

[208] James Shorter and Susan Lindquist. Prions as adaptive conduits of memory and inheritance. *Nat Rev Genet*, **6**: 435–50, 2005. (see p. 51)

[209] Kausik Si, Susan Lindquist, and Eric R Kandel. A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell*, **115**: 879–91, 2003. (see p. 51)

[210] Heidi Olzscha et al. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell*, **144**: 67–78, 2011. (see p. 52)

[211] Monica Bucciantini et al. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**: 507–11, 2002. (see p. 52)

[212] Dennis J Selkoe. Folding proteins in fatal ways. *Nature*, **426**: 900–4, 2003. (see p. 52)

[213] L Mucke et al. High-level neuronal expression of abeta 1-42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. *J Neurosci*, **20**: 4050–8, 2000. (see p. 52)

[214] J Collinge. Prion diseases of humans and animals: their causes and molecular basis. *Annu Rev Neurosci*, **24**: 519–50, 2001. (see p. 52)

[215] S. B. Prusiner, ed. *Prion biology and diseases*. Second Edition. Cold Spring Harbor Monograph Series New York, USA: Cold Spring Harbor Laboratory Press, 2004. (see p. 52)

[216] Damian C Crowther. Familial conformational diseases and dementias. *Hum Mutat*, **20**: 1–14, 2002. (see p. 52)

[217] R R Kopito and D Ron. Conformational disease. *Nat Cell Biol*, **2**: E207–9, 2000. (see p. 52)

[218] R W Carrell and D A Lomas. Conformational disease. *Lancet*, **350**: 134–8, 1997. (see p. 52)

[219] Niels Gregersen, Lars Bolund, and Peter Bross. Protein misfolding, aggregation, and degradation in disease. *Mol Biotechnol*, **31**: 141–50, 2005. (see p. 52)

[220] Virginie Bernier et al. Pharmacological chaperones: potential treatment for conformational diseases. *Trends Endocrinol Metab*, **15**: 222–8, 2004. (see p. 52)

[221] Manel Camps et al. Genetic constraints on protein evolution. *Crit Rev Biochem Mol Biol*, **42**: 313–26, 2007. (see p. 52)

[222] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, **450**: 964–72, 2007. (see pp. 53–55, 60, 61)

[223] R Bryn Fenwick, Santi Esteban-Martín, and Xavier Salvatella. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J*, **40**: 1339–55, 2011. (see pp. 53, 54, 62)

[224] Friedrich Schotte et al. Picosecond time-resolved X-ray crystallography: probing protein function in real time. *J Struct Biol*, **147**: 235–46, 2004. (see p. 53)

[225] Virginia A Jarymowycz and Martin J Stone. Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem Rev*, **106**: 1624–71, 2006. (see p. 53)

[226] Andrew L Lee et al. Temperature dependence of the internal dynamics of a calmodulin-peptide complex. *Biochemistry*, **41**: 13814–25, 2002. (see p. 53)

[227] D Yang and L E Kay. Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. *Journal of Molecular Biology*, **263**: 369–82, 1996. (see p. 53)

[228] Jordan H Chill et al. Measurement of 15N relaxation in the detergent-solubilized tetrameric KcsA potassium channel. *J Biomol NMR*, **36**: 123–36, 2006. (see p. 53)

[229] D A Doyle et al. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**: 69–77, 1998. (see p. 53)

[230] Elan Z Eisenmesser et al. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438**: 117–21, 2005. (see p. 53)

[231] Elan Zohar Eisenmesser et al. Enzyme dynamics during catalysis. *Science*, **295**: 1520–3, 2002. (see p. 53)

[232] I Schlichting et al. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science*, **287**: 1615–22, 2000. (see pp. 53, 55)

[233] A Ansari et al. Protein states and proteinquakes. *Proc Natl Acad Sci USA*, **82**: 5000–4, 1985. (see p. 54)

[234] Jörg Pieper and Gernot Renger. Protein dynamics investigated by neutron scattering. *Photosyn Res*, **102**: 281–93, 2009. (see p. 54)

[235] M Marconi et al. Comparative study of protein dynamics in hydrated powders and in solutions: a neutron scattering investigation. *Chemical Physics*, 2008. (see p. 54)

[236] Alexandra Buchsteiner et al. Relationship between structure, dynamics and function of hydrated purple membrane investigated by neutron scattering and dielectric spectroscopy. *Journal of Molecular Biology*, **371**: 914–23, 2007. (see p. 54)

[237] Fritz G Parak. Proteins in action: the physics of structural fluctuations and conformational changes. *Curr Opin Struct Biol*, **13**: 552–7, 2003. (see pp. 54, 55)

[238] Hans Frauenfelder et al. A unified model of protein dynamics. *Proc Natl Acad Sci USA*, **106**: 5129–34, 2009. (see p. 54)

[239] F Parak et al. Protein dynamics on different timescales. *Journal of non- . . .*, 2006. (see p. 54)

[240] Helén Jansson, Rikard Bergman, and Jan Swenson. Relation between solvent and protein dynamics as studied by dielectric spectroscopy. *J Phys Chem B*, **109**: 24134–41, 2005. (see p. 54)

[241] Adalberto Bonincontro and Gianfranco Risuleo. Dielectric spectroscopy as a probe for the investigation of conformational properties of proteins. *Spectrochim Acta A Mol Biomol Spectrosc*, **59**: 2677–84, 2003. (see p. 54)

[242] S H Chong et al. Dynamical transition of myoglobin in a crystal: comparative studies of X-ray crystallography and Mössbauer spectroscopy. *Eur Biophys J*, **30**: 319–29, 2001. (see p. 54)

[243] R Vogel and F Siebert. Vibrational spectroscopy as a tool for probing protein function. *Curr Opin Chem Biol*, **4**: 518–23, 2000. (see p. 54)

[244] X Hu et al. New light on allostery: dynamic resonance Raman spectroscopy of hemoglobin kempsey. *Biochemistry*, **38**: 3462–7, 1999. (see p. 54)

[245] Javier Sancho. The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques... plus the structure of the equilibrium intermediates at the same time. *Arch Biochem Biophys*, **531**: 4–13, 2013. (see pp. 54, 55)

[246] Olav Schiemann and Thomas F Prisner. Long-range distance determinations in biomacromolecules by EPR spectroscopy. *Q Rev Biophys*, **40**: 1–53, 2007. (see p. 54)

[247] P P Borbat et al. Electron spin resonance in studies of membranes and proteins. *Science*, **291**: 266–9, 2001. (see p. 54)

[248] Christopher M Johnson. Differential scanning calorimetry as a tool for protein folding and stability. *Arch Biochem Biophys*, **531**: 100–9, 2013. (see p. 55)

[249] Luis A Campos et al. Structure of stable protein folding intermediates by equilibrium phi-analysis: the apoflavodoxin thermal intermediate. *Journal of Molecular Biology*, **344**: 239–55, 2004. (see p. 55)

[250] M Jäger et al. The folding mechanism of a beta-sheet: the WW domain. *Journal of Molecular Biology*, **311**: 373–93, 2001. (see p. 55)

[251] M Oliveberg. Characterisation of the transition states for protein folding: towards a new level of mechanistic detail in protein engineering analysis. *Curr Opin Struct Biol*, **11**: 94–100, 2001. (see p. 55)

[252] J M Matthews and A R Fersht. Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry*, **34**: 6805–14, 1995. (see p. 55)

[253] A R Fersht, A Matouschek, and L Serrano. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *Journal of Molecular Biology*, **224**: 771–82, 1992. (see p. 55)

[254] A Matouschek et al. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, **340**: 122–6, 1989. (see p. 55)

[255] Xavier Michalet, Shimon Weiss, and Marcus Jäger. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem Rev*, **106**: 1785–813, 2006. (see p. 55)

[256] Haw Yang et al. Protein conformational dynamics probed by single-molecule electron transfer. *Science*, **302**: 262–6, 2003. (see p. 55)

[257] S Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nat Struct Biol*, **7**: 724–9, 2000. (see p. 55)

[258] Anthony Mittermaier and Lewis E Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, **312**: 224–8, 2006. (see p. 55)

[259] Joanna F Swain and Lila M Gierasch. The changing landscape of protein allostery. *Curr Opin Struct Biol*, **16**: 102–8, 2006. (see p. 55)

[260] Ryo Kitahara, Shigeyuki Yokoyama, and Kazuyuki Akasaka. NMR snapshots of a fluctuating protein structure: ubiquitin at 30 bar-3 kbar. *Journal of Molecular Biology*, **347**: 277–85, 2005. (see p. 55)

[261] Joan J Englander et al. Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proc Natl Acad Sci USA*, **100**: 7057–62, 2003. (see p. 55)

[262] Haripada Maity et al. Protein hydrogen exchange mechanism: local fluctuations. *Protein Sci*, **12**: 153–60, 2003. (see p. 55)

[263] R Ishima and D A Torchia. Protein dynamics from NMR. *Nat Struct Biol*, **7**: 740–3, 2000. (see p. 55)

[264] Veronique Receveur-Brechot and Dominique Durand. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci*, **13**: 55–75, 2012. (see p. 55)

[265] Pau Bernadó and Dmitri I Svergun. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst*, **8**: 151–67, 2012. (see p. 55)

[266] Alexander V Shkumatov et al. Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering. *Proteins*, **79**: 2122–31, 2011. (see p. 55)

[267] Peter Tompa. Unstructural biology coming of age. *Curr Opin Struct Biol*, **21**: 419–25, 2011. (see p. 55)

[268] Pau Bernadó et al. Structure and Dynamics of Ribosomal Protein L12: An Ensemble Model Based on SAXS and NMR Relaxation. *Biophys J*, **98**: 2374–82, 2010. (see p. 55)

[269] Lee Makowski. Characterization of proteins with wide-angle X-ray solution scattering (WAXS). *J Struct Funct Genomics*, **11**: 9–19, 2010. (see p. 55)

[270] Marco Cammarata et al. Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat Methods*, **5**: 881–6, 2008. (see p. 55)

[271] R F Fischetti et al. Wide-angle X-ray solution scattering as a probe of ligand-induced conformational changes in proteins. *Chem Biol*, **11**: 1431–43, 2004. (see p. 55)

[272] Dominique Bourgeois and Antoine Royant. Advances in kinetic protein crystallography. *Current opinion in structural biology*, **15**: 538–47, 2005. (see p. 55)

[273] Andrew Aquila et al. Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Opt Express*, **20**: 2706–16, 2012. (see p. 55)

[274] Marius Schmidt et al. Five-dimensional crystallography. *Acta Crystallogr, A, Found Crystallogr*, **66**: 198–206, 2010. (see p. 55)

[275] Ahmed H Zewail. 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu Rev Phys Chem*, **57**: 65–103, 2006. (see p. 55)

[276] Lin X Chen. Probing transient molecular structures in photochemical processes using laser-initiated time-resolved X-ray absorption spectroscopy. *Annu Rev Phys Chem*, **56**: 221–54, 2005. (see p. 55)

[277] V Srajer et al. Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochemistry*, **40**: 13802–15, 2001. (see p. 55)

[278] U K Genick et al. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science*, **275**: 1471–5, 1997. (see p. 55)

[279] V Srajer et al. Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. *Science*, **274**: 1726–9, 1996. (see p. 55)

[280] Aitziber L Cortajarena, Simon G J Mochrie, and Lynne Regan. Mapping the energy landscape of repeat proteins using NMR-detected hydrogen exchange. *Journal of Molecular Biology*, **379**: 617–26, 2008. (see p. 56)

[281] Patrick L Wintrode et al. An obligatory intermediate controls the folding of the alpha-subunit of tryptophan synthase, a TIM barrel protein. *Journal of Molecular Biology*, **347**: 911–9, 2005. (see p. 56)

[282] Teerapat Rojsajjakul et al. Multi-state unfolding of the alpha subunit of tryptophan synthase, a TIM barrel protein: insights into the secondary structure of the stable equilibrium intermediates by hydrogen exchange mass spectrometry. *Journal of Molecular Biology*, **341**: 241–53, 2004. (see p. 56)

[283] B A Schulman et al. A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nat Struct Biol*, **4**: 630–4, 1997. (see p. 56)

[284] K Kuwajima. The molten globule state of alpha-lactalbumin. *FASEB J*, **10**: 102–9, 1996. (see p. 56)

[285] Y Bai et al. Protein folding intermediates: native-state hydrogen exchange. *Science*, **269**: 192–7, 1995. (see p. 56)

[286] Sara B-M Whittaker et al. NMR analysis of the conformational properties of the trapped on-pathway folding intermediate of the bacterial immunity protein Im7. *Journal of Molecular Biology*, **366**: 1001–15, 2007. (see p. 56)

[287] Chiaki Nishimura, H Jane Dyson, and Peter E Wright. Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *Journal of Molecular Biology*, **355**: 139–56, 2006. (see p. 56)

[288] T L Religa et al. Solution structure of a protein denatured state and folding intermediate. *Nature*, **437**: 1053–6, 2005. (see p. 56)

[289] Hanqiao Feng, Zheng Zhou, and Yawen Bai. A protein folding pathway with multiple folding intermediates at atomic resolution. *Proc Natl Acad Sci USA*, **102**: 5026–31, 2005. (see p. 56)

[290] Sara Ayuso-Tejedor et al. Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. *Journal of Molecular Biology*, **406**: 604–19, 2011. (see p. 56)

[291] Sara Ayuso-Tejedor et al. Design and structure of an equilibrium protein folding intermediate: a hint into dynamical regions of proteins. *Journal of Molecular Biology*, **400**: 922–34, 2010. (see p. 56)

[292] M Sunde and C Blake. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv Protein Chem*, **50**: 123–59, 1997. (see p. 56)

[293] J D Harper, C M Lieber, and P T Lansbury. Atomic force microscopic imaging of seeded fibril formation and fibril branching by the Alzheimer's disease amyloid-beta protein. *Chem Biol*, **4**: 951–9, 1997. (see p. 56)

[294] L C Serpell et al. The protofilament substructure of amyloid fibrils. *Journal of Molecular Biology*, **300**: 1033–9, 2000. (see p. 56)

[295] Christopher P Jaroniec et al. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc Natl Acad Sci USA*, **99**: 16748–53, 2002. (see p. 56)

[296] Aneta T Petkova et al. A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proc Natl Acad Sci USA*, **99**: 16742–7, 2002. (see p. 56)

[297] Haripada Maity et al. Protein folding: the stepwise assembly of foldon units. *Proc Natl Acad Sci USA*, **102**: 4741–6, 2005. (see p. 56)

[298] M Ohgushi and A Wada. 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett*, **164**: 21–4, 1983. (see p. 56)

[299] Xavier Salvatella et al. Determination of the folding transition states of barnase by using PhiI-value-restrained simulations validated by double mutant PhiIJ-values. *Proc Natl Acad Sci USA*, **102**: 12389–94, 2005. (see p. 56)

[300] Tongye Shen et al. Scanning malleable transition state ensembles: comparing theory and experiment for folding protein U1A. *Biochemistry*, **44**: 6433–9, 2005. (see p. 56)

[301] Linda Hedberg and Mikael Oliveberg. Scattered Hammond plots reveal second level of site-specific information in protein folding: phi' (beta++). *Proc Natl Acad Sci USA*, **101**: 7606–11, 2004. (see p. 56)

[302] Emanuele Paci et al. Comparison of the transition state ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with phi value restraints. *Proteins*, **54**: 513–25, 2004. (see pp. 56, 62)

[303] M Vendruscolo et al. Three key residues form a critical contact network in a protein folding transition state. *Nature*, **409**: 641–5, 2001. (see p. 56)

[304] T Ternström et al. From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc Natl Acad Sci USA*, **96**: 14854–9, 1999. (see p. 56)

[305] M P Schwartz, S Huang, and A Matouschek. The structure of precursor proteins during import into mitochondria. *J Biol Chem*, **274**: 12759–64, 1999. (see p. 56)

[306] F G van der Goot et al. A 'molten-globule' membrane-insertion intermediate of the pore-forming domain of colicin A. *Nature*, **354**: 408–10, 1991. (see p. 56)

[307] Katie L Thoren et al. Lethal factor unfolding is the most force-dependent step of anthrax toxin translocation. *Proc Natl Acad Sci USA*, **106**: 21555–60, 2009. (see p. 56)

[308] Vincent J Hilser and E Brad Thompson. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc Natl Acad Sci USA*, **104**: 8311–5, 2007. (see pp. 56, 60)

[309] Vladimir N Uversky and Anthony L Fink. Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta*, **1698**: 131–53, 2004. (see p. 56)

[310] Maksym Tsytlonok and Laura S Itzhaki. The how's and why's of protein folding intermediates. *Arch Biochem Biophys*, **531**: 14–23, 2013. (see p. 56)

[311] Mattia Belli, Matteo Ramazzotti, and Fabrizio Chiti. Prediction of amyloid aggregation in vivo. *EMBO Rep*, **12**: 657–63, 2011. (see p. 57)

[312] Virginia Castillo et al. Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J*, **6**: 674–85, 2011. (see p. 57)

[313] Fabrizio Chiti et al. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**: 805–8, 2003. (see p. 57)

[314] Gian Gaetano Tartaglia and Michele Vendruscolo. The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev*, **37**: 1395–401, 2008. (see p. 57)

[315] Shahin Zibaee et al. A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci*, **16**: 906–18, 2007. (see pp. 57, 58)

[316] Oscar Conchillo-Solé et al. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**: 65, 2007. (see p. 57)

[317] Natalia Sánchez de Groot et al. Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Structural Biology*, **5**: 18, 2005. (see p. 57)

[318] Amol P Pawar et al. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *Journal of Molecular Biology*, **350**: 379–92, 2005. (see p. 57)

[319] Ana-Maria Fernandez-Escamilla et al. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, **22**: 1302–6, 2004. (see pp. 57, 58)

[320] Kateri F Dubay et al. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *Journal of Molecular Biology*, **341**: 1317–26, 2004. (see p. 57)

[321] Gian Gaetano Tartaglia et al. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci*, **13**: 1939–41, 2004. (see p. 57)

[322] Sukjoon Yoon et al. CSSP2: an improved method for predicting contact-dependent secondary structure propensity. *Comput Biol Chem*, **31**: 373–7, 2007. (see p. 57)

[323] Michael J Thompson et al. The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA*, **103**: 4074–8, 2006. (see p. 57)

[324] Allen W Bryan et al. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol*, **5**: e1000333, 2009. (see pp. 57, 58)

[325] Antonio Trovato, Flavio Seno, and Silvio C E Tosatto. The PASTA server for protein aggregation prediction. *Protein engineering, design & selection : PEDS*, **20**: 521–3, 2007. (see pp. 57, 58)

[326] Antonio Trovato et al. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol*, **2**: e170, 2006. (see p. 57)

[327] Oxana V Galzitskaya, Sergiy O Garbuzynskiy, and Michail Yurievich Lobanov. Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol*, **2**: e177, 2006. (see p. 57)

[328] Sebastian Maurer-Stroh et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods*, **7**: 237–42, 2010. (see p. 57)

[329] R J Williams. The conformation properties of proteins in solution. *Biol Rev Camb Philos Soc*, **54**: 389–437, 1979. (see pp. 57, 58)

[330] P Romero, Z Obradovic, and C Kissinger. . . Identifying disordered regions in proteins from amino acid sequence. *Neural Networks*, 1997. (see p. 57)

[331] Jaime Prilusky et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**: 3435–8, 2005. (see p. 58)

[332] V N Uversky, J R Gillespie, and A L Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**: 415–27, 2000. (see p. 58)

[333] Jonathan J Ward et al. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**: 2138–9, 2004. (see p. 58)

[334] Rune Linding et al. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, **31**: 3701–8, 2003. (see p. 58)

[335] David T Jones and Jonathan J Ward. Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53 Suppl 6**: 573–8, 2003. (see p. 58)

[336] Oxana V Galzitskaya, Sergiy O Garbuzynskiy, and Michail Yu Lobanov. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, **22**: 2948–9, 2006. (see p. 58)

[337] Sergiy O Garbuzynskiy, Michail Yu Lobanov, and Oxana V Galzitskaya. To be folded or to be unfolded? *Protein Sci*, **13**: 2871–7, 2004. (see p. 58)

[338] Bo He et al. Predicting intrinsic disorder in proteins: an overview. *Cell Res*, **19**: 929–49, 2009. (see p. 58)

[339] Jessica Walton Chen et al. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res*, **5**: 879–87, 2006. (see p. 58)

[340] J J Ward et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, **337**: 635–45, 2004. (see p. 58)

[341] Yvonne J K Edwards et al. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome Biol*, **10**: R50, 2009. (see p. 58)

[342] Chad Haynes et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*, **2**: e100, 2006. (see p. 58)

[343] Josephine C Dorsman et al. Strong aggregation and increased toxicity of polyleucine over polyglutamine stretches in mammalian cells. *Hum Mol Genet*, **11**: 1487–96, 2002. (see p. 58)

[344] Paul M Harrison and Mark Gerstein. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biol*, **4**: R40, 2003. (see p. 58)

[345] M D Michelitsch and J S Weissman. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc Natl Acad Sci USA*, **97**: 11910–5, 2000. (see p. 58)

[346] Simon Alberti et al. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell*, **137**: 146–58, 2009. (see p. 59)

[347] James A Toombs et al. De novo design of synthetic prion domains. *Proc Natl Acad Sci USA*, **109**: 6519–24, 2012. (see p. 59)

[348] Aqeel Ahmed, Saskia Villinger, and Holger Gohlke. Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins*, **78**: 3341–52, 2010. (see p. 59)

[349] Ivet Bahar et al. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem Rev*, **110**: 1463–97, 2010. (see p. 59)

[350] Lei Yang, Guang Song, and Robert L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys J*, **93**: 920–9, 2007. (see p. 59)

[351] Manuel Rueda, Pablo Chacón, and Modesto Orozco. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**: 565–75, 2007. (see p. 59)

[352] Ivet Bahar and A J Rader. Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol*, **15**: 586–92, 2005. (see p. 59)

[353] A R Atilgan et al. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*, **80**: 505–15, 2001. (see p. 59)

[354] P Doruker, A R Atilgan, and I Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, **40**: 512–24, 2000. (see p. 59)

[355] F Tama and Y H Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng*, **14**: 1–6, 2001. (see p. 59)

[356] Ivet Bahar, Chakra Chennubhotla, and Dror Tobi. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr Opin Struct Biol*, **17**: 633–40, 2007. (see p. 59)

[357] Dror Tobi and Ivet Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci USA*, **102**: 18908–13, 2005. (see p. 59)

[358] Wenjun Zheng and Sebastian Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proc Natl Acad Sci USA*, **100**: 13253–8, 2003. (see p. 59)

[359] Zheng Yang, Peter Májek, and Ivet Bahar. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput Biol*, **5**: e1000360, 2009. (see p. 59)

[360] James O Wrabl et al. The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys Chem*, **159**: 129–41, 2011. (see p. 60)

[361] Vincent J Hilser et al. A statistical thermodynamic model of the protein ensemble. *Chem Rev*, **106**: 1545–58, 2006. (see p. 60)

[362] Jason Vertrees et al. COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures. *Bioinformatics*, **21**: 3318–9, 2005. (see p. 60)

[363] V J Hilser and E Freire. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *Journal of Molecular Biology*, **262**: 756–72, 1996. (see p. 60)

[364] Travis P Schrank, D Wayne Bolen, and Vincent J Hilser. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proc Natl Acad Sci USA*, **106**: 16984–9, 2009. (see p. 60)

[365] H Pan, J C Lee, and V J Hilser. Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. *Proc Natl Acad Sci USA*, **97**: 12020–5, 2000. (see p. 60)

[366] Tong Liu et al. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *J Am Soc Mass Spectrom*, **23**: 43–56, 2012. (see p. 60)

[367] Bruno Rizzuti and Valerie Daggett. Using simulations to provide the framework for experimental protein folding studies. *Arch Biochem Biophys*, **531**: 128–35, 2013. (see pp. 60, 62)

[368] Shina C L Kamerlin, Maciej Haranczyk, and Arieh Warshel. Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pKa, redox reactions and solvation free energies. *J Phys Chem B*, **113**: 1253–72, 2009. (see p. 61)

[369] Demian Riccardi et al. Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J Phys Chem B*, **110**: 6458–69, 2006. (see p. 61)

[370] R Murphy and D Philipp. . . A mixed quantum mechanics/molecular mechanics (QM/MM) method for largescale modeling of chemistry in protein environments. *Journal of Computational . . .*, 2000. (see p. 61)

[371] Marissa G Saunders and Gregory A Voth. Coarse-graining of multiprotein assemblies. *Current opinion in structural biology*, **22**: 144–50, 2012. (see p. 61)

[372] Shoji Takada. Coarse-grained molecular simulations of large biomolecules. *Current opinion in structural biology*, **22**: 130–7, 2012. (see p. 61)

[373] Andrew R. Leach. *Molecular modelling: principles and applications*. Second Edition. Pearson Education, 2001. (see p. 61)

[374] B R Brooks et al. CHARMM: the biomolecular simulation program. *J Comput Chem*, **30**: 1545–614, 2009. (see p. 61)

[375] David A Case et al. The Amber biomolecular simulation programs. *J Comput Chem*, **26**: 1668–88, 2005. (see p. 61)

[376] K Bowers, R Dror, and D Shaw. Zonal methods for the parallel execution of range-limited N-body simulations. *Journal of Computational Physics*, **221**: 303–329, 2007. (see p. 61)

[377] Blake G Fitch et al. "Blue Matter: Strong scaling of molecular dynamics on Blue Gene/L" in: *Computational Science–ICCS 2006*. Springer, 2006. 846–854 (see p. 61)

[378] David E Shaw. A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. *J Comput Chem*, **26**: 1318–28, 2005. (see p. 61)

[379] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophys J*, **100**: L47–9, 2011. (see p. 61)

[380] Kresten Lindorff-Larsen et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**: 1950–8, 2010. (see p. 61)

[381] Viktor Hornak et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, **65**: 712–25, 2006. (see p. 61)

[382] Alexander D Mackerell, Michael Feig, and Charles L Brooks. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*, **25**: 1400–15, 2004. (see p. 61)

[383] D Shaw, R Dror, and J Salmon. . . Millisecond-scale molecular dynamics simulations on Anton. . . . , 2009. (see p. 61)

[384] M Taiji et al. Protein explorer: A petaflops special-purpose computer system for molecular dynamics simulations. . . . , 2003. (see p. 61)

[385] John E Stone et al. GPU-accelerated molecular modeling coming of age. *J Mol Graph Model*, **29**: 116–25, 2010. (see p. 61)

[386] Mark S Friedrichs et al. Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem*, **30**: 864–72, 2009. (see p. 61)

[387] James C Phillips, John E Stone, and Klaus Schulten. "Adapting a message-driven parallel application to GPU-accelerated clusters" in: *High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008. International Conference for*. IEEE 2008. 1–9 (see p. 61)

[388] Vincent A Voelz et al. Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. *J Am Chem Soc*, **134**: 12565–77, 2012. (see p. 62)

[389] Gregory R Bowman, Vincent A Voelz, and Vijay S Pande. Atomistic folding simulations of the five-helix bundle protein (685). *J Am Chem Soc*, **133**: 664–7, 2011. (see p. 62)

[390] David E Shaw et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, **330**: 341–6, 2010. (see p. 62)

[391] Vincent A Voelz et al. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc*, **132**: 1526–8, 2010. (see p. 62)

[392] Thomas J Lane et al. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr Opin Struct Biol*, **23**: 58–65, 2013. (see p. 62)

[393] Daniel M Rosenbaum et al. Structure and function of an irreversible agonist-(2) adrenoceptor complex. *Nature*, **469**: 236–40, 2011. (see p. 62)

[394] Ron O Dror et al. Activation mechanism of the 2-adrenergic receptor. *Proc Natl Acad Sci USA*, **108**: 18684–9, 2011. (see p. 62)

[395] Tod D Romo, Alan Grossfield, and Michael C Pitman. Concerted interconversion between ionic lock substates of the beta(2) adrenergic receptor revealed by microsecond timescale molecular dynamics. *Biophys J*, **98**: 76–84, 2010. (see p. 62)

[396] Anna Berteotti et al. Protein conformational transitions: the closure mechanism of a kinase explored by atomistic simulations. *J Am Chem Soc*, **131**: 244–50, 2009. (see p. 62)

[397] Yibing Shan et al. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proc Natl Acad Sci USA*, **106**: 139–44, 2009. (see p. 62)

[398] Sichun Yang, Nilesh K Banavali, and Benoît Roux. Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Natl Acad Sci USA*, **106**: 3776–81, 2009. (see p. 62)

[399] José D Faraldo-Gómez and Benoît Roux. On the importance of a funneled energy landscape for the assembly and regulation of multidomain Src tyrosine kinases. *Proc Natl Acad Sci USA*, **104**: 13643–8, 2007. (see p. 62)

[400] Ron O Dror et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci USA*, **108**: 13118–23, 2011. (see p. 62)

[401] Yibing Shan et al. How does a drug molecule find its target binding site? *J Am Chem Soc*, **133**: 9181–3, 2011. (see p. 62)

[402] Morten Ø Jensen et al. Principles of conduction and hydrophobic gating in K+ channels. *Proc Natl Acad Sci USA*, **107**: 5833–8, 2010. (see p. 62)

[403] Isaiah T Arkin et al. Mechanism of Na+/H+ antiporting. *Science*, **317**: 799–803, 2007. (see p. 62)

[404] Sergei Yu Noskov, Simon Bernèche, and Benoît Roux. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature*, **431**: 830–4, 2004. (see p. 62)

[405] Kresten Lindorff-Larsen et al. How fast-folding proteins fold. *Science*, **334**: 517–20, 2011. (see p. 62)

[406] Michelle E McCully et al. Refolding the engrailed homeodomain: structural basis for the accumulation of a folding intermediate. *Biophys J*, **99**: 1628–36, 2010. (see p. 62)

[407] Michelle E McCully, David A C Beck, and Valerie Daggett. Microscopic reversibility of protein folding in molecular dynamics simulations of the engrailed homeodomain. *Biochemistry*, **47**: 7079–89, 2008. (see p. 62)

[408] M Marvin Seibert et al. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *Journal of Molecular Biology*, **354**: 173–83, 2005. (see p. 62)

[409] Ryan Day et al. Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *Journal of Molecular Biology*, **322**: 189–203, 2002. (see p. 62)

[410] Tae Yeon Yoo et al. The folding transition state of protein L is extensive with nonnative interactions (and not small and polarized). *Journal of Molecular Biology*, **420**: 220–34, 2012. (see p. 62)

[411] Joerg Gsponer et al. Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proc Natl Acad Sci USA*, **103**: 99–104, 2006. (see p. 62)

[412] Kathryn A Scott et al. Importance of context in protein folding: secondary structural propensities versus tertiary contact-assisted secondary structure formation. *Biochemistry*, **45**: 4153–63, 2006. (see p. 62)

[413] James Gumbart et al. Mechanisms of SecM-mediated stalling in the ribosome. *Biophys J*, **103**: 331–41, 2012. (see p. 62)

[414] Paula M Petrone et al. Side-chain recognition and gating in the ribosome exit tunnel. *Proc Natl Acad Sci USA*, **105**: 16549–54, 2008. (see p. 62)

[415] Harianto Tjong and Huan-Xiang Zhou. The folding transition-state ensemble of a four-helix bundle protein: helix propensity as a determinant and macromolecular crowding as a probe. *Biophys J*, **98**: 2273–80, 2010. (see p. 62)

[416] Sanbo Qin and Huan-Xiang Zhou. Atomistic modeling of macromolecular crowding predicts modest increases in protein folding and binding stability. *Biophys J*, **97**: 12–9, 2009. (see p. 62)

[417] Thomas J Piggot, Richard B Sessions, and Steven G Burston. Toward a detailed description of the pathways of allosteric communication in the GroEL chaperonin through atomistic simulation. *Biochemistry*, **51**: 1707–18, 2012. (see p. 62)

[418] Giorgio Colombo et al. Understanding ligand-based modulation of the Hsp90 molecular chaperone dynamics at atomic resolution. *Proc Natl Acad Sci USA*, **105**: 7976–81, 2008. (see p. 62)

[419] Akashi Ohtaki et al. Structure and molecular dynamics simulation of archaeal prefoldin: the molecular mechanism for binding and recognition of nonnative substrate proteins. *Journal of Molecular Biology*, **376**: 1130–41, 2008. (see p. 62)

[420] Bosco K Ho, David Perahia, and Ashley M Buckle. Hybrid approaches to molecular simulation. *Current opinion in structural biology*, **22**: 386–93, 2012. (see p. 62)

[421] Olivier Fisette et al. Synergistic applications of MD and NMR for the study of biological systems. *J Biomed Biotechnol*, **2012**: 254208, 2012. (see p. 62)

[422] Nicoletta Calosci et al. Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc Natl Acad Sci USA*, **105**: 19241–6, 2008. (see p. 62)

[423] Christopher J Francis et al. Characterization of the residual structure in the unfolded state of the Delta131Delta fragment of staphylococcal nuclease. *Proteins*, **65**: 145–52, 2006. (see p. 62)

[424] Michele Vendruscolo et al. Structures and relative free energies of partially folded states of proteins. *Proc Natl Acad Sci USA*, **100**: 14817–21, 2003. (see p. 62)

[425] M Cargill et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, **22**: 231–8, 1999. (see p. 62)

[426] D G Wang et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**: 1077–82, 1998. (see p. 62)

[427] M Krawczak et al. Human gene mutation database-a biomedical information and research resource. *Hum Mutat*, **15**: 45–51, 2000. (see p. 62)

[428] Joanna Amberger et al. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Research*, **37**: D793–6, 2009. (see p. 62)

[429] Peter D Stenson et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*, **21**: 577–81, 2003. (see p. 62)

[430] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, **5**: 16–8, 2008. (see p. 62)

[431] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet*, **24**: 133–41, 2008. (see p. 62)

[432] Yongwook Choi et al. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, **7**: e46688, 2012. (see p. 63)

[433] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, **4**: 1073–81, 2009. (see p. 63)

[434] Carles Ferrer-Costa, Modesto Orozco, and Xavier De La Cruz. Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. *Proteins*, **61**: 878–87, 2005. (see p. 63)

[435] C Ferrer-Costa, M Orozco, and X De La Cruz. Sequence-based prediction of pathological mutations. *Proteins*, **57**: 811–819, 2004. (see p. 63)

[436] P C Ng and S Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, **11**: 863–74, 2001. (see p. 63)

[437] S Sunyaev et al. Prediction of deleterious human alleles. *Hum Mol Genet*, **10**: 591–7, 2001. (see p. 63)

[438] Lei Bao and Yan Cui. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**: 2185–90, 2005. (see p. 63)

[439] Nathan O Stitziel et al. Structural location of disease-associated single-nucleotide polymorphisms. *Journal of Molecular Biology*, **327**: 1021–30, 2003. (see p. 63)

[440] Carles Ferrer-Costa, Modesto Orozco, and Xavier De La Cruz. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of Molecular Biology*, **315**: 771–86, 2002. (see p. 63)

[441] Christopher T Saunders and David Baker. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology*, **322**: 891–901, 2002. (see p. 63)

[442] D Chasman and R M Adams. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, **307**: 683–706, 2001. (see p. 63)

[443] S Sunyaev, V Ramensky, and P Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, **16**: 198–200, 2000. (see p. 63)

[444] Z Wang and J Moult. SNPs, protein structure, and disease. *Hum Mutat*, **17**: 263–70, 2001. (see p. 63)

[445] Abel González-Pérez and Nuria López-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, **88**: 440–9, 2011. (see p. 63)

[446] Yana Bromberg and Burkhard Rost. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, **35**: 3823–35, 2007. (see p. 63)

[447] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, **30**: 3894–900, 2002. (see p. 63)

[448] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, **7**: 61–80, 2006. (see p. 64)

[449] Man Hoang Viet et al. Effect of the Tottori Familial Disease Mutation (D7N) on the Monomers and Dimers of A40 and A42. *ACS Chem Neurosci*, 2013. (see p. 64)

[450] Yu-Shan Lin and Vijay S Pande. Effects of familial mutations on the monomer structure of A. *Biophys J*, **103**: L47–9, 2012. (see p. 64)

[451] S Cuesta-López, F Falo, and J Sancho. Computational diagnosis of protein conformational diseases: short molecular dynamics simulations reveal a fast unfolding of r-LDL mutants that cause familial hypercholesterolemia. *Proteins*, **66**: 87–95, 2007. (see p. 64)

Objectives

As exposed in the previous chapter, computational approaches are very useful and can be considered as a synergistic and necessary complement to experimental methods. Thus the need to generate new and more accurate computational algorithms. Following this idea, and taking into consideration all the previous experimental and computational background for the study of protein conformational instability, protein aggregation and misfolding, we have defined the following objectives for this Thesis:

General Objective

> The study of protein conformational flexibility and misfolding from a bioinformatics point of view at the sequence, structural and dynamical levels in three different case studies

Specific Objectives

1. Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains
2. Prediction of local unstable regions of proteins based on physicochemical and geometric characteristics of buried protein interfaces
3. Predicting abnormal phenotypes caused by **S**ingle **N**ucleotide **P**olymorphisms (SNPs) in a Conformational Disease : Familial Hypercholesterolemia

CHAPTER 2

Discovering Prion-like Proteins in Complete Proteomes Based on Probabilistic Representations of Q/N-rich Domains

Contents

THE RESULTS PRESENTED IN THIS CHAPTER ARE A COMPENDIUM OF THE FOLLOWING RESEARCH ARTICLES:

- Angarica, V.E., Angulo, A., Giner, A., Losilla, G., Ventura, S., and Sancho, J. (2014). **PrionScan: an online database of predicted prion domains in complete proteomes**. *BMC Genomics*. **15:** 102
- Angarica, V.E., Ventura, S. and Sancho, J. (2013). **Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains**. *BMC Genomics*. **14:** 316

2.1 Introduction

The formation of intracellular amyloid fibrils is a widespread phenomenon in eukaryotes[1–4] and it has been found related to a number of beneficial adaptive cellular functions[5–11], to protein-encoded heritable information transmission in yeast[12–15], and to a variety of important diseases in mammals[16–20]. Amyloidogenesis is mediated by a diverse group of evolutionarily unrelated proteins from different organisms, all sharing the propensity to form β-sheet aggregates in their complete or fragmented forms[16]. A subset of these aggregation-prone proteins is characterized by the presence of regions that comprise homopolymeric tracts, also named 'single sequence repeats'[21]. It has been reported that the presence of these low complexity stretches, and more specifically that of Q/N-rich regions, strongly influences the aggregation potential of eukaryotic proteins[22–24]. In several neurodegenerative disorders, such as spinocerebellar ataxias and Huntington's disease, long pure glutamine repeats are generated by the instability of CAG codons[25–27], and cause the abnormal proteins to form intracellular inclusions in specific neuron types. However, prionogenic Q/N-rich regions usually contain additional amino acids and form sequentially heterogeneous domains responsible for the main properties of prions, including self-propagating amyloid aggregation.

Much research has been devoted to determine the structural and sequential basis of prion formation, and the compositional determinants of prionogenic domains. Studies from different groups have concluded that both amino acid composition and the length of such regions play important roles in prion induction[28–30]. Additional sequential requirements such as the number and distribution of prolines and charged residues have been recently found to be relevant in the formation of prionic aggregates[28]. Mutational studies, in which the sequence of yeast prions ***Ure2p*** and ***Sup35p*** were randomly shuffled, proved that the [*PSI*+] phenotype is mainly determined by the amino acid composition of the domain independently of the primary sequence, as most of the shuffled species generated were able to form prions *in vivo*[29,30]. This knowledge has been used to try to predict putative prions in biological sequence databases, though the available methodologies to carry out the task are just a few. A first group of algorithms intend to estimate the propensity of peptides of a given length to form amyloid aggregates based on their primary sequence[31–34]. This kind of methods, based on more or less complex models of parallel β-sheets, have proven quite ineffective for coping with Q/N-rich stretches since these domains do not share the common characteristics of β-sheet-amyloid forming peptides[35] –*e.g.* high hydrophobicity.

A second group of methodologies try to predict Q/N-rich domains from the primary sequence based on the strong amino acid compositional bias of these segments. Proteome-wide identification of Q/N-rich regions was successfully achieved in 30 proteomes from eukarya, archaea and eubacteria using a quite straightforward algorithm, based on the estimation of the significance of occurrence of regions with a high proportion of Q and N[36]. A similar methodology for assessing compositional bias in biological sequences was also tested to find proteins enriched in Q and N[37]. However, these two algorithms only take into consideration the frequency of a specific group of biased amino acids in a given sequence segment –*i.e.* Q/N, hydrophobic or charged amino acids– instead of considering the relative contribution of all the residues present in the segment to the prionogenicity of the domain[29]. Furthermore, they failed to generate a statistical model and a scoring function that would allow the systematic evaluation of protein segments and sorting the predicted domains according to their prionogenicity. A recent report has proposed an interesting alternative procedure to generate a bioinformatics model to predict prions at genomic scale. Starting from the sequences of four known yeast prions, a hidden Markov model (HMM) was generated to assess the compositional similarity of proteins from the yeast proteome to the model. This yielded up to 200 proteins with candidate prionogenic domains (PrD), from which the top scoring 100 were tested experimentally *in vitro* and *in vivo*[38]. Finally, a total of 19 new proteins that proved switching behavior and amyloid formation were identified, in addition to the four prions previously described in this organism. Notwithstanding the remarkable outcomes from this work, the inherent bias of the predictive model built, generated from just a few sequences[38], apparently hampers its ability to correctly score proteins sequences, as roughly half of the high scoring predictions were false positives exhibiting no prion-like behavior.

A complementary strategy went farther in an attempt to define the compositional features that influence prion formation. Libraries of ***Sup35p*** mutants expressed *in vivo* were used to comprehensively analyze the sequence compositional determinants of prions[28]. This study ultimately produced an experimental technique to measure the prion propensities of individual amino acids, showing that there is a strong bias against prolines and charged residues, a strong bias favoring the presence of hydrophobic residues and no significant bias for or against Q/N residues[28]. With this methodology, the scoring of the putative prions made by Alberti *et al.* could be improved. A recent follow up by the same group has used this methodology to design *de novo* synthetic prionogenic sequences capable, not only of forming amyloids, but also to stably propagate over many generations[39]. However, this and the other approaches available to date for identifying and predicting Q/N-rich segments with prionogenic activity, lack a detailed statistical benchmarking of their performances at a genomic scale. Thus, a methodology able not only to identify putative prion domains in large databases of protein sequences, but also to correctly classify the predictions in terms of precision and accuracy would be of high interest.

Here we present a bioinformatics approach to create a statistical representation of prion domains that allows scoring protein sequences according to their likelihood of being prions. Starting from a list of 29 proteins reported experimentally to exhibit conformational conversion and amyloid formation in yeast[38], we have developed a probabilistic model of PrDs to discover Q/N-rich prionogenic proteins in complete proteomes. The independent probability of occurrence of all amino acids in prion domains were estimated and a log-likelihood model was built to assign uncalibrated scores to sequence fragments of variable length. We first benchmarked our model against a list of 18 proteins that were tested in the same experimental conditions and showed no *SUP35C* activity *in vivo*[38]. From this assay we obtained the predictive cutoff that should be used and the confidence intervals of the predictions. Our classifier performed fairly well filtering prions from proteins with no prionogenicity with an accuracy higher than 0.83 and a precision of 80% at the predictive cutoff set. In these conditions the fraction of false positives was rather low, corresponding to less than 16% of the total predictions. We also tested the ability of our model to scan large sequence datasets from Uniprot[40], the PDB[41] and intrinsically disordered proteins (IDPs) annotated in Disprot[42]. Our results proved that the model is well suited to handle datasets with a high proportion of negative instances without recovering an excessive amount of false positives, which is important to perform predictive assays in complete proteomes. Our scoring model was effective to almost completely separate the distributions of real prion domains from the Uniprot and PDB datasets, while the sequence of some IDPs proved more alike Q/N-rich prion forming domains.

We have used this methodology to scan all the known proteomes annotated in public databases, which yielded 27925 predictions in 3236 different organisms from all taxa[*]. In order to provide a functional and public framework for analyzing the large amount

[*] These figures correspond to the predictions generated analyzing Uniprot release of June 2013

of data generated in our study, we have developed PrionScan[†], as an open source of up-to-date prion predictions for all the proteins annotated in public databases. The site is designed as a simple and flexible querying system suitable for data mining by combining different sorts of information included in our database to recover, for instance, prion predictions in the complete genome of an organism or for proteins belonging to a specific functional family or related to a specific biological process. This is to our knowledge the most extensive effort to predict PrD sequences performed so far, reporting putative prions in the proteomes of a diverse group of organisms, most of which have been poorly studied. We mined the information stored in PrionScan to make global analyses of the distribution of prion proteins in different functional categories, and localization in different cellular compartments. We have observed some interesting trends in the distribution of PrDs in different protein functional families. From our results it appears that prions are associated with different cellular components and to function in different biological processes depending on the taxon and organisms group. The present predictive approach uncovers a large set of putative prionogenic proteins whose further experimental characterization might contribute significantly to understanding prion biology from a genome-wide perspective.

2.2 Results

2.2.1 Amino Acid Composition of Prion-forming Domains

Based on the sequence of a group of experimentally tested protein domains that showed prion-like behavior *in vivo* and *in vitro* in yeast[38], we trained an unsupervised classifier relying on the amino acid propensities in PrD domains, see the Methodology section for more details. The estimated relative abundance of each amino acid type in a group of well-characterized prion domains with respect to the expected frequency of occurrence in proteins is shown in Table 2.1. Some residues, such as G, H, M and P, are equally frequent in PrD and proteins. Other residues, including C, E, D, K and W, appear to be underrepresented in prion forming domains, while Q and N and also Y and S, have a significant positive bias. Unlike previous approaches[36,37], this model allows us to obtain a representation of prionogenic domains accounting for the relative statistical significance of each residue in the scoring function. The high odds ratios observed for Q (4.1) and N (5.7), which represent the previously reported favorable bias for these residues in PrDs, can be combined with the statistical potentials obtained for amino acids such as C and W, which are 14 and 10 times less frequent in these regions than in proteins.

[†] Available at: http://webapps.bifi.es/prionscan

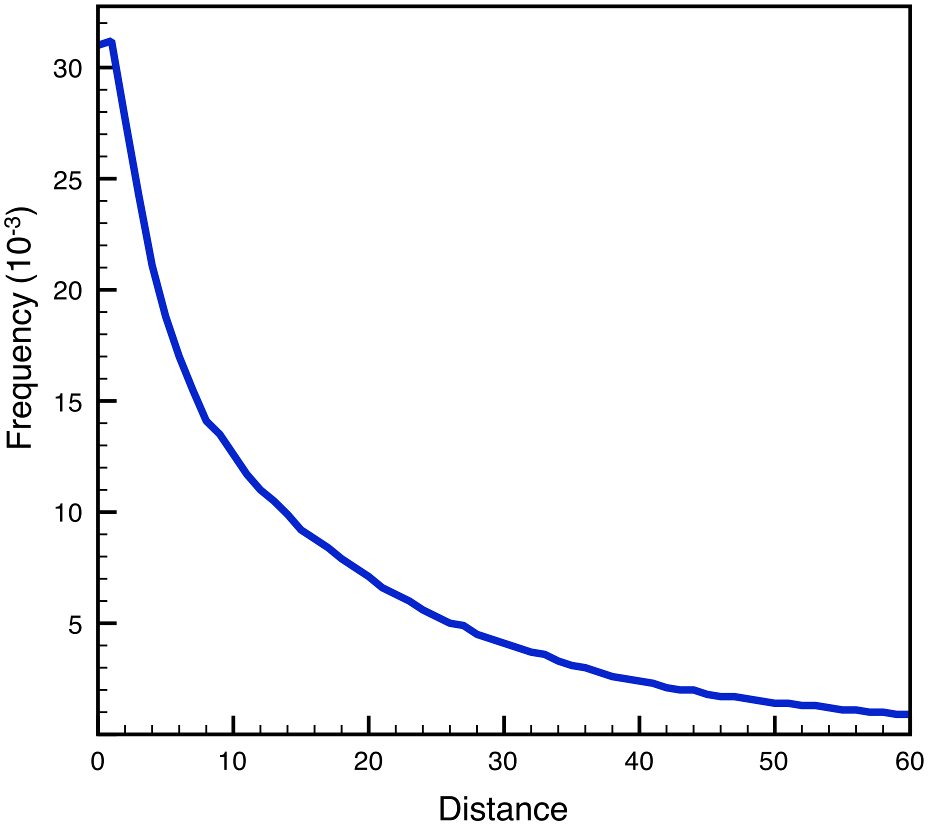
TABLE 2.1: Amino Acid Propensities in Prion Domains

Residue	Prion domains		Prion domains (Library 1)	
	Odds ratio	LOr	Odds ratio	LOr
A	0.675	−0.568	0.670	−0.578
C	0.071	−3.807	1.520	0.604
D	0.352	−1.507	0.280	−1.837
E	0.147	−2.766	0.550	−0.862
F	0.718	−0.478	2.310	1.208
G	1.028	0.040	0.960	−0.059
H	0.913	−0.131	0.760	−0.396
I	0.350	−1.515	2.260	1.176
K	0.271	−1.883	0.210	−2.252
L	0.340	−1.556	0.960	−0.059
M	1.125	0.170	1.960	0.971
N	5.700	2.511	1.080	0.111
P	1.170	0.227	0.300	−1.737
Q	4.125	2.044	1.070	0.098
R	0.436	−1.196	0.670	−0.578
S	1.662	0.733	1.140	0.189
T	0.830	−0.268	0.890	−0.168
V	0.304	−1.716	2.260	1.176
W	0.091	−3.459	1.950	0.963
Y	1.724	0.786	2.180	1.124

The observed frequencies of occurrence of the different amino acid residues were transformed into the corresponding statistical potentials using the Equation 2.1. Columns 2 and 3 show the calculated odds-ratio for the complete prion and the statistical potentials corresponding to the odds-ratios of PrD respectively (LOr). Columns 4 and 5 contain the ratio and log-odds obtained experimentally by means of a random mutagenesis assay, as described in Toombs *et al.*[28]

The analysis of the ratios reported in a previous work[28], resulting from a random mutagenesis assay of two specific segments of ***Sup35p*** protein, reveals significant differences with our results. They include, as can be inferred from the comparison included in Table 2.1, differences in the relative log-odds for some important residues such as E, 3.8 times less frequent in PrDs according to our results and P, which is 3.9 times more likely to be found in these domains according to our model. The more remarkable differences are obtained for some key residues such as Q and N, for which we found a marked favorable bias. For other residues such as K, Y, S and D no significant differences were found between our model and the results from Toombs *et al*[28]. The contribution of P to the prionogenicity of a given sequence stretch, unlike those of other amino acids, appears to be related not just to its abundance in PrDs. As it has been previously noted, prolines in prions tend to appear in clusters while, in non-prionogenic Q/N-rich proteins, they are usually scattered along the complete sequence of the stretch[28]. However, there were no experimental or theoretical models to relate the existence of specific proline patterns

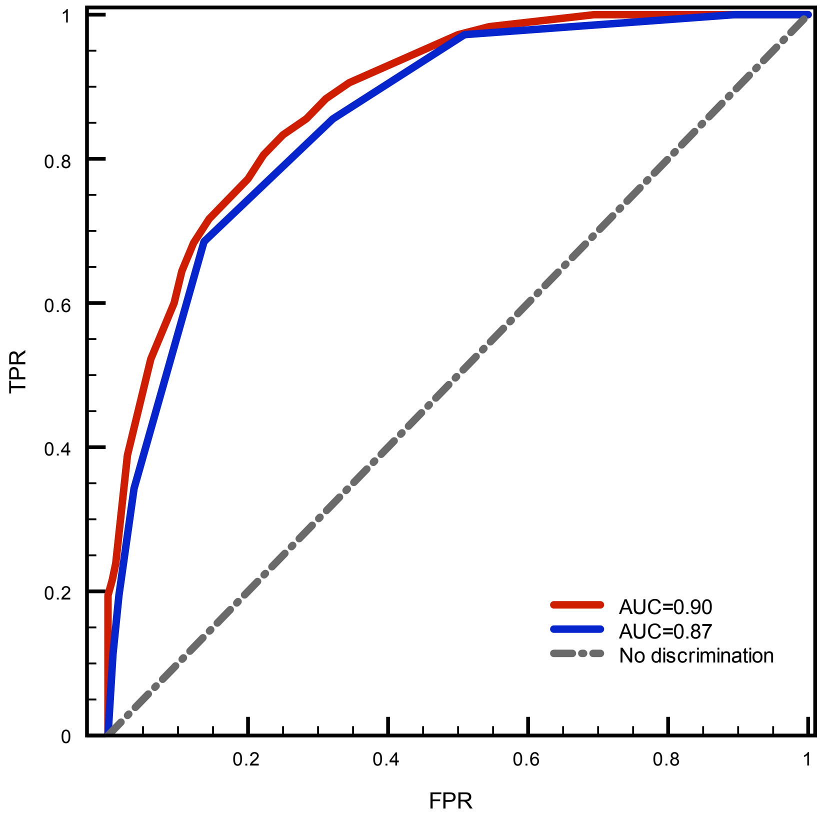
FIGURE 2.1: Observed Frequency of $P-(X)_n-P$ Patterns in Proteins



A representative non-redundant dataset of 4606913 proteins from Uniref 50 were analyzed in the search for the significance of proline patterns in the protein universe. In the chart we plot the trend of the observed frequency of each pattern of two prolines separated a given distance between 1 and 60 residues

in a given PrD with the prionogenicity of the sequence. In our model we use an approach to correct the score calculated for a given stretch from the relative propensities of the amino acids (see Equations (2.1) and (2.2)), taking into account the number of non-contiguous prolines found in the segment, as described in Equation (2.3). In this approach, we first estimated the relative frequency of pairs of prolines separated a given distance in a non-redundant dataset of protein sequences and convert those frequencies into log-likelihoods, see Figure 2.1. We then use those log-likelihoods to assess the significance of finding a pattern of prolines, separated a given distance in the window of sixty residues used for the scanning, and the resulting support value is used to correct the compositional score, see Equation (2.3). In this way, using solely sequence information, we generate for a given sequence, a corrected score which takes into account both the relative propensities of the amino acids and the unfavorable contribution of non-contiguous prolines to prion formation.

FIGURE 2.2: ROC Plots of the PrD Recovery and Bootstrapping Assays



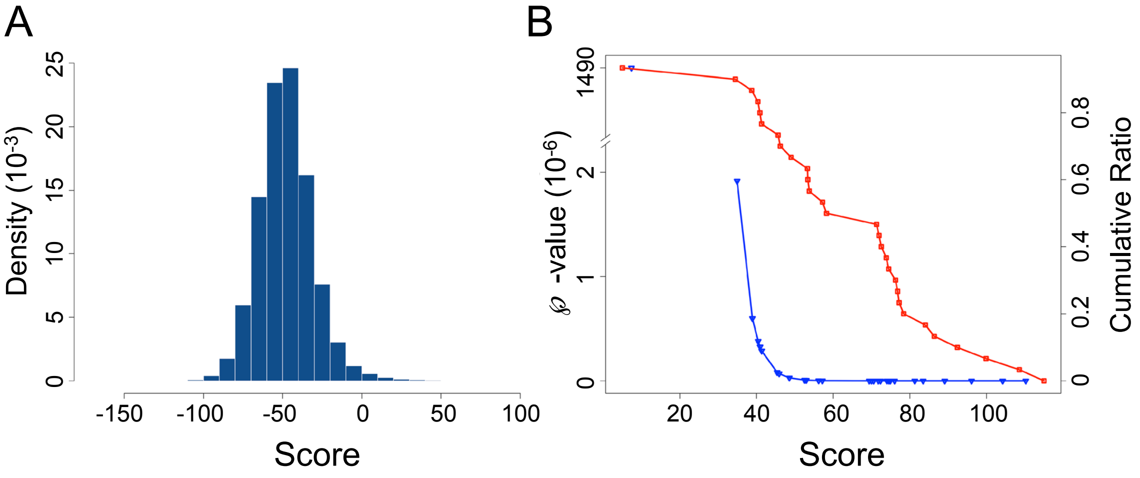
The scoring histogram distributions of the negative and positive datasets were processed and the true positive rate (TPR, defined in Equation (2.4)) was plotted against the false positive rate (FPR, defined in Equation (2.5)) in a tryout in which the known PrDs –*i.e.* positives in all four experimental tests– are picked up from a test dataset of non prions –*i.e.* negatives in all four experimental tests[38]. In red we show the plot obtained using our model which has an area under the curve $AUC = 0.90$. We also include the result of a bootstrap assay in which the 18 prions used as the training set were resampled 1×10^6 times forming partial training sets of 9 prions and generating positive test sets for the ROC plot analysis with the remaining 9 prions. One million ROC plots were generated always using the same negative set and the average ROC curve was calculated (shown in blue), the area under the curve $AUC = 0.87$

2.2.2 Using Compositional Bias to Assess the Prionogenicity of Protein Sequences

We used the model obtained for the PrD domains to scan protein sequences. In order to ease the analysis at the benchmarking stage, we selected the highest scoring stretch in a given sequence as the putative PrD, assuming only one prionogenic region per protein. Though there are evidences of proteins that bear more than one prion-forming domain and in some cases the PrD is a diffuse region of more than 60 residues[38], this approximation significantly reduces the number of sequence fragments to be analyzed, without affecting the number of true positive predictions. A detailed assessment of the predictive potential of our model is shown in Figure 2.2. The ROC plot obtained from the analysis of known PrDs and the negative dataset used in benchmarking illustrates the good performance of the algorithm, with an area under the ROC curve $AUC = 0.90$. The AUC is a global estimator of the statistical significance of a classification test, representing the probability that, each time a pair of positive and negative instances is randomly retrieved from the pool, the scoring function will assign a higher score to the positive example. The non-parametric Mann-Withney-Wilcoxon rank-sum test for distributions comparison[43], is rather low ($\wp - value = 6.7 \times 10^{-6}$) with a significance $\wp - value < 0.05$. We did not have access to the absolute scores in the HMM-based prediction of the yeast prions[38], which were subsequently used to implement our method. This previous work described in detail an extensive experimental assessment of the predictions, but few details were available on the scoring and benchmarking procedures, thus impeding a quantitative evaluation of the performances of the two methods. We addressed this comparison indirectly investigating how our predictor scored the *bona fide* prions identified in the above mentioned work with respect to the complete yeast proteome. The analysis is described in Figure 2.3, where we include the density distribution of the scoring of all the proteins annotated in the genome of *Saccharomyces cerevisiae* and the corresponding $\wp - values$ of each of the 29 known prions in this organism. This chart indicates that our methodology is able to discriminate PrDs from the rest of the proteins in the proteome. Except for ***RBS1*** PrD, whose $\wp - value$ of 1.49×10^{-3} locates it in a more or less confusion zone in the scoring distribution, the $\wp - values$ for the rest of real PrDs are well below 1×10^{-6}. This means that PrDs can be retrieved as a completely different distribution from the proteome score distribution, with a significance level of 0.1%. In addition, at a score of 50 $bits$, 63% of the real PrD have a $\wp - value$ lower than 3.4×10^{-8} (Figure 2.3, panel B).

We also decided to test the wealth of the amino acid propensities calculated in our model and check whether there is a high rate of redundancy within the training set, which could hamper the predictive potential of the model. Thus, we performed a thorough bootstraping assay in which we randomly resampled 1×10^6 training sets from the 18 sequences that are positives in all the experimental assays, leaving out 9 PrDs each time, see the Methodology section for details. In each case we recalculated the propensities and used the excluded PrDs as positive test set in the ROC plot tryouts, maintaining the same negative set. The results of this experiment are also shown in Figure 2.2, where the average ROC curve calculated from the million plots generated is depicted. As expected, the AUC decreases, but only to 0.87, which still corresponds to a fairly good classifier performance, reflecting that the deviation from the most common classification behavior is marginal. This finding means that the estimated propensities

FIGURE 2.3: Scoring of PrDs in Yeast with Respect to the Complete Proteome



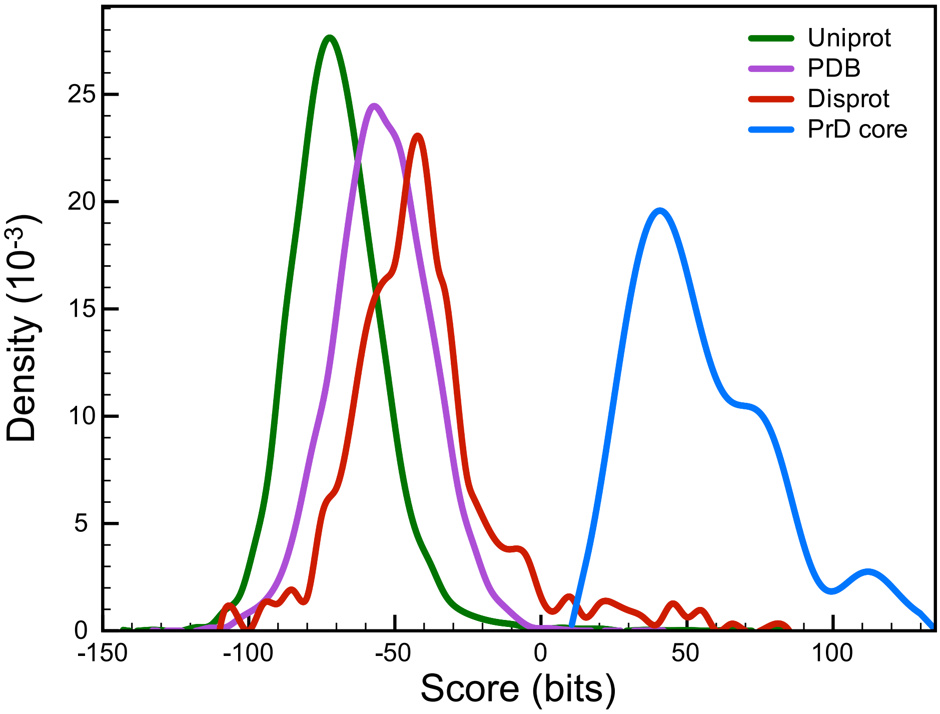
The density histogram of the score of all the proteins in the yeast genome is shown in panel **A.** In panel **B**, in the left ordinate axis we include the observed $\wp - value$ for the 29 known prions in this organism (blue line connecting open triangles) and the cumulative ratio representing the percent of known prions with a $\wp - value$ equal or less than a given value is shown in the right ordinate axis (red line connecting open squares)

calculated from the training set are unbiased and are significant enough to correctly separate the population of positive and negative instances.

2.2.3 Testing the Suitability of our Algorithm to Process Large Sequence Databases

The ROC plot analysis is an excellent technique to evaluate the predictive potential of a classification methodology, since it is insensitive to changes in the class distributions –*i.e.* the TPR *vs* FPR dependence remains the same if the proportion of positive to negative instances changes. Nevertheless, this property becomes a limitation when the number of negative instances is considerably higher than the population of positives, which is quite common in the analysis of large biological sequence databases. In this scenario, a classifier corresponding to a reasonably well-shaped ROC plot with a high AUC might return an elevated number of false positives along with the putative predictions at a specific cutoff score. Therefore, it is very important to complement ROC trials with other performance metrics that combine different classes of the *confusion matrix* and are consequently sensitive to class skew. In Figures 2.4 and 2.5 we inspected the dependence of the $Precision$ of our classifier and the recovery rate of known PrD for the three test datasets. Firstly, Figure 2.4 confirms that our classifier is able to differentiate the prions from the sequences included in each of the test sets, without significant tail

FIGURE 2.4: Histogram Density Plots of the Scoring of Protein Sequences Using our Probabilistic Model



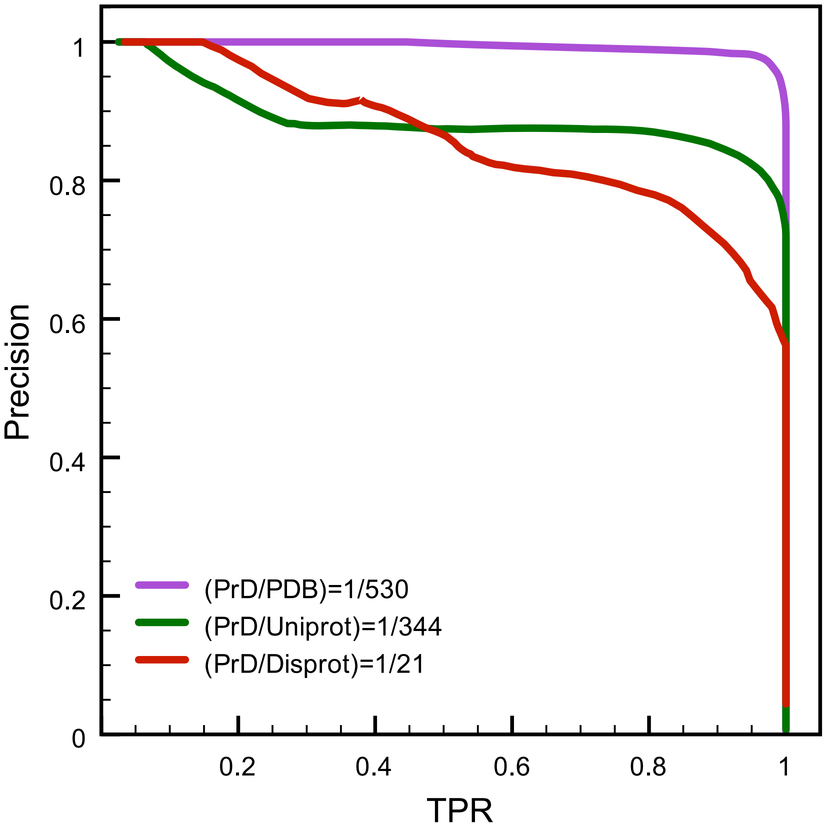
Protein sequences from Uniprot/Swissprot, PDB, Disprot and experimentally tested yeast prions (PrD) were scanned using a window size of 60 amino acids. The corresponding histograms were transformed into the normalized density distributions (ordinate), plotted against the score in the abscissa

superposition at lower scores. However, even if the superposition of the scoring distribution around zero $bits$ might seem insignificant, the number of false positives could be significant enough just because the sizes of the distributions of negative sequences are various orders of magnitude higher than that of prions. However, as depicted in Figure 2.5, our results confirm that our algorithm also performed very well for processing large sequence datasets. It is clear in this Precision-Recall chart that despite the proportions of the distribution of prion-forming domains and the corresponding distributions of the three test sets –*e.g.* Disprot is 21 times larger than PrD dataset while the PDB dataset is 530 times larger– we were able to pick up almost 90% of the true positives yielding $Precision$ values above 80%.

2.2.4 Selection of a Cutoff Value for Predicting in Complete Proteomes

The classification accuracy of the method can be taken into account to select the predictive cutoff, see Figure 2.6. The evaluation of the rate of correctly mapped instances

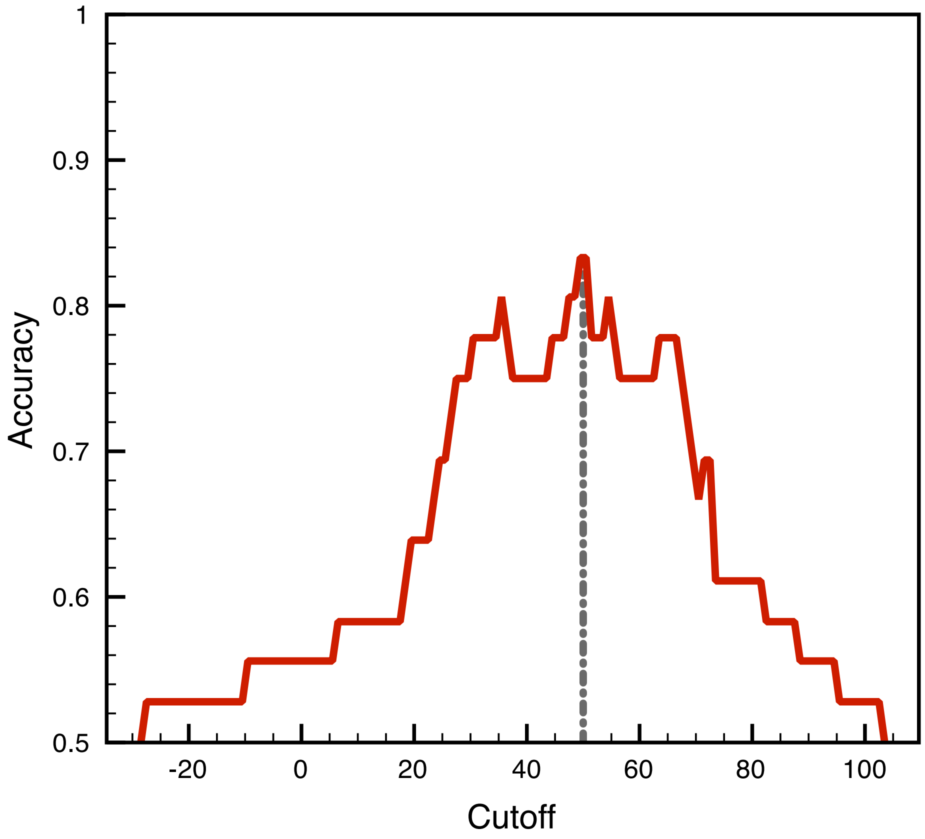
FIGURE 2.5: Precision-Recall Plots for the Comparison of PrD and Non-prionogenic Sequence Distributions



For each one of the three negative additional datasets including proteins from Uniprot, Disprot and the PDB we follow the evolution of the classifier's $Precision$, defined in Equation (2.6), to correctly make a positive mapping of known PrD segments from a pool of non-prionogenic sequences. These values are plotted against the TPR – *i.e.* recall– of the corresponding classification step. The ratio between the number of instances in each positive and negative distribution is also shown

from both positive and negative distributions prove that our method is able to both correctly scoring and separating sequences that experimentally showed prion-like activity from other sequences with no such an activity in the same assays, but also handling at the same time disproportionate positive and negative datasets, see Figure 2.5. As can be inferred from Figure 2.6, in our model the cutoff value of 50 $bits$ marks the maximum predictive accuracy. This was the cutoff score set for performing prediction assays in complete proteomes as will be described below. With this cutoff we guarantee both an $Accuracy$ of 83% and a $Precision$ of classification as high as 80%. These values of classification efficiency are comparable with those obtained with a methodology reported recently used for *de novo* design of synthetic prion domains[39]. We also obtained estimations of the proportion of false positives that our algorithm will necessarily recover along with the putative predictions. The false discovery rate (FDR, defined in Equation (2.8))

FIGURE 2.6: Accuracy-Cutoff Plot of the Classifier Against the Negative Test Set



The $Accuracy$, defined in Equation (2.7), obtained for the correct classification of TP and TN is graphed against decreasing cutoffs spanning the score range of the corresponding negative and positive distributions. We highlighted the highest accuracy of the assay, used to set the predictive cutoff of 50 $bits$

is quite an interesting metric in classification problems, corresponding to the proportion of events in which the null hypothesis is incorrectly rejected, or in other words, the likelihood of incurring in *type I error* in a test[44,45]. In our benchmarking tryouts, the FDR obtained for the selected cutoff of 50 $bits$ is 16%. This value indicates that our methodology produces fairly clean recovery sets with a rather low proportion of false positives.

2.2.5 Proteome-wide Predictions of Proteins Bearing Putative PrDs

After a comprehensive benchmarking of our model we used it to predict proteins containing PrD in the complete proteomes of organisms. As described in the Methodology section, we performed a scanning of all the proteins annotated in complete proteomes, and the predictions obtained in this search are available in the PrionScan database. Our methodology yielded 27925 predictions of putative prions in 3236 different organisms

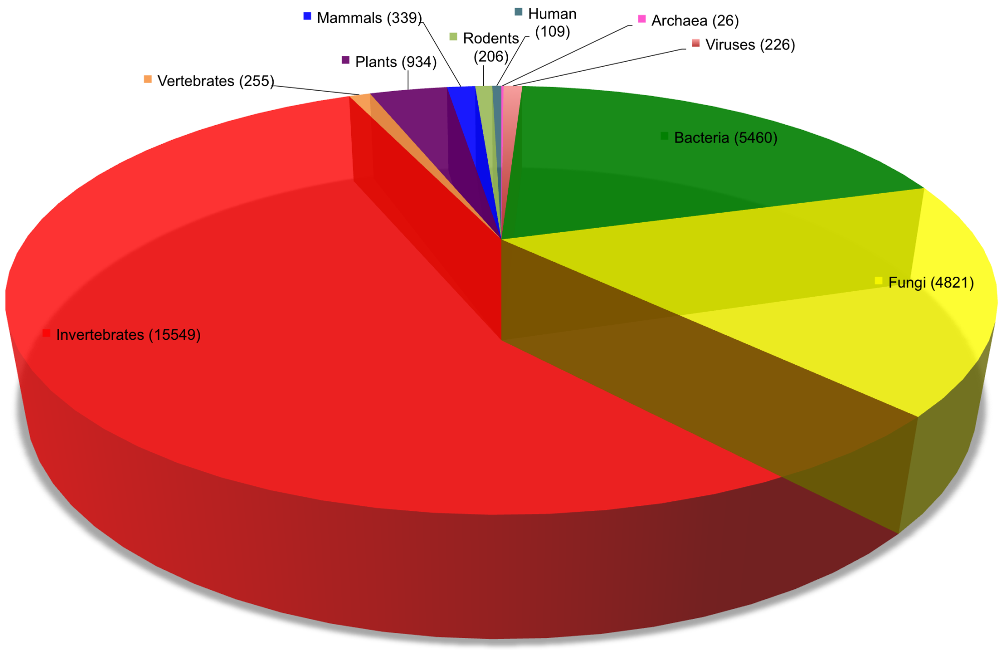
TABLE 2.2: Summary of Prion Predictions in Different Taxa

Taxon	**# Organisms**	**# Proteins**	**# Predictions**
Archaea	18	708135	26
Bacteria	2531	31097600	5460
Viruses	69	1773811	226
Fungi	196	1976771	4821
Invertebrates	228	2293203	15549
Vertebrates	27	857235	255
Plants	96	2027308	934
Rodents	13	222750	206
Mammals	57	797045	339
Human	1	133798	109

The predictions obtained for all the organisms analyzed is organized by taxon and the following information is included in the table: in the first column the taxon; in column 2, the number of organisms for which we obtained predictions; in column 3, the number of proteins scanned in the search for PrDs; and column 4 shows the number of predicted proteins bearing prion-forming domains

from all taxa, from viruses and archaea to plants and higher eukaryotes. A summary of the predictions obtained in all taxa is shown in Figure 2.7 and in Table 2.2 we include a detailed description of the predictions obtained in each taxon.

FIGURE 2.7: Distribution of the Predictions in PrionScan in all Taxa



The pie chart depicts the distribution of prion predictions in all different taxa, from archaea to humans. In each case, besides the color code, we also include the number of predictions in each taxon in parenthesis

The inspection of some selected organisms shown in Table 2.3 illustrates some interesting trends of prion content in proteomes. In most cases the percent of proteins bearing prion-forming domains is less than 1% of the size of the proteome. In Archaea and Viruses the number of putative prion proteins is less than 10 per proteome (with the sole exceptions of *Acanthamoeba polyphaga mimivirus* and *Porcine epidemic diarrhea virus*), while in Bacteria, Fungi, Plants and animals it might range from a few tens to a few hundreds in some specific organisms. Among Bacteria there exist important exceptions such as *Staphylococcus aureus*, for which the number of prionogenic proteins correspond to almost 20% of the genome. In Protozoa we observe important differences in the ratio of PrDs in the proteome of different organisms of this class. While for *Cryptosporidium parvum*, *Theileria parva*, *Trypanosoma brucei* the percent of PrD proteins in the genome is relatively low, for *Dictyostelium discoideum*, *Dictyostelium purpureum* and *Plasmodium falciparum* the proportions of putative prions are as high as 20%, 8% and 10% respectively. This is in agreement with previous reports proving the abundance of hydrophilic low-complexity regions in the proteome of these organisms[46,47]. This tendency is also present in other species from the genus Plasmodium, such as *Plasmodium yoelii*, which has 137 PrD proteins in its proteome. Another noticeable examples correspond to Fungi, which have a relatively high number of prions in their genomes. Previous reports have found this trend in the genomes of yeasts in which these repetitive stretches are generated by DNA tandem duplication[48], rendering protein domains that were thought to have no function[49], but that according to our results might indeed be prion proteins involved in homeostatic processes. In Dipterans, there are also a significant number of predictions, amounting to 1–4% of the genome for *Anopheles gambiae*, *Drosophila mojavensis* and *melanogaster*.

2.2.6 Building an online Database of Predicted Prion Domains in Complete Proteomes

There are a few examples of repositories with information on prion proteins, prionogenic sequences, prion-related diseases, prion protein interactions and orthologs and paralogs of prion proteins in multiple organisms. For example, the Prion Disease Database[50] contains a sort of experimental data on prion sequences and multi-level data on diseases caused by prions, combined with a set of tools for data analysis and systems biology studies in mouse. PrionHome[51] is a non-redundant database containing approximately 2000 prion-related sequences obtained from different public and private sources, in some cases with experimental support or inferred using different predictive algorithms[38,52,53]. There is yet another similar resource, set up as a web application for predicting prion forming propensity[39]. Though not a database in the strict sense of

TABLE 2.3: Ratio of Prion Domains in the Proteomes of some Model Organisms

Species	Predictions	% of the proteome
Listeria monocytogenes^1	146	4.86
Bacillus cereus^1	218	4.01
Staphylococcus aureus^1	515	19.70
Cryptosporidium parvum^2	60	1.57
Dictyostelium discoideum^2	2692	20.10
Dictyostelium purpureum^2	992	8.01
Plasmodium falciparum^2	853	10.20
Theileria parva^2	11	0.50
Trypanosoma brucei^2	15	0.16
Candida albicans^3	169	2.62
Saccharomyces cerevisiae^3	746	12.63
Lodderomyces elongisporus^3	150	2.58
Arabidopsis thaliana^4	56	0.20
Oryza sativa^4	50	0.08
Drosophila melanogaster^5	765	3.72
Drosophila mojavensis^5	486	3.33
Anopheles gambiae^5	160	1.16
Caenorhabditis elegans^6	98	0.42
Homo sapiens^7	111	0.29

The percent of the proteome corresponding to proteins bearing putative prion-domain (column 3) is shown for a representative group of model organisms (column 1), from different evolutionary classifications, some of which have been extensively studied and whose complete genomes have been well characterized. The organisms included correspond to different species of $^{(1)}$ bacteria, $^{(2)}$ protozoans, $^{(3)}$ yeasts, $^{(4)}$ plants, $^{(5)}$ dipterans, $^{(6)}$ nematodes and $^{(7)}$ human. The number of predictions obtained for each organism is shown in column 2. *For some cases, the number of predictions annotated in PrionScan can be higher because we also included the predictions for proteins of subspecies

the term, the PAPA site[‡] allows the analysis of protein sequences based on amino acid propensities in prion sequences inferred from *in vivo* aggregation analysis. In contrast to these available resources, PrionScan provides genomic-scale prion predictions for the proteomes of all organisms –*i.e* almost 28000 putative prion proteins– in a framework that allows an easy way to study the sequential/structural determinants of prionogenicity, as well as transversal comparative studies of the implication of prions in cell biology in different groups of organisms. The site is organized in an easy way for mining our data and also offering a web functionality for the high-throughput analysis of sequence variants not annotated in public databases, see the Methodology section for details. A comparison of the predictions obtained from different updates of Uniprot gives an idea of the increasing pace of prion prediction in proteins annotated in this database. In fact, since the predictions obtained by us in the initial paper describing our method, in which we used the Uniprot[54] (update 2012_02) from February 2012, to the predictions

[‡] Available at: `http://combi.cs.colostate.edu/supplements/papa/`

annotated now in PrionScan, which correspond to the Uniprot (update 2013_07) of June 2013, there has been an increment of more than a 61% from the approximately 17400 predictions obtained at that time.

2.3 Discussion

2.3.1 From Amino Acid Composition to a Comprehensive Model of Prion-forming Domains

Great effort has been devoted in recent years to the experimental characterization of prion proteins, with a special interest in defining the sequential and structural determinants of aggregate formation and prion transmission. To date, the number of prions studied is still limited and little is known regarding the approximate number of prion-like proteins in complete proteomes or the cellular processes in which they might be involved. Nevertheless, several studies have shed some light into the general characteristics of prions[1,17,55–57] and how this information can be used to try to identify novel Q/N-rich candidates in protein databases[28,36–38]. Only recently the availability of high-throughput experimental procedures to study prions *in vitro* and *in vivo*[38,58–60], and the feasibility of extensive mutational studies[28–30,61], have provided deeper insights into the characteristics of protein domains that mediate aggregation and prion induction. It is now clear that methodologies relying on approximating the likelihood of contiguous protein stretches to form parallel β-sheets[31–34] cannot be successfully used to predict Q/N-rich prion domains. Among other examples, these methods are unable to predict β-aggregation nuclei in known yeast prions such as ***Ure2p*** and ***Sup35p***[62]. Instead, prediction of PrDs using the distinctive amino acid composition of these domains[28,36,37] and assuming primary sequence independence for prion formation[29,30,39,61], appears more promising. A recent comparison of most of the methods currently used to predict prion propensity has proved that approaches that focus largely on composition –*e.g.* PAPA and Zyggregator– show far more predictive accuracy than those focusing on primary sequence[39].

Following this idea, we have generated here a reliable model that uses the compositional bias of PrDs, taking special care on thoroughly benchmarking the algorithm in order to establish realistic confidence intervals for predicting in large biological sequence databases. The results from the work by Alberti *et al.*[38] were very valuable to provide an ample enough training set from which we obtained the statistical potentials summarized in Table 2.1. The odds-ratios calculated by us embody the previously described bias observed in prion-forming domains[28,36,37], and enable the inspection of protein sequences to find putative PrDs. Our method relies solely on amino acid propensities calculated using compositional bias, plus a correction to the score which accounts for the unfavorable existence of certain proline patters in the sequences analyzed, see Figure 2.1. The variance of the score distribution of candidate prions for which there is strong experimental evidence[38], reflects the high sequential variability that aggregation-prone domains can accommodate. In their work, Alberti and coworkers do not make a statistical evaluation of the predictive power of the model used. Instead, they rely on the potentiality of the high-scale experimental assays performed to classify the predictions. They acknowledge the bias of the hidden Markov model built[38], which might be related to the scant scoring capability of the method that ranks highest a number of sequences that showed no aggregation propensity. The training stage is very important in the construction of HMMs[63], and this is probably why this model, generated from just a few examples, is able to identify probable candidates but is unable to score them correctly. We believe our model improves the scoring of these sequences, as can be inferred from the scoring of known PrDs in the complete yeast genome, see Figure 2.3.

Another recent study aimed at modeling and predicting prions[28] has produced interesting results. The authors carried out random mutagenesis assays of the ***Sup35p*** sequence in specific locations and tested for amyloidogenesis in the expressed cultures, resulting in estimations of the propensities of amino acids in PrDs. A two dimensional analysis, complementing the prion propensity estimations with calculations of intrinsic disorder, was also used to improve the classification method. This methodology has been successfully used to generate synthetic prion-like sequences that were able to form aggregates and propagate on *in vivo* experiments[39]. As stated above, this methodology by Toombs *et al.*, displaying a fairly high classification accuracy when compared to other available methodologies, rely on the random mutation of just two short segments of 19 and 7 amino acids of ***Sup35p***, a domain of almost 100 residues with long glutamine and asparagine-rich stretches. As a consequence, it is possible that the mutational space is not completely explored, which could result in a model not well suited to scan large sets of protein sequences. This is evident from the results they present in their reports[28,39], on which they are unable to perform genome-wide searches with their methodology, but instead rescue prions from sets previously filtered based on intrinsic disorder. In contrast, our model is based in the sequences of almost all the known proteins displaying prion-like behavior and we have demonstrated that our method can perform as well as PAPA[§] for differentiating real and false prions. The bootstrapping assay, see Figure 2.2, also proves that the propensities obtained are unbiased.

[§] Available at: `http://combi.cs.colostate.edu/supplements/papa/`

2.3.2 Putting the Algorithm in Context: Analyzing Real Sequence Datasets

Most of the algorithms used to predict Q/N-rich prion candidates[28,36–38] have a common downside: they lack a proper statistical calibration of the methodology and thus an estimation of the predictive capability of the model to scan sequence databases. In some cases, protein sequences have been modeled as a Poisson[36] or a binomial[37] distributions to calculate the probability of occurrence of glutamine and asparagine in a peptide, and its statistical significance. These approximations have two main problems; the first is that they exclude the positive or negative contributions of all other amino acids to the prionogenicity of the domain. And the second is that not even a normalized probability of occurrence for the Q/N composition of a stretch guaranties a good classification performance in terms of number of false positive prions that will be returned to rescue a desired number of true prions. Our position-independent model accounts for the positive contribution of Q and N to prion induction, but also for the favorable contribution given by S and Y, and for the unfavorable contribution of C, E or W, among others (see Table 2.1). Our model corresponds to an unsupervised learning classifier that represents almost all the rules describing real prion-forming domains, also appending the negative contribution of uncontiguous prolines. An increase in the number of PrDs sequences available for the training, as well as the inclusion of supervised training to add biologically relevant information to the model, such as organism-specific information of the distribution of prolines in the domains or the intrinsic β-aggregation propensity of the sequence, might improve the predictive potential of our model.

We have confirmed here that our strategy performs reasonably well at recovering known prions from large datasets of protein sequences, which makes it very appropriate to make predictions at genomic scale. The method shows a consistent performance even for 500-fold skews towards the negative instances distribution, see Figures 2.4 and 2.5, suggesting that the compositional information embodied in the model can efficiently discriminate between prions and non-prions in variable-size protein sequences databases. This is important if the goal is to predict Q/N-rich domains in small genomes of just a few hundred proteins, as well as in the larger eukaryotic genomes.

The benchmarking of our algorithm also gives us the opportunity to obtain statistically the confidence intervals within which we can predict prions in complete proteomes. The choice of a classification cutoff score is always subjective, but an analytical approach permits to ascertain the composition of the recovery sets during the search of a database, and also enables controlling the inherent tradeoff between $Precision$ and recall[64] –*i.e.* TPR– defined in Equations (2.6) and (2.4). Here we decided to set the cutoff high at 50 $bits$, as depicted in Figure 2.6, in accordance to the maximum prediction accuracy and to diminish as much as possible the rate of false positives included in the predictions.

We were primarily concerned about obtaining a high number of fall-outs that could mislead the implications of our work. The false discovery rates obtained support the fairly good classification ability of the algorithm, that minimizes down to 16% the proportion of non-prions passing the cutoff.

It is also interesting that with our scoring model we found compositional similarities between some IDPs[65–67] and prions. Amino acid composition has been used in the past to predict IDPs[65,68–70], and those studies have concluded that such domains are enriched in K, E, P, S and Q, and depleted in W, C, Y, G and N[68]. The propensities calculated in this study represent in some cases a compositional bias similar to those found in IDPs, –*i.e.* enrichment in Q and S and the depletion in C and W. This might be the reason causing the superposition of the right tail of the Disprot score distribution with that of PrDs, see Figure 2.4. Based in those similarities, we can argue that most of the false positive predictions recovered in a predictive tryout would be natively disordered proteins. There are also experimental evidences suggesting that certain intrinsically disordered proteins might in fact propagate like prions[71,72], including α-synuclein[73], the Aβ peptide[74] and huntingtin[75], involved in Parkinson, Alzheimer and Huntington diseases, respectively. Huntingtin is predicted to posses a PrD, whereas Aβ and α-synuclein are not included in our dataset. However, it is still a matter of debate whether these two proteins are disordered or contain a significant α-helical content[76,77]. Therefore, it could be that our method can correctly classify proteins in the superposed zone between the two distributions, and that some of the predictions tagged as false positives could be in fact prions. However, in general terms, the amino acid propensities of the rest of residues is rather different between IDPs and PrDs, which determines that, in most cases, our algorithm can accurately discriminate between these domain types.

2.3.3 Discovering Putative Prion-like Domains in Complete Proteomes

Although generally thought as linked to disease, prions are also associated with central cellular functions and have been well studied in fungi and some microorganisms, where they play important roles as epigenetic elements[78,79], evolutionary capacitors[14,80] and bet-hedging devices[81,82] in the processes of adaptation to environmental fluctuations. There are also evidences suggesting that, even in invertebrates, prions take part in mechanisms crucial to maintain long-term physiological states[83–85]. However, our knowledge of prions in higher organisms is limited to a handful of examples associated to serious illnesses, thereby the need for strategies that can point out new putative candidates that might be coupled to other cellular functions. The decisive step of a predictive methodology is always the discovery of new instances resembling a given model under some statistical restrictions. Our model, and most importantly the outcomes of the calibration process that proves that our methodology can be used to scan large databases without losing accuracy, gave us the opportunity to scan all the available proteomes. This distinguishes our work from previous attempts in a few specific organisms. The 27925 predictions in 3236 different organisms from all the evolutionary classes[¶] represents, to our best knowledge, the most extensive set of PrD predictions obtained so far, which will help to attain a global view of the distribution of prion domains in the proteomes of organisms, and to unravel the cellular processes in which proteins containing different prion-forming domains might be involved.

Our results show that, in general terms, the number of prions per genome is low, though there are organisms in which prion-like self-assembly might play important functions, as can be inferred from the rather high number of prions in their genomes. It is important to bear in mind that there could be a significant bias in these estimations, when associated with annotation problems of some genomes. The analysis of incomplete sequenced genomes of some members of the genus *Plasmodium* proved that they contain abundant hydrophilic low-complexity segments, which correspond to species-specific, rapidly diverging regions that might be forming non-globular domains that help the parasites to evade the host's immune response[47]. Here we demonstrate this trend by analyzing the complete proteomes of various members of this genus, and propose that most of these stretches may correspond to PrDs, see Table 2.3 for some examples of this case and those that follows. We also found a similar tendency in the genome of *Dictyostelium discoideum*, by far the organism with more predicted prions in its proteome, which implies that most of the low-complexity stretches found in the sequencing of the genome of this organism[46] could be prions, though the functional implications of such an amount of aggregation-prone proteins is unclear. Having a high number of low-complexity stretches appears to be characteristic of these organisms[86]. Accordingly, despite being less represented than in *Dictyostelium discoideum*, the number of PrDs in *Dictyostelium purpureum* genome is fairly high in comparison with that in other organisms. It is known that *Plasmodium* is able to survive with an aggregation-prone proteome even under the periodic heat shock stress that characterizes malaria, where patients suffer recurrent episodes of fever exceeding $40°$C. This is possible thanks to the presence of specialized chaperones, which are essential for parasite survival within red cells[87]. So far, only one of our *Plasmodium* PrDs candidates has been characterized experimentally: ***PFI115w*** (Uniprot ID **Q8I2S1_PLAF7**). In agreement with our prediction, the protein aggregates intracellularly when expressed in human cells[87]. *Plasmodium* chaperones act as cellular capacitors allowing the accumulation of potentially deleterious PrDs, whose

[¶] These figures correspond to the predictions generated analyzing Uniprot release of June 2013 presence should therefore provide certain advantage to the organism. It is still to discover whether *Dictyostelium* exploits a similar strategy to cope with the high aggregation load of its proteome.

Saccharomyces cerevisiae is the most studied organism regarding amyloid formation, and there are various predictive strategies reporting putative PrDs in its complete proteome[28,38,88]. Here we have not only improved the scoring capability of previous methodologies[38], but have also provided an ample list of PrD predictions, including more than 500 completely new predictions in the yeast proteome. The molecular chaperone ***Hsp104*** is essential for the propagation of known yeast prions, which cannot be propagated in cells devoid of the chaperone. The current model of amyloid propagation suggests that the prion fibrils need to be shortened or cleaved by ***Hsp104*** in order to be transmitted to the progeny during cell division[89]. Therefore, one should expect a certain correlation between the ability of ***Hsp104*** to propagate prionogenic species and the number of PrDs in the proteome of this organism. Despite its homology with the *S. cerevisiae* chaperone, it has been shown that the *Schizosaccharomyces pombe* ***Hsp104*** is unable to propagate the [*PSI*+] prion[90]. Interestingly enough, only 3 putative PrDs were identified in the genome of *S. pombe*. This is in contrast with *Candida albicans*, the yeast with the largest number of predicted PrDs after *S. cerevisiae* (169 domains), whose ***Hsp104*** chaperone supports [*PSI*+] prion propagation[91].

Prions can be defined as proteins able to shift between their soluble and aggregated states. This equilibrium should be tightly regulated in the cell, since the accumulation of aggregated species is inherently toxic and linked to the onset of a variety of human disorders. We explored the GeneCards database[92] to identify links between PrD predictions and human disorders. Remarkably, most of the human proteins for which protein function has been reported appear to be strongly linked to severe diseases, including different neuropathies and cancers, see Table 2.4. This suggests that physiological conditions or genetic mutations disrupting the balance between soluble and insoluble species in human prion candidates might lead to localized pathological conditions. Moreover, owing to the predicted prion-like nature of these proteins, it is possible that, once formed, the seeds might spread to other locations. Thus, impeding the aggregation and/or subsequent dissemination of the identified candidates might constitute a way to tackle these, in most cases, intractable disorders.

TABLE 2.4: Association Between Proteins Bearing PrD Predictions and Diseases in Human

Gene	Disease
ATXN1	Spinocerebellar Ataxia Huntington's disease
ATXN3	Machado-Joseph disease Spinocerebellar Ataxias
ATXN8	Spinocerebellar Ataxia type 8
BMP2K	Internuclear ophthalmoplegia Ulnar neuropathy
FOXP2	Speech-language disorders Blepharophimosis Premature ovarian failure Autism Dyslexia
HTT	Huntington's disease Spinocerebellar Ataxia
MAML	Mucoepidermoid carcinoma Hidradenoma Lipoadenoma Epithelial-myoepithelial carcinoma
MED12	FG syndrome Intellectual disability Schizophrenia
MED15	Epicondylitis
NCOA3	Breast cancer Ovarian carcinoma
PAXIP1	Spinocerebellar Ataxia
TAF15	Chondrosarcoma Peripheral primitive neuroectodermal tumor Amyotrophic lateral sclerosis Sarcoma Liposarcoma
TOX3	Breast cancer
TPB	Spinocerebellar ataxia Tuberculosis

Continued on next page...

TABLE 2.4: (continued)

Gene	Disease
	Huntington's disease

We compiled the different diseases associated with the genes in humans for which we found PrD predictions from GeneCards database[92]

2.3.4 Prion-like Domains are Associated to Specific Protein Functions, Processes and Locations in Different Organisms

The analysis of the predictions generated in this study, a large amount of data for an ample set of proteins from different organisms, would be quite difficult from an operative perspective if it is not organized in an efficient way for data mining. From our data, testing if a given protein is a prion is straightforward, but more interesting questions, such as which proteins from those encoded in the genome of an organism are prions and which are the main biological processes, molecular functions and cellular components they are involved or located in, would be more difficult to answer. Even more complex transversal questions could come to mind, such as trying to analyze the distribution of prion proteins in different groups of organisms corresponding to different or close evolutionary categories, combining this information with information of function and spatial localization in the cells, could be of great importance. To try to answer those questions and provide the scientific community with a resource in which to extend our initial ideas to study prion biology at a genomic scale we have developed PrionScan[‖]. An example of the use of the wealth of the information stored in PrionScan is presented here for the analysis of our prion predictions in the different proteomes combined to Gene Ontology[93] annotations, which contains the more complete classification of proteins into functional classes, biological processes and cellular locations. This analysis has uncovered similarities and differences in PrDs distribution among taxa or evolutionary related organisms (Appendix Figures A.1, A.2 and A.3). A first surprising observation is that the predicted PrDs appear to be associated with different cellular components and to work in different biological processes in different taxa and organism groups. These data are consistent with the view that the common switching mechanism underlying prion behavior can be exploited for different physiological purposes[13].

In bacteria, PrDs are depleted in the intracellular space and significantly enriched at the cell wall. Accordingly, bacterial PrDs appear to be essentially involved in metabolic and catabolic processes resulting in construction and disassembly of the cell wall. No

[‖] Available at: http://webapps.bifi.es/prionscan prion protein has been characterized yet in bacteria. However, many bacterial species form extracellular biofilms[94–97], which are constituted, among other components, by proteins assembled into amyloid structures identical to those in neurodegenerative disorders[98–100]. Amyloidogenic proteins in biofilms are constituents or interact with the bacterial cell wall[98–100] and it is known that biofilms are important virulence factors for bacteria, favoring the attachment to eukaryotic cells[94,95,101–104]. Importantly, biofilm forming pathogens such as *Staphylococcus aureus*[101–103], presents the highest content in PrDs among bacteria, suggesting that the identified proteins might contribute to form or sustain the network of amyloid contacts that stabilize the biofilm. Preliminary experimental data support this view since the predicted *S. aureus* PrD ***SSAA2*** forms *bona fide* amyloid fibrils *in vitro* (unpublished results from our group). Bacterial amyloids can initiate the formation of pathogenic or misfolded amyloid upon interaction with diverse host proteins[105]. This template-directed process resembles prion transmission and brings up a possible relationship between bacterial infections and neurodegenerative diseases. Accordingly, bacterial amyloids cause the development of amyloidosis when they are injected in susceptible mice[106].

In eukaryotes, PrDs are intracellular and preferentially localized in the nucleus, as previously suggested[107]. In yeasts and plants, PrDs are found associated with the transcription factor II D component[108–110] (TFIID), a protein complex composed of the TATA binding protein (TBP) and a set of TBP associated factors (TAFs), well conserved across species. Binding of TFIID to DNA is necessary for transcription initiation in most RNA polymerase II promoters. Accordingly, in both taxa, a large number of PrDs are linked to the transcriptional function. In fungi 86 PrDs are involved in catalyzing release of nascent polypeptide chains from the ribosome, a function similar to that exerted by ***Sup35***. Overall, both in fungi and plantae PrDs are enriched in DNA and RNA-binding proteins, controlling apparently unrelated processes such as nitrogen utilization in fungi and hormone –*e.g.* auxin and ethylene– signaling pathways in plants.

In animals, PrDs are also essentially nuclear and depleted in both the mitochondrial and plasmatic membrane, consistent with a soluble nature under physiological conditions. They are also underrepresented in mitochondrion, in agreement with the observation that bacteria contain a reduced number of PrDs. Also, in animals the majority of PrDs corresponds to DNA and RNA-binding proteins. In vertebrates, PrDs are overrepresented in two important functional components: the mediator[111–114] and the histone acetyltransferase[115–117] complexes. Mediator is a multiprotein complex that functions as a transcriptional coactivator in all eukaryotes[118]. In fact, we also find PrDs linked to mediator in yeast. The mediator complex is required for activation of transcription of most protein-coding genes, but can also act as a transcriptional co-repressor. In humans, it includes proteins such as ***MED12*** and ***MED15***[111,114,118], which, as discussed previously, are linked to debilitating disorders. Histone acetylation is also linked to transcriptional activation and associated to euchromatin[115–117]. Histone acetyl-transferases can also acetylate non-histone proteins, such as transcription factors and nuclear receptors to facilitate gene expression. The DNA/RNA binding properties of mammal PrDs determine that most of them act in the control of transcriptional and translational processes. In humans, these proteins include transcriptional factors (PAX-interacting protein 1, ***TOX3***), tumor suppressor proteins (***MN1***), histone methyl/acetyl-transferases (Histone-lysine N-methyl-transferase ***MLL2***, E1A-binding protein ***p400***) and nuclear receptors (***NCOA3***), and they function in essential pathways such as beta cadherin mediated *Wnt* signaling or estrogen response.

Overall, in animals, protein bearing PrDs appear to work in the upstream regulation of central biological processes and more specifically in development. In almost all cases putative prion proteins appear related to biological regulatory processes involving the formation of supramolecular complexes implicating an ample number of proteins and DNA[111–118], in which these prionogenic domains could play an important role in establishing the interactions stabilizing those complexes[108,109,119–121], and providing great versatility allowing the formation of complexes with different composition depending on the environmental conditions. In vertebrates, prions might act in the development of central nervous regions such as the putamen, caudate nucleus or the neural crest. This regulatory activity of neuronal development is conserved between mammals and humans, where prion proteins may additionally play a role in cerebellum and cerebral cortex development. Therefore, it is likely that PrDs malfunction might be intimately linked to the apparition of neurodegenerative diseases, as previously discussed (Table 2.4). Mammal and human PrDs are also involved in embryonic development and more generally in cell differentiation, which might explain the association of PrDs with different types of cancer (Table 2.4).

Interestingly, 30% of the predictions in humans were found in proteins of unknown function. If we combine all the predictions obtained in this study for all the analyzed organisms, the percentage of PrDs predictions in proteins of unknown function raises to 56%. Therefore, our results could be of help to uncover new potential targets for experimental analysis and to unravel the yet-to-discover functional implications of these proteins. As one major challenge in the field of prion prediction is the lack of good datasets on which to train and test potential algorithms[122], methodologies and databases like the ones presented here could be of great help for providing ample sets of putative proteins and domains for experimental tests. This would help increasing the reliability of predictive approaches, and guide us to a better definition of the sequence determinants that lead prion formation. As our knowledge of the fundamental features of prion formation and propagation, and the relationship between prion activity and disease grows, application of this knowledge to prion prediction will lead to more accurate prediction methods and identification of new prions or prion-like proteins, potentially resulting in additional targets for treating human neurodegenerative disorders[122].

2.4 Conclusions

In this work, we have developed a probabilistic model to predict prion domains based on the primary sequence of proteins. By using this model, which is combined with a thorough benchmarking and calibration to handle genome-size sequence databases, we have been successful on predicting prions in all the proteomes available, which to our knowledge constitutes the most extensive study in this direction performed so far. We have disclosed an ample list of proteins containing stretches with a fairly high compositional similarity to those of known prions, including proteins from almost all the evolutionary classifications and taxa, from archaea and viruses to mammals and human. Our results also show that this kind of domains is found in an ample and diverse group of evolutionarily unrelated proteins. In fact, our predictions highlight some interesting trends in the distribution of prion domains in different protein functional families, different cellular compartments and involved in dissimilar biological processes depending on the taxonomic classification. In a time in which prion biology is a rather unexplored field, and the number of prion proteins confirmed experimentally is scarce, predictive approaches such as ours could be of great help to pinpoint putative prionogenic proteins for further experimental characterization. We have included all our predictions in a database with a simple and flexible query system, which allows the mining of our data for study the distribution and functional implication of prions at a genomic scale in all the proteomes annotated in sequence databases. Thus, the free distribution of these predictions, as well as the continuous updating and improvement of the predictive models based on new experimental evidence, might significantly contribute to increase the understanding of prion biology, and to reach a deeper understanding of prions' functional and regulatory mechanisms.

2.5 Methodology

2.5.1 Sequence Datasets

A group of 29 proteins that proved heritable switch and significant *in vivo* amyloid formation in yeast[38] was used as the training set for obtaining the amino acid propensities in prion domains. We calculated the propensities based on the complete sequences that were cloned and tested experimentally in this work, which we believe, is more credible than using the predicted PrD-cores, which are inferred solely based in statistical precepts, see Appendix Table B.1 for a complete list. Another set of 18 high scoring prion predictions, all of which had also been experimentally tested and showed no prion-forming propensity in any of the four assays[38], was used as the negative evaluation set in the benchmarking of the methodology, see Appendix Table B.2. The positive evaluation set for the ROC plot analysis was formed with the 18 out of the 29 prions used to construct the model that resulted positive in all the four assays described in the work by Alberti *et al*[38]. In order to avoid artifacts due to the use of intersected sets of positive instances for training and testing, we also performed an exhaustive jackknife bootstrap assay to estimate the significance of the amino acid propensities obtained. In this bootstrap assay, we resampled with replacement one million subsets from the positive set of 18 prion proteins, randomly excluding half of the prions each time. We then regenerated the model with the remaining 9 prions and used the excluded instances as the positive test set for the ROC plot construction, while the negative set was the same set of 18 negative sequences in all cases. Accordingly, a million ROC plots were built and processed to obtain the average curve and the errors associated to the estimations in each point of the curve, as depicted in Figure 2.2.

We also defined three additional evaluation datasets, comprising the Uniprot/Swissprot database[40] (release from February 2012), a culled list of proteins with solved tridimensional structure annotated in SCOP (version 1.75) obtained from the ASTRAL compendium[123] (including proteins with less than 95% sequence similarity) and all the intrinsically disordered proteins annotated in Disprot[42] (version 5.7). In the case of the Uniprot/Swissprot dataset we randomly generated a million sets that were used in the benchmarking, while for the other two databases we used all the protein sequences annotated. In all cases, the known prions were removed from the negative datasets. These three test sets were used to measure the ability of the model to handle sequence datasets with a high number of negative instances, as it is the case of the scanning of complete proteome databases.

2.5.2 Construction of the Probabilistic Model

The amino acid frequency propensities obtained from the known PrD training dataset described above were used to build an independent log-likelihood model of prion-forming domains. In this model we assume that composition and not primary sequence determines the principal properties of PrD[29,30], thus we choose a model in which the position of amino acids in a given sequence is irrelevant. The observed frequencies were transformed into statistical potentials by using the following expression:

$$LO_{r_i} = \log_2 \frac{f_i}{p_i} \tag{2.1}$$

in which LO_{r_i} is the log-odds ratio of amino acid r_i in $bits$, f_i is the observed frequency of this amino acid in the training set and p_i is the corresponding expected frequency in the protein universe –*i.e.* frequency of amino acids in all known proteins reported in Swissprot. The resulting statistical potentials for all the amino acids are shown in Table 2.1. Assuming complete independence among the positions of a sequence fragment of a certain length, these log-odds can be summed up to return an uncalibrated score associated to the fragment, for which the higher the score the higher the probability that the sequence is a PrD. With this model, that is essentially a 'classifier' for mapping instances into a specific class, we scanned protein sequences with a sliding-window approach using the expression:

$$Score_L = \sum_{l=1}^{L} LO_{r_l} \tag{2.2}$$

where the $Score$ of a protein sequence segment of length L is obtained accounting for the relative support of each amino acid independently.

We added a correction to the score based on the number and distance between non-contiguous prolines found in the PrD. It has been previously reported that the relative abundances of the different amino acids, and not the specific sequence, is related to the prionogenicity of a given sequence stretch[28–30]. However, prolines display important differences with the other amino acids because they cause a characteristic structural disruption of secondary structures, and it has been suggested that the abundance of non-contiguous prolines decreases the prionogenicity of a given sequence[28]. Thus, we set up a strategy in which we estimated the relative abundance of proline pairs separated a given distance –*i.e.* between one and sixty residues in accordance with the scanning window defined. In order to do so we parsed a set of 4606913 sequences included in UniRef 50, release of February 2012. This database contains clusters of sequences extracted from Uniprot/Swissprot[40] and is both representative of the protein universe and non-redundant, as it only contains sequences with less than 50% sequence identity. From this assay we were able to obtain the relative frequency of proline patterns, see Figure 2.1, and we used those frequencies to obtain the corresponding log-likelihoods for each proline pattern, taking into consideration the corresponding expected frequencies. Then, we obtained the final corrected score using the following formula:

$$Score_{L_{(corr)}} = \sum_{l=1}^{L} LOr_{r_l} + \sum_{p=1}^{P-1} LOr_{(d_p - d_{p+1})} \tag{2.3}$$

in which the second addend accounts for the significance of non-contiguous prolines in the sequence. The resulting corrected scores were used in the benchmarking and predictive stages of our methodology.

2.5.3 Benchmarking of the Classification Methodology

The classifier performance was assessed with the positive and negative sets described above in this Methodology section. The real prionogenic sequences –*i.e.* positive test set as included in Appendix Table B.1– were analyzed in combination with a set of non-prion sequences –*i.e.* negative test set as included in Appendix Table B.2– and the ability of the classifier to correctly rank the positive instances in the pool of negative cases was tested. The following statistical performance metrics were calculated to follow the benchmarking progress:

$$TPR = \frac{TP}{TP + FN} \tag{2.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$Accuracy = \frac{TP + TN}{P + N} \tag{2.7}$$

$$FDR = \frac{FP}{FP + TP} \tag{2.8}$$

where TP, FN, FP, TN stands for true positives, false negatives, false positives and true negatives respectively. These variables were used to calculate the false positive (FPR) and true positive (TPR) rates, needed for constructing the receiver operating characteristics (ROC) curves. The $Accuracy$, $Precision$ and false discovery rate (FDR) were also calculated. The areas under the ROC curves (AUC) were calculated non-parametrically using the trapezoid algorithm. All the statistical analysis was done using the R suite[124] and a library of *ad hoc* Perl scripts developed by us.

2.5.4 Predicting Q/N-rich Putative Prion Proteins in Complete Proteomes

We downloaded the complete proteomes of all the organisms sequenced so far from the Uniprot/Knowledgebase database[40] to identify novel proteins bearing prion-forming domains. These repositories include four-weekly updates of proteins resulting from genome sequencing and annotation projects and are subdivided in two complementary and non-redundant datasets: **a**) Swissprot, for fully annotated curated entries and **b**) TrEMBL, formed by computer-generated entries enriched with automated classification and annotation. This subsection of Uniprot is organized in separate files for different taxonomic divisions, which give us the opportunity to study the compositional characteristics of our predictions in each evolutionary clade. In this dataset, there is a file for each taxon, including all the proteins for organisms belonging to that taxon, except for rodents, mammals and human, which are distributed in individual files each. These files were processed with an *ad hoc* perl script included in Appendix Script C.1. The proteins passing the cutoff defined in the predictive methodology based on the amino acid composition of a continuous stretch of sixty residues[38] –*i.e.* what was proposed to be a typical length of PrD-cores– were accepted as predictions. All the predictions are stored and are publicly available in the PrionScan site[**]. The predictions obtained were analyzed to estimate the number of proteins with PrDs in all the taxa studied, belonging to different ontology classifications[93] in the following sub-categories: Molecular Function, Biological Process and Cellular Component. Also, in order to estimate the significance of the number of predictions in a given classification, we set up a tryout in which we calculated the expected number of each GO term by randomizing the selection 1×10^6 times and then estimating the $z-scores$ for each GO term parametrically. These results are included in Appendix Figures A.1, A.2 and A.3.

2.5.5 Construction and Design of a Database of Predicted Prion Proteins

2.5.5.1 Data Acquisition and Database Organization

Our primary source of information is Uniprot[40], the standard and most complete repository of protein sequences freely available. Following each update of this database once a month, we thoroughly scan all the entries included both in Swissprot and TrEMBL in the search for prion-like domains according to our model. In parallel, we also extract some relevant information from Uniprot for those entries containing putative prion domains, and store it in our database. The data generated during the prediction process comprises the score of the highest scoring window during the scan of a protein sequence,

[**] Available at: `http://webapps.bifi.es/prionscan`

the sequence of the highest scoring domain, the localization of the highest scoring putative prion-domain and the complete scanning profile of the protein sequence. This data is merged with the information extracted from Uniprot entries, including the entry identifier and accession number, the organism and taxon, the protein names, the Gene Ontology[93] GO Terms for the molecular functions, biological processes and cellular component in which the protein is related/located and finally, cross-references to other databases with relevant information for the protein bearing putative prion domains.

All this information is stored in a MySQL database environment, allowing linking all the information at all possible levels to enable the efficient querying of the database for fine-grained data retrieval. A description of the data in the present version of PrionScan, including the predictions for the Uniprot (update 2013_07) of June 2013, is shown in Figure 2.7.

2.5.5.2 The PrionScan Website

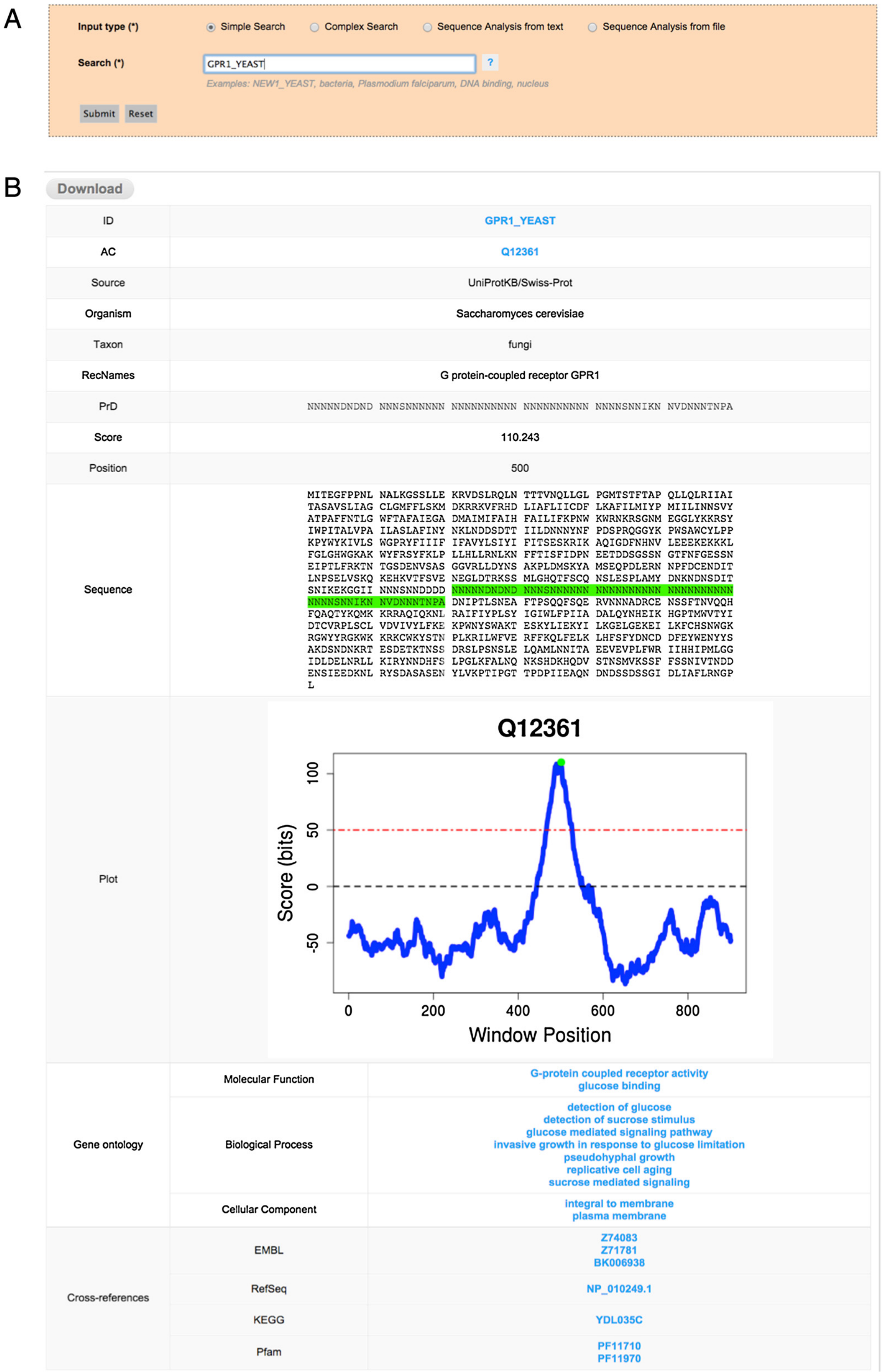
PrionScan is hosted in an Apache web server that relies on a PHP bundle to connect the client query patterns with the database and a set of *ad hoc* Perl scripts that perform some functions, such as the prediction of prion domains in the client's own sequences and the connection to our computer cluster for processing a large number of client sequences. The system processes the client searches and data submission and generates dynamic HTML pages designed to be completely functional in the main web browsers. The home page of PrionScan contains a short introduction to our method and the functionalities of the site to guide the users in a glimpse, and the **Submission Form** organized in checkboxes to easily select the different searching alternatives (**Simple** or **Complex Searches**) and the two different ways of submitting sequences to be analyzed with our method (**Sequence Analysis** from **text** or **file**), please see Figure 2.8, panel A. There is also a link in the leftmost vertical menu to a page containing detailed help and guidance on the methodology, the searchable fields of the database and the output generated. Furthermore, in order to facilitate the use of our site without the requirement of a full reading of the Help page, we also enabled the auto completion utility in the **Simple Search** tab and added hover help buttons for in-site help.

2.5.5.3 Querying the Database

PrionScan is configured to be searched in two different ways:

- **Simple Searches:** The easiest way for retrieving information when the user wants to find out whether a specific protein contains prion-like domains. In this case

FIGURE 2.8: The Simple Search Option for Direct Access to Protein Prion Prediction Information



A

Input type (*) ● Simple Search ○ Complex Search ○ Sequence Analysis from text ○ Sequence Analysis from file

Search (*) GPR1_YEAST ?

Examples: NEW1_YEAST, bacteria, Plasmodium falciparum, DNA binding, nucleus

Submit Reset

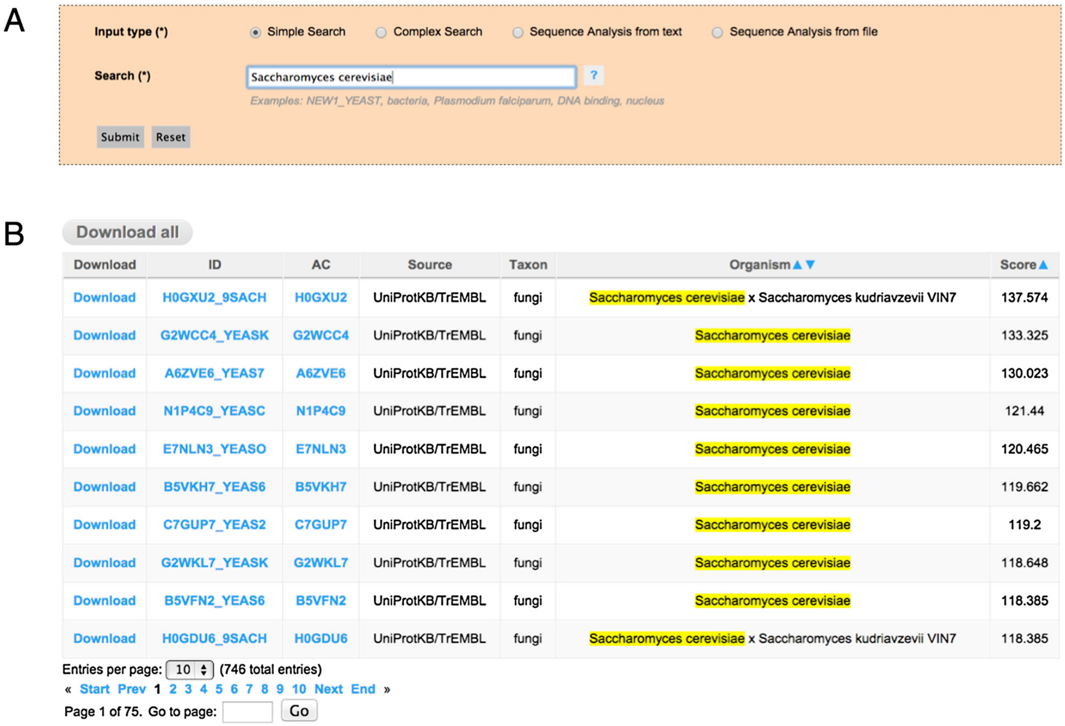
B

Download

ID	GPR1_YEAST	
AC	Q12361	
Source	UniProtKB/Swiss-Prot	
Organism	Saccharomyces cerevisiae	
Taxon	fungi	
RecNames	G protein-coupled receptor GPR1	
PrD	NNNNNDNDND NNNSNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNSNNIKN NVDNNNTNPA	
Score	110.243	
Position	500	
Sequence	MITEGFPPNL NALKGSSLLE KRVDSLRQLN TTTVNQLLGL PGMTSTFTAP QLLQLRIIAI TASAVSLIAG CLGMFFLSKM DKRRKVFRHD LIAFLIICDF LKAFILMIYP MIILINNSVY ATPAFFNTLG WFTAFAIEGA DMAIMIFAIH FAILIFKPNW KWRNKRSGNM EGGLYKKRSY IWPITALVPA ILASLAFINY NKLNDDSDTT IILDNNNYNF PDSPRQGGYK PWSAWCYLPP KPYWYKIVLS WGPRYFIIIF IFAVYLSIYI FITSESKRIK AQIGDFNHNV LEEEKEKKKL FGLGHWGKAK WYFRSYFKLP LLHLLRNLKN FFTISFIDPN EETDDSGSSN GTFNFGESSN EIPTLFRKTN TGSDENVSAS GGVRLLDYNS AKPLDMSKYA MSEQPDLERN NPFDCENDIT LNPSELVSKQ KEHKVTFSVE NEGLDTRKSS MLGHQTFSCQ NSLESPLAMY DNKNDNSDIT SNIKEKGGII NNNSNNDDDD NNNNNDNDND NNNSNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNSNNIKN NVDNNNTNPA DNIPTLSNEA FTPSQQFSQE RVNNNADRCE NSSFTNVQQH FQAQTYKQMK KRRAQIQKNL RAIFIYPLSY IGIWLFPIIA DALQYNHEIK HGPTMWVTYI DTCVRPLSCL VDVIVYLFKE KPWNYSWAKT ESKYLIEKYI LKGELGEKEI LKFCHSNWGK RGWYYRGKWK KRKCWKYSTN PLKRILWFVE RFFKQLFELK LHFSFYDNCD DFEYWENYYS AKDSNDNKRT ESDETKTNSS DRSLPSNSLE LQAMLNNITA EEVEVPLFWR IIHHIPMLGG IDLDELNRLL KIRYNNDHFS LPGLKFALNQ NKSHDKHQDV STNSMVKSSF FSSNIVTNDD ENSIEEDKNL RYSDASASEN YLVKPTIPGT TPDPIIEAQN DNDSSDSSGI DLIAFLRNGP L	
Plot		
Gene ontology	Molecular Function	G-protein coupled receptor activity glucose binding
	Biological Process	detection of glucose detection of sucrose stimulus glucose mediated signaling pathway invasive growth in response to glucose limitation pseudohyphal growth replicative cell aging sucrose mediated signaling
	Cellular Component	integral to membrane plasma membrane
Cross-references	EMBL	Z74083 Z71781 BK006938
	RefSeq	NP_010249.1
	KEGG	YDL035C
	Pfam	PF11710 PF11970

A) The Simple Search option is used for querying the database using the Uniprot identifier of a given protein. **B**) The Detailed Output Page retrieved by the query for a protein with putative prion domains. At the top there is a button for complete download of the results

FIGURE 2.9: The Simple Search Option for Searching with Text Keywords



Input type (*) ● Simple Search ○ Complex Search ○ Sequence Analysis from text ○ Sequence Analysis from file

Search (*) Saccharomyces cerevisiae

Examples: NEW1_YEAST, bacteria, Plasmodium falciparum, DNA binding, nucleus

Download	ID	AC	Source	Taxon	Organism ▲▼	Score ▲
Download	H0GXU2_9SACH	H0GXU2	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7	137.574
Download	G2WCC4_YEASK	G2WCC4	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	133.325
Download	A6ZVE6_YEAS7	A6ZVE6	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	130.023
Download	N1P4C9_YEASC	N1P4C9	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	121.44
Download	E7NLN3_YEASO	E7NLN3	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	120.465
Download	B5VKH7_YEAS6	B5VKH7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	119.662
Download	C7GUP7_YEAS2	C7GUP7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	119.2
Download	G2WKL7_YEASK	G2WKL7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	118.648
Download	B5VFN2_YEAS6	B5VFN2	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	118.385
Download	H0GDU6_9SACH	H0GDU6	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7	118.385

Entries per page: 10 (746 total entries)

« Start Prev 1 2 3 4 5 6 7 8 9 10 Next End »

Page 1 of 75. Go to page: Go

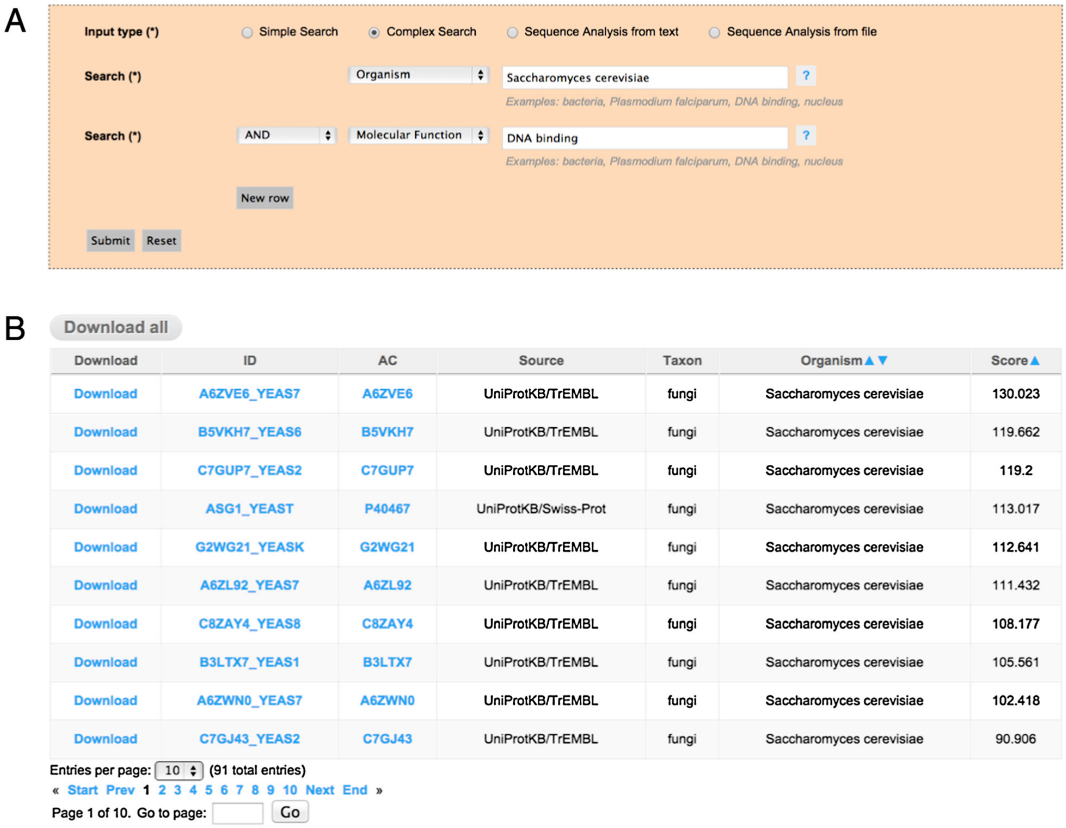
A) The Simple Search option is used for querying the database with a text keyword –*i.e.* the complete name of an organism in this case. **B**) The General Output Page retrieved by the query with rows corresponding to multiple entries in the database –*i.e.* putative prion proteins in the genome of this organism– each one redirecting to a Detailed Output Page. At the top there is a button for complete download of the results and at the bottom there is a summary of the number of predictions and a functionality for browsing throughout multiple pages containing all the results returned

it is possible to directly access the information of a single protein providing its Uniprot **identifier** or principal **accession number**, as shown in Figure 2.8, panel A. This option is also the best alternative for querying the database with information from one of the searchable fields **Taxon**, **Organism Name**, **Protein Name** (Recommended Name, Alternative Name and Submitted Name) and the Gene Ontology Terms for **Molecular Function**, **Biological Process** and **Cellular Component**. For example, it is possible to retrieve all the putative prion proteins in the genome of an organism by providing the complete or partial organism name, as shown in Figure 2.9, panel A.

- **Complex Searches:** Sometimes, however, more complex searches are needed, especially when the user has more detailed information of the set of proteins to be retrieved. In those cases the search can be refined by combining multiple fields from the database –*i.e.* Taxon, Organism Name, Protein Name (Recommended Name, Alternative Name and Submitted Name) and the Gene Ontology Terms for Molecular Function, Biological Process and Cellular Component. These fields can be combined when needed, by introducing the search terms in the rightmost tabs, and selecting the appropriate field that should be considered in the leftmost tabs. You can also choose the logical operators combining the query instances. Using this option, it is possible, for example, to retrieve all the prion-like proteins having a similar Molecular Function or related to a specific Biological Process in the genome of a specific organism, as depicted in Figure 2.10, panel A.

- **The Output:** After performing a search for a specific protein using its Uniprot identifier or principal accession number, if the protein selected has prion-like domains the output will be a **Detailed Output Page** including the Uniprot identifier (ID) and principal accession number (AC), the source (Source) of the protein (coming from Swissprot or TrEMBL), the organism name (Organism) and taxon (Taxon), the names of the protein (recommended names: RecName and/or alternative names: AltName and/or submission names: Subname), the highest scoring prion domain in the sequence (PrD), the score of the highest scoring prion domain (Score), the position in the protein sequence of the highest scoring prion domain (Position), a representation of the complete protein sequence with the highest scoring prion domain highlighted in green (Sequence), and a graphical representation of the scanning of the complete protein sequence (Plot), corresponding to a chart with the score profile along the sequence, also showing the score used for making the predictions (Figure 2.8, panel B). In addition to these fields, the **Detailed Output Page** might also include information regarding the Gene Ontology Terms associated to the protein for the Molecular Function, Biological Processes and/or Cellular Component and the Cross-references to other databases like the EMBL, Refseq, Pfam and so on, lower part of Figure 2.8, panel B. However, if the search, either a **Simple Search** or a **Complex Search**, retrieves more than one entry, the output will be a **General Output Page** with columns and rows that could contain different information depending on the search conducted, with some columns enabled to be dynamically ordered in ascending or decreasing manner (Figures 2.9 and 2.10, panel B). Every row shown in this **General Output Page** redirects to a **Detailed Output Page** as described above. At the bottom part of the **General Output Page** we include a short summary of the number of results retrieved by the query, which is also useful for browsing forward and backwards to different pages in the **General Output Page** by using the page links, or just introducing the exact page in the 'Go to page' box (lower part of Figures 2.9 and 2.10, panel B). Independently of the type of query, it is possible to download the results retrieved in the form of a compressed file containing all the information displayed in the web version, which includes all the information of entries and the associate scanning plots. This information is in HTML format and can be displayed locally using any web browser. We also include a download version in flat text file format with the same information, that could also be easily parsed by *ad hoc* scripts written by the users for performing in-house massive offline analysis of our data.

FIGURE 2.10: Complex Search Option for Searching with Multiple Text Keywords



Input type (*) ○ Simple Search ● Complex Search ○ Sequence Analysis from text ○ Sequence Analysis from file

Search (*) Organism: Saccharomyces cerevisiae ?

Examples: bacteria, Plasmodium falciparum, DNA binding, nucleus

Search (*) AND Molecular Function: DNA binding ?

Examples: bacteria, Plasmodium falciparum, DNA binding, nucleus

New row

Submit Reset

Download all

Download	ID	AC	Source	Taxon	Organism ▲▼	Score ▲
Download	A6ZVE6_YEAS7	A6ZVE6	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	130.023
Download	B5VKH7_YEAS6	B5VKH7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	119.662
Download	C7GUP7_YEAS2	C7GUP7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	119.2
Download	ASG1_YEAST	P40467	UniProtKB/Swiss-Prot	fungi	Saccharomyces cerevisiae	113.017
Download	G2WG21_YEASK	G2WG21	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	112.641
Download	A6ZL92_YEAS7	A6ZL92	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	111.432
Download	C8ZAY4_YEAS8	C8ZAY4	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	108.177
Download	B3LTX7_YEAS1	B3LTX7	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	105.561
Download	A6ZWN0_YEAS7	A6ZWN0	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	102.418
Download	C7GJ43_YEAS2	C7GJ43	UniProtKB/TrEMBL	fungi	Saccharomyces cerevisiae	90.906

Entries per page: 10 (91 total entries)

« Start Prev 1 2 3 4 5 6 7 8 9 10 Next End »

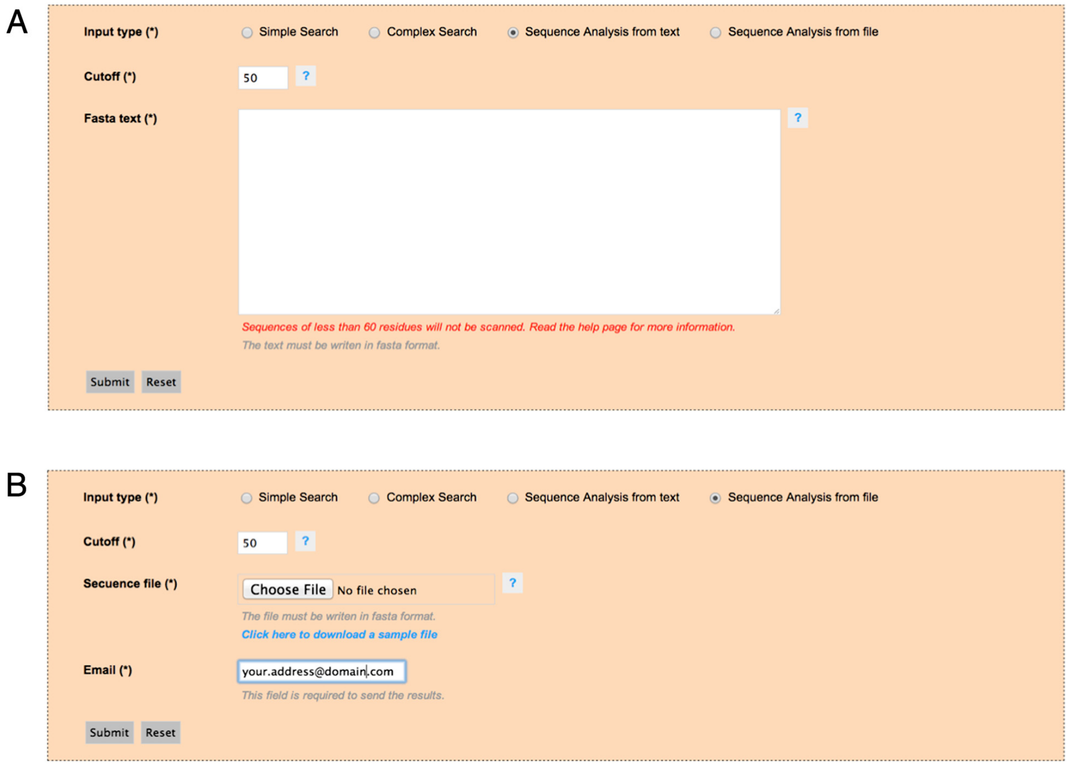
Page 1 of 10. Go to page: Go

A) The Complex Search option is used for querying the database combining the information of two columns of the database –*e.g.* in this case we search the putative prions in the genome of an organism related to a specific molecular function as described in Gene Ontology. **B**) The General Output Page retrieved by the query with rows corresponding to multiple entries in the database –*i.e.* putative prion proteins in the genome of this organism– each one redirecting to a Detailed Output Page. At the top there is a button for complete download of the results and at the bottom there is a summary of the number of predictions and a functionality for browsing throughout multiple pages containing all the results returned

2.5.5.4 Analyzing Your Own Sequences

In this case the user has complete flexibility for testing the prionogenicity of protein sequences using the (**Sequence Analysis** from **text** or **file**) functionalities, as depicted in Figure 2.11. First, the right option in the **Submission Form** is selected in order to enable the option for pasting a limited number of sequences in FASTA format or for

FIGURE 2.11: Sequence Analysis from File and Text for Processing User's Own Sequences



Input type (*) ○ Simple Search ○ Complex Search ● Sequence Analysis from text ○ Sequence Analysis from file

Cutoff (*) 50

Fasta text (*)

Sequences of less than 60 residues will not be scanned. Read the help page for more information.

The text must be writen in fasta format.

Input type (*) ○ Simple Search ○ Complex Search ○ Sequence Analysis from text ● Sequence Analysis from file

Cutoff (*) 50

Secuence file (*) Choose File No file chosen

The file must be writen in fasta format.

Click here to download a sample file

Email (*) your.address@domain.com

This field is required to send the results.

A) The Sequence Analysis from text option in which the user can modify the cutoff used in our methodology to scan the sequences that can be pasted in the text box below. **B**) The Sequence Analysis from file option useful for processing a high number of sequences by submitting the file to be processed in our server. In this case there is an obligatory text box for providing the e-mail address for sending the results upon completion and the cutoff could also be adjusted at users discretion.

uploading a file with a high number of protein sequences, which can be either a flat file or a compressed file in FASTA format (the limit is $500\ MB$ for compressed files, which we estimate can contain approximately one million sequences). We also provide the possibility that the user can select the best cutoff for prediction according to his/her needs. In this case, if only one among the sequences introduced by the user happens to bear prion-like domains, the output will correspond to a **Detailed Output Page** with the specific information for the protein. On the other hand if the analysis of the sequences results in more than one protein with prion domain predictions, then the output will be a **General Output Page** with one row for each protein with predictions. As in the case of results obtained while searching the database, each row redirects to a **Detailed Output Page** with the specific information for the selected sequence. If the number of sequences is less than 5000, the output will be generated in a few seconds in HTML format as just described here, but when the number of sequences is higher than this value, then the job will be submitted to our computer cluster for processing. In this last case the results will be submitted by e-mail to the user upon completion.

2.6 Bibliography

[1] Sergey G Inge-Vechtomov, Galina A Zhouravleva, and Yury O Chernoff. Biological roles of prion domains. *Prion*, **1**: 228–35, 2007. (see pp. 80, 95)

[2] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, **75**: 333–66, 2006. (see p. 80)

[3] Thomas R Jahn and Sheena E Radford. The Yin and Yang of protein folding. *FEBS J*, **272**: 5962–70, 2005. (see p. 80)

[4] Dennis J Selkoe. Folding proteins in fatal ways. *Nature*, **426**: 900–4, 2003. (see p. 80)

[5] Samir K Maji et al. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science*, **325**: 328–32, 2009. (see p. 80)

[6] Douglas M Fowler et al. Functional amyloid formation within mammalian tissue. *PLoS Biol*, **4**: e6, 2006. (see p. 80)

[7] Steffen P Graether, Carolyn M Slupsky, and Brian D Sykes. Freezing of a fish antifreeze protein results in amyloid fibril formation. *Biophys J*, **84**: 552–7, 2003. (see p. 80)

[8] Matthew R Chapman et al. Role of Escherichia coli curli operons in directing amyloid fiber formation. *Science*, **295**: 851–5, 2002. (see p. 80)

[9] J E Podrabsky, J F Carpenter, and S C Hand. Survival of water stress in annual fish embryos: dehydration avoidance and egg envelope amyloid fibers. *Am J Physiol Regul Integr Comp Physiol*, **280**: R123–31, 2001. (see p. 80)

[10] V A Iconomidou, G Vriend, and S J Hamodrakas. Amyloids protect the silkmoth oocyte and embryo. *FEBS Lett*, **479**: 141–5, 2000. (see p. 80)

[11] V Coustou et al. The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. *Proc Natl Acad Sci USA*, **94**: 9773–8, 1997. (see p. 80)

[12] Susan W Liebman and Yury O Chernoff. Prions in yeast. *Genetics*, **191**: 1041–72, 2012. (see p. 80)

[13] Gemma L Staniforth and Mick F Tuite. Fungal prions. *Prog Mol Biol Transl Sci*, **107**: 417–56, 2012. (see pp. 80, 102)

[14] James Shorter and Susan Lindquist. Prions as adaptive conduits of memory and inheritance. *Nat Rev Genet*, **6**: 435–50, 2005. (see pp. 80, 98)

[15] P Chien and J S Weissman. Conformational diversity in a yeast prion dictates its seeding specificity. *Nature*, **410**: 223–7, 2001. (see p. 80)

[16] Eric Karran, Marc Mercken, and Bart De Strooper. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat Rev Drug Discov*, **10**: 698–712, 2011. (see p. 80)

[17] Adriano Aguzzi and Anna Maria Calella. Prions: protein aggregation and infectious diseases. *Physiol Rev*, **89**: 1105–52, 2009. (see pp. 80, 95)

[18] Vittorio Bellotti and Fabrizio Chiti. Amyloidogenesis in its biological environment: challenging a fundamental issue in protein misfolding diseases. *Current opinion in structural biology*, **18**: 771–9, 2008. (see p. 80)

[19] Christopher A Ross and Michelle A Poirier. Protein aggregation and neurodegenerative disease. *Nat Med*, **10 Suppl**: S10–7, 2004. (see p. 80)

[20] S B Prusiner et al. Prion protein biology. *Cell*, **93**: 337–48, 1998. (see p. 80)

[21] Noel G Faux et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Research*, **15**: 537–51, 2005. (see p. 80)

[22] Randal Halfmann et al. Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. *Mol Cell*, **43**: 72–84, 2011. (see p. 80)

[23] Marcus Fändrich and Christopher M Dobson. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J*, **21**: 5682–90, 2002. (see p. 80)

[24] Josephine C Dorsman et al. Strong aggregation and increased toxicity of polyleucine over polyglutamine stretches in mammalian cells. *Hum Mol Genet*, **11**: 1487–96, 2002. (see p. 80)

[25] Helen M Saunders and Stephen P Bottomley. Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng Des Sel*, **22**: 447–51, 2009. (see p. 80)

[26] J Michael Andresen et al. The relationship between CAG repeat length and age of onset differs for Huntington's disease patients with juvenile onset or adult onset. *Ann Hum Genet*, **71**: 295–301, 2007. (see p. 80)

[27] S Choudhry et al. CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum Mol Genet*, **10**: 2437–46, 2001. (see p. 80)

[28] James A Toombs, Blake R McCarty, and Eric D Ross. Compositional determinants of prion formation in yeast. *Molecular and Cellular Biology*, **30**: 319–32, 2010. (see pp. 80–82, 84, 95–97, 100, 107)

[29] Eric D Ross et al. Primary sequence independence for prion formation. *Proc Natl Acad Sci USA*, **102**: 12825–30, 2005. (see pp. 80, 81, 95, 106, 107)

[30] Eric D Ross, Ulrich Baxa, and Reed B Wickner. Scrambled prion domains form prions and amyloid. *Molecular and Cellular Biology*, **24**: 7206–13, 2004. (see pp. 80, 95, 106, 107)

[31] Allen W Bryan et al. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol*, **5**: e1000333, 2009. (see pp. 81, 95)

[32] Antonio Trovato, Flavio Seno, and Silvio C E Tosatto. The PASTA server for protein aggregation prediction. *Protein engineering, design & selection : PEDS*, **20**: 521–3, 2007. (see pp. 81, 95)

[33] Shahin Zibaee et al. A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci*, **16**: 906–18, 2007. (see pp. 81, 95)

[34] Ana-Maria Fernandez-Escamilla et al. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, **22**: 1302–6, 2004. (see pp. 81, 95)

[35] Amol P Pawar et al. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *Journal of Molecular Biology*, **350**: 379–92, 2005. (see p. 81)

[36] M D Michelitsch and J S Weissman. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc Natl Acad Sci USA*, **97**: 11910–5, 2000. (see pp. 81, 83, 95, 97)

[37] Paul M Harrison and Mark Gerstein. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biol*, **4**: R40, 2003. (see pp. 81, 83, 95, 97)

[38] Simon Alberti et al. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell*, **137**: 146–58, 2009. (see pp. 81–83, 86, 87, 93, 95–97, 100, 105, 106, 109)

[39] James A Toombs et al. De novo design of synthetic prion domains. *Proc Natl Acad Sci USA*, **109**: 6519–24, 2012. (see pp. 82, 90, 93, 95, 96)

[40] UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, **41**: D43–7, 2013. (see pp. 82, 106, 107, 109)

[41] Peter W Rose et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research*, **41**: D475–82, 2013. (see p. 82)

[42] Megan Sickmeier et al. DisProt: the Database of Disordered Proteins. *Nucleic Acids Research*, **35**: D786–93, 2007. (see pp. 82, 106)

[43] N Graham. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal . . .*, 2002. (see p. 87)

[44] J Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 2003. (see p. 91)

[45] Y Benjamini. . . The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 2001. (see p. 91)

[46] L Eichinger et al. The genome of the social amoeba Dictyostelium discoideum. *Nature*, **435**: 43–57, 2005. (see pp. 93, 99)

[47] E Pizzi and C Frontali. Low-complexity regions in Plasmodium falciparum proteins. *Genome Research*, **11**: 218–29, 2001. (see pp. 93, 99)

[48] M Nishizawa and K Nishizawa. Local-scale repetitiveness in amino acid use in eukaryote protein sequences: a genomic factor in protein evolution. *Proteins*, **37**: 284–92, 1999. (see p. 93)

[49] G B Golding. Simple sequence is abundant in eukaryotic proteins. *Protein Sci*, **8**: 1358–61, 1999. (see p. 93)

[50] Nils Gehlenborg et al. The Prion Disease Database: a comprehensive transcriptome resource for systems biology research in prion diseases. *Database (Oxford)*, **2009**: bap011, 2009. (see p. 93)

[51] Djamel Harbi et al. PrionHome: a database of prions and other sequences relevant to prion phenomena. *PLoS ONE*, **7**: e31785, 2012. (see p. 93)

[52] Paul M Harrison, Amit Khachane, and Manish Kumar. Genomic assessment of the evolution of the prion protein gene family in vertebrates. *Genomics*, **95**: 268–77, 2010. (see p. 93)

[53] Luke B Harrison et al. Evolution of budding yeast prion-determinant sequences across diverse fungi. *Journal of Molecular Biology*, **368**: 273–82, 2007. (see p. 93)

[54] Vladimir Espinosa Angarica, Salvador Ventura, and Javier Sancho. Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains. *BMC Genomics*, **14**: 316, 2013. (see p. 94)

[55] Roger A Moore, Lara M Taubner, and Suzette A Priola. Prion protein misfolding and disease. *Curr Opin Struct Biol*, **19**: 14–22, 2009. (see p. 95)

[56] Annalisa Pastore and Adriana Zagari. A structural overview of the vertebrate prion proteins. *Prion*, **1**: 185–97, 2007. (see p. 95)

[57] David Eisenberg et al. The structural biology of protein aggregation diseases: Fundamental questions and some answers. *Acc Chem Res*, **39**: 568–75, 2006. (see p. 95)

[58] Randal Halfmann and Susan Lindquist. Screening for amyloid aggregation by Semi-Denaturing Detergent-Agarose Gel Electrophoresis. *J Vis Exp*, 2008. (see p. 95)

[59] Motomasa Tanaka et al. The physical basis of how prion conformations determine strain phenotypes. *Nature*, **442**: 585–9, 2006. (see p. 95)

[60] N Sondheimer and S Lindquist. Rnq1: an epigenetic modifier of protein function in yeast. *Mol Cell*, **5**: 163–72, 2000. (see p. 95)

[61] Eric D Ross, Allen Minton, and Reed B Wickner. Prion domains: sequences, structures and interactions. *Nat Cell Biol*, **7**: 1039–44, 2005. (see p. 95)

[62] Rune Linding et al. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *Journal of Molecular Biology*, **342**: 345–53, 2004. (see p. 95)

[63] S R Eddy. Hidden Markov models. *Curr Opin Struct Biol*, **6**: 361–5, 1996. (see p. 96)

[64] M Buckland. . . The relationship between recall and precision. *Journal of the American society for . . .*, 1994. (see p. 97)

[65] Bo He et al. Predicting intrinsic disorder in proteins: an overview. *Cell Res*, **19**: 929–49, 2009. (see p. 98)

[66] David Eliezer. Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol*, **19**: 23–30, 2009. (see p. 98)

[67] A Keith Dunker et al. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol*, **18**: 756–64, 2008. (see p. 98)

[68] Kang Peng et al. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**: 208, 2006. (see p. 98)

[69] P Romero et al. Sequence complexity of disordered protein. *Proteins*, **42**: 38–48, 2001. (see p. 98)

[70] P Romero, Z Obradovic, and C Kissinger. . . Identifying disordered regions in proteins from amino acid sequence. *Neural Networks*, 1997. (see p. 98)

[71] Christian Münch and Anne Bertolotti. Propagation of the prion phenomenon: beyond the seeding principle. *Journal of Molecular Biology*, **421**: 491–8, 2012. (see p. 98)

[72] Patrik Brundin, Ronald Melki, and Ron Kopito. Prion-like transmission of protein aggregates in neurodegenerative diseases. *Nat Rev Mol Cell Biol*, **11**: 301–7, 2010. (see p. 98)

[73] Christian Hansen et al. -Synuclein propagates from mouse brain to grafted dopaminergic neurons and seeds aggregation in cultured human cells. *J Clin Invest*, **121**: 715–25, 2011. (see p. 98)

[74] Melanie Meyer-Luehmann et al. Exogenous induction of cerebral beta-amyloidogenesis is governed by agent and host. *Science*, **313**: 1781–4, 2006. (see p. 98)

[75] Pei-Hsien Ren et al. Cytoplasmic penetration and persistent infection of mammalian cells by polyglutamine aggregates. *Nat Cell Biol*, **11**: 219–25, 2009. (see p. 98)

[76] Tim Bartels, Joanna G Choi, and Dennis J Selkoe. -Synuclein occurs physiologically as a helically folded tetramer that resists aggregation. *Nature*, **477**: 107–10, 2011. (see p. 98)

[77] C Nerelius et al. Alpha-helix targeting reduces amyloid-beta peptide toxicity. *Proc Natl Acad Sci USA*, **106**: 9191–6, 2009. (see p. 98)

[78] Heather L True, Ilana Berlin, and Susan L Lindquist. Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. *Nature*, **431**: 184–7, 2004. (see p. 98)

[79] H L True and S L Lindquist. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, **407**: 477–83, 2000. (see p. 98)

[80] Joanna Masel and Mark L Siegal. Robustness: mechanisms and consequences. *Trends Genet*, **25**: 395–403, 2009. (see p. 98)

[81] Olivier Namy et al. Epigenetic control of polyamines by the prion [PSI+]. *Nat Cell Biol*, **10**: 1069–75, 2008. (see p. 98)

[82] M M Patino et al. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science*, **273**: 622–6, 1996. (see p. 98)

[83] Sven U Heinrich and Susan Lindquist. Protein-only mechanism induces self-perpetuating changes in the activity of neuronal Aplysia cytoplasmic polyadenylation element binding protein (CPEB). *Proc Natl Acad Sci USA*, **108**: 2999–3004, 2011. (see p. 98)

[84] Paromita Banerjee et al. Short- and long-term memory are modulated by multiple isoforms of the fragile X mental retardation protein. *J Neurosci*, **30**: 6782–92, 2010. (see p. 98)

[85] Kausik Si, Susan Lindquist, and Eric R Kandel. A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell*, **115**: 879–91, 2003. (see p. 98)

[86] Richard Sucgang et al. Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum. *Genome Biol*, **12**: R20, 2011. (see p. 99)

[87] Vasant Muralidharan et al. Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. *Nat Commun*, **3**: 1310, 2012. (see p. 99)

[88] Randal Halfmann, Simon Alberti, and Susan Lindquist. Prions, protein homeostasis, and phenotypic diversity. *Trends Cell Biol*, **20**: 125–33, 2010. (see p. 100)

[89] James Shorter and Susan Lindquist. Hsp104 catalyzes formation and elimination of self-replicating Sup35 prion conformers. *Science*, **304**: 1793–7, 2004. (see p. 100)

[90] Patrick Sénéchal et al. The Schizosaccharomyces pombe Hsp104 disaggregase is unable to propagate the [PSI] prion. *PLoS ONE*, **4**: e6939, 2009. (see p. 100)

[91] Joanna F Zenthon et al. The [PSI+] prion of Saccharomyces cerevisiae can be propagated by an Hsp104 orthologue from Candida albicans. *Eukaryotic Cell*, **5**: 217–25, 2006. (see p. 100)

[92] Gil Stelzer et al. In-silico human genomics with GeneCards. *Hum Genomics*, **5**: 709–17, 2011. (see pp. 100, 102)

[93] M Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**: 25–9, 2000. (see pp. 102, 109, 110)

[94] Luanne Hall-Stoodley and Paul Stoodley. Evolving concepts in biofilm infections. *Cell Microbiol*, **11**: 1034–43, 2009. (see p. 103)

[95] Luanne Hall-Stoodley, J William Costerton, and Paul Stoodley. Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol*, **2**: 95–108, 2004. (see p. 103)

[96] G O'Toole, H B Kaplan, and R Kolter. Biofilm formation as microbial development. *Annu Rev Microbiol*, **54**: 49–79, 2000. (see p. 103)

[97] J W Costerton, P S Stewart, and E P Greenberg. Bacterial biofilms: a common cause of persistent infections. *Science*, **284**: 1318–22, 1999. (see p. 103)

[98] Diego Romero et al. Amyloid fibers provide structural integrity to Bacillus subtilis biofilms. *Proc Natl Acad Sci USA*, **107**: 2230–4, 2010. (see p. 103)

[99] Poul Larsen et al. Amyloid-like adhesins produced by floc-forming and filamentous bacteria in activated sludge. *Appl Environ Microbiol*, **74**: 1517–26, 2008. (see p. 103)

[100] Poul Larsen et al. Amyloid adhesins are abundant in natural biofilms. *Environ Microbiol*, **9**: 3077–90, 2007. (see p. 103)

[101] Jeremy M Yarwood et al. Generation of virulence factor variants in Staphylococcus aureus biofilms. *J Bacteriol*, **189**: 7961–7, 2007. (see p. 103)

[102] Karen E Beenken et al. Global gene expression in Staphylococcus aureus biofilms. *J Bacteriol*, **186**: 4665–84, 2004. (see p. 103)

[103] Jeremy M Yarwood et al. Quorum sensing in Staphylococcus aureus biofilms. *J Bacteriol*, **186**: 1838–50, 2004. (see p. 103)

[104] D G Davies et al. The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science*, **280**: 295–8, 1998. (see p. 103)

[105] D Otzen and P H Nielsen. We find them here, we find them there: functional bacterial amyloid. *Cell Mol Life Sci*, **65**: 910–27, 2008. (see p. 103)

[106] Katarzyna Lundmark et al. Protein fibrils in nature can enhance amyloid protein A amyloidosis in mice: Cross-seeding as a disease mechanism. *Proc Natl Acad Sci USA*, **102**: 6098–102, 2005. (see p. 103)

[107] Julien Couthouis et al. A yeast functional screen predicts new candidate ALS disease genes. *Proc Natl Acad Sci USA*, **108**: 20881–90, 2011. (see p. 103)

[108] K Ferreri, G Gill, and M Montminy. The cAMP-regulated transcription factor CREB interacts with a component of the TFIID complex. *Proc Natl Acad Sci USA*, **91**: 1210–3, 1994. (see pp. 103, 104)

[109] G Gill et al. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc Natl Acad Sci USA*, **91**: 192–6, 1994. (see pp. 103, 104)

[110] B F Pugh and R Tjian. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes Dev*, **5**: 1935–45, 1991. (see p. 103)

[111] Tilman Borggrefe and Xiaojing Yue. Interactions between subunits of the Mediator complex with gene-specific transcription factors. *Semin Cell Dev Biol*, **22**: 759–68, 2011. (see pp. 103, 104)

[112] Stefan Björklund and Claes M Gustafsson. The yeast Mediator complex and its regulation. *Trends Biochem Sci*, **30**: 240–4, 2005. (see pp. 103, 104)

[113] Ronald C Conaway et al. The mammalian Mediator complex and its role in transcriptional regulation. *Trends Biochem Sci*, **30**: 250–5, 2005. (see pp. 103, 104)

[114] Sohail Malik and Robert G Roeder. Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci*, **30**: 256–63, 2005. (see pp. 103, 104)

[115] Michael J Carrozza et al. The diverse functions of histone acetyltransferase complexes. *Trends Genet*, **19**: 321–9, 2003. (see pp. 103, 104)

[116] Shelley L Berger. Histone modifications in transcriptional regulation. *Curr Opin Genet Dev*, **12**: 142–8, 2002. (see pp. 103, 104)

[117] B D Strahl and C D Allis. The language of covalent histone modifications. *Nature*, **403**: 41–5, 2000. (see pp. 103, 104)

[118] Amelia Casamassimi and Claudio Napoli. Mediator complexes and eukaryotic transcription regulation: an overview. *Biochimie*, **89**: 1439–46, 2007. (see pp. 103, 104)

[119] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**: 197–208, 2005. (see p. 104)

[120] Richard N Freiman and Robert Tjian. Neurodegeneration. A glutamine-rich trail leads to transcription factors. *Science*, **296**: 2149–50, 2002. (see p. 104)

[121] Hitoshi Okazawa. "Glutamine/Asparagine-Rich Regions in Proteins and Polyglutamine Diseases" in: *Protein Misfolding, Aggregation, and Conformational Diseases* ed. by Vladimir N. Uversky and Anthony L. Fink. vol. 6 Protein Reviews Springer US, 2007. 451–463 (see p. 104)

[122] Sean M Cascarina and Eric D Ross. Yeast prions and human prion-like proteins: sequence features and prediction methods. *Cell Mol Life Sci*, 1–17, 2014. (see pp. 104, 105)

[123] John-Marc Chandonia et al. The ASTRAL Compendium in 2004. *Nucleic Acids Research*, **32**: D189–92, 2004. (see p. 106)

[124] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0 R Foundation for Statistical Computing, Vienna, Austria, 2012. (see p. 108)

CHAPTER 3

Prediction of Local Unstable Regions of Proteins Based on Physicochemical and Geometric Characteristics of Buried Protein Interfaces

Contents

THE RESULTS PRESENTED IN THIS CHAPTER ARE A COMPENDIUM OF THE FOLLOWING RESEARCH ARTICLES:

- Angarica, V.E. and Sancho, J. (2012). **Protein dynamics governed by interfaces of high polarity and low packing density**. (2012). *PLoS ONE*. **7(10):** e48212
- Martinez-Julvez, M., Rojas, A., Olekhnovich, I., Angarica, V.E., Hoffman, P.S., and Sancho, J. (2012). **Structure of RdxA: an oxygen insensitive nitroreductase essential for metronidazole activation in *Helicobacter pylori***. *FEBS J*. **279(23):** 430617
- Ayuso-Tejedor, S., Angarica, V.E., Bueno, M., Campos, L.A., Abian, O., Bernado, P., Sancho, J., Jimenez, M.A. (2010). **Design and structure of a protein folding intermediate. A hint into dynamical regions of proteins**. *J. Mol. Biol*. **400(4):** 922–34

3.1 Introduction

Protein dynamics range from local fluctuations of specific regions[1–3] to large-scale rearrangements involving partial or global unfolding of the native state[4–6]. Fluctuations between alternative structures within the native basin are thought essential for enzyme catalysis and protein recognition[1,3,7], while larger rearrangements may lead to protein misfolding and aggregation[6]. Dynamism and conformational variability are intrinsic to polypeptides and play a central role in protein folding and function[7,8], and also mediate the wide functional versatility of some specialized proteins that exert a remarkable functional multiplicity, and that participate, among others, in central regulatory cell processes[9–15]. Besides, protein dynamics has been proposed to constitute an essential feature of protein evolvability[16]. In fact, the increasing cumulation of more experimental evidence by the recent developed high-throughput techniques, reveals that there is a varied set of cell mechanisms to generate and take advantage of protein structural instability and protein multitasking[17–28]. Traditional views that the biological functions of proteins are carried out by single, well-defined conformations have been abandoned and there is mounting evidence suggesting that function is mediated by ensembles of alternative structures in equilibrium with the 'native state'[29]. Local structural fluctuations have been reported for some enzymes and promiscuous proteins in which multiple conformers contribute to binding a wide range of substrates or partners[1,3,7]. Remarkable flexibility involving wider rearrangements, and even fold transitions, has been described in some proteins where different folding species in equilibrium regulate their biological functions[4,5], in prions that undergo a switch between the soluble and aggregated forms[3], or in proteins that tend to aggregate in specific conditions, causing severe diseases[6].

At present, the intrinsic flexibility and dynamic behavior of individual proteins can be investigated at atomic or residue level, in a one-to-one basis, by using well established techniques, such as a multitude of different NMR approaches[30–38], time resolved X-ray diffraction crystallography[39–45], small and wide angle X-ray scattering[46–53], Φ-analysis[54,55], and pure Molecular Dynamics simulations[56–63] or strategies combining multiple computational and/or experimental approaches[64–69], among others. While these approaches have provided a wealth of information relating structure and dynamics, they are painstaking and cannot be easily applied on a proteome scale, nor can they reveal evolutionary relationships without extreme effort. Free energy estimation-based models, such as COREX[70,71] are useful to predict local properties, such as hydrogen exchange rates[72]. The approach nevertheless requires extensive calculations and the estimation at residue level of a thermodynamic quantity: the free energy of folding, that is very difficult to calculate accurately even using careful parameterizations[73]. Coarse-grained computational models[74–77], such as Elastic Network Models[7,75,78–81], have proven very useful describing slow motions of proteins and have provided strong evidence that those motions are dictated to some extent by the fold geometry. These models, however, do not take into account specific interactions within the protein and, therefore, can offer limited insight into the key physicochemical characteristics of highly dynamic protein *loci*. Thus, there is a need for simple and reliable methods of computational analysis that could help to identify and delineate the boundaries of such regions.

Proteins are generally organized into folding domains, some proteins consisting of just one. The interior of protein domains is well packed on average but significantly heterogeneous, such that tightly-packed regions, usually hydrogen bond-rich[82], coexists with others containing packing defects and cavities[83]. On the other hand, many important cellular processes are mediated by molecular recognition events occurring at protein surfaces that are constantly reshaped by internal motions. Since these motions should be governed by the relative stability of buried interfaces, we hypothesize here that domain cores will contain regions with physicochemical properties specifically suited to ease the reorganization of the contacting segments, hence allowing functionally relevant intradomain motions. We show here that proteins contain buried interfaces of high polarity and low packing density, coined as LIPs: **L**ight **I**nterfaces of high **P**olarity, whose physicochemical properties make them unstable. The structures of well-characterized equilibrium and kinetic folding intermediates indicate that the LIPs of the corresponding native proteins fold late and are involved in local unfolding events. Importantly, LIPs can be identified using very fast and uncomplicated computational analysis of protein three-dimensional structures, which provides an easy way to delineate the protein segments involved in dynamics. Since LIPs can be retained while the sequences of the interacting segments diverge significantly, proteins could in principle evolve new functional features reusing pre-existing encoded dynamics. Large-scale identification of LIPs may contribute to understanding evolutionary constraints of proteins and the way protein intrinsic dynamics are encoded.

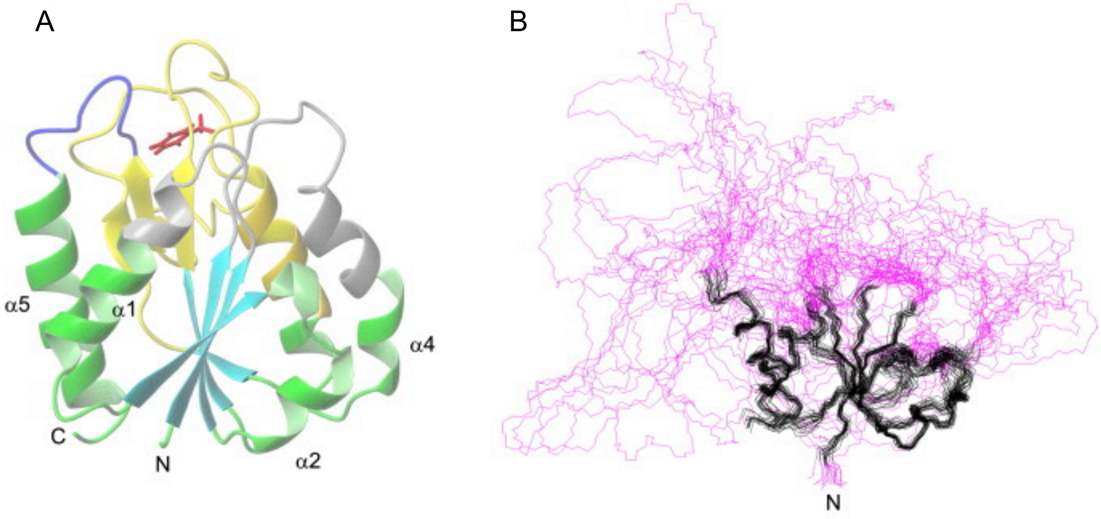
3.2 Results

3.2.1 Identification of LIPs by Means of 3D-structure Analysis

To test our hypothesis, we define protein interfaces as surface patches buried by the interaction of a contiguous protein segment of an arbitrarily defined length with the rest of the protein. Stable protein cores are characterized by a high content of hydrophobic residues, a fine matching of buried polar groups through hydrogen bonding, and tight packing. Most protein buried interfaces should, therefore, be highly apolar and the protein segments at the interface should be tightly packed. In contrast, protein interfaces involved in dynamics need to be intrinsically unstable and should display physicochemical features indicative of a low stability, such as a higher polarity and a lower packing density than stable interfaces.

Our first approximation to this problem came from previous experimental studies of the folding intermediates of the Apoflavodoxin from *Anabaena* PCC 7119[54,59,84]. From the low-resolution characterization of the Flavodoxin intermediate using Φ-analysis[54] we obtained some tentative clues pointing out that regions partially unstructured in the intermediate might be related with structural elements that pack concealing buried interfaces not as apolar as expected for interfaces in the protein interior. The subsequent structural details obtained using NMR[84] and SAXS[59,84] gave further support to that idea, and allowed us to give shape to our initial postulations, because we obtained a precise assessment of the protein *loci* that are unfolded in the intermediate. As can be observed in Figure 3.1, the intermediate adopts a partially unfolded conformation, with a disordered region with no definite structure, as can be inferred from NMR chemical shifts[84]. On the other hand, there is another region that remains in a conformation quite similar to that observed in the native structure. A closer inspection to our proposition can be drawn from the inspection of Figure 3.2. A preliminary study of the physicochemical characteristics of the buried interfaces connecting different segments of the Apoflavodoxin (cyan, green and and yellow subdomains in Figure 3.2) shows that the two dynamic segments of the thermal intermediate (yellow subdomain) correspond to regions exhibiting high polarity ratios, and thus conforming buried interfaces with the rest of the protein of higher polarities than other regions. This observation qualitatively indicates

FIGURE 3.1: 3D Structures of the Wild-type Apoflavodoxin and the F98N Mutant

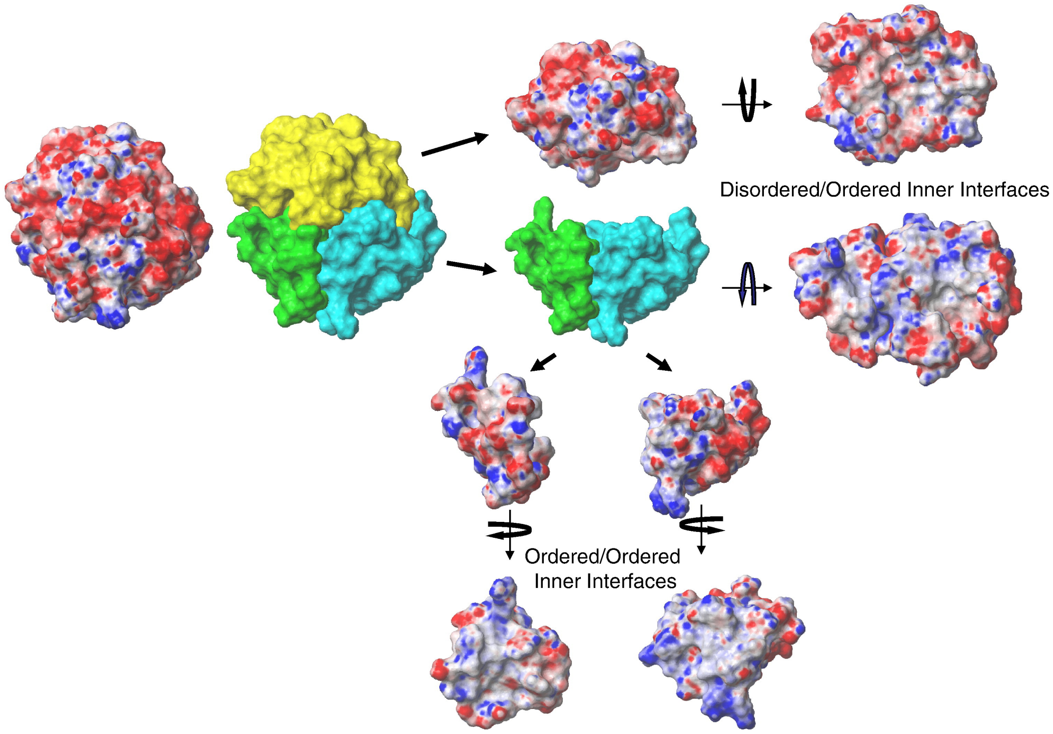


(**A**) Ribbon representation of the wild-type Apoflavodoxin structure (PDB id: 1FTG). Residues assigned for F98N-Apoflavodoxin are displayed in green except for β-strands that are displayed in cyan; residues assigned for wild-type but not for F98N mutant are shown in yellow, and those not found in either wild-type or F98N mutant are colored in gray. The loop corresponding to residues 144–151, whose chemical shifts differ significantly between F98N mutant and wild-type protein, is in blue. Side-chain atoms for residue F98 are shown in red. (**B**) Backbone atoms of the structure calculated for F98N-Apoflavodoxin superposed over residues 2–8, 18–53, 71–86, 109–117, and 153–169. Disordered regions are shown in magenta

that the buried interface in the boundary between the folded and the unfolded regions in the thermal intermediate is unusually polar. Indeed, the electrostatic surface of the well-folded region in the thermal intermediate (cyan and green subdomains) displays a fairly even distribution, on which the buried interface in contact with the disordered region is only slightly less polar than the outer surfaces. In accordance, the buried interface shown by the disordered region when it appears folded in the native structure is rather polar. In contrast, a similar analysis of two halves resulting from splitting the ordered region of the thermal intermediate (cyan and green subdomains) exhibit quite apolar buried electrostatic surfaces, lower panels of Figure 3.2, as is characteristic of well-folded protein cores.

In order to develop a quantitative methodology to analyze the properties of protein buried interfaces for the assessment of these physicochemical singularities, we have probed protein structures using a sliding-window approach. To that end, the three-dimensional structure of a given monomeric protein, as defined by a PDB file, is scanned from end-to-end using a contiguous peptide probe. For each peptide probe, two relevant properties of the probed interface are computed: the ratio of polar/apolar buried area ($Pr_{interface}$), and the packing density ($\rho_{interface}$), as defined on Equations 3.1 and 3.2.

FIGURE 3.2: Polarity of Wild-type Apoflavodoxin Buried Interfaces

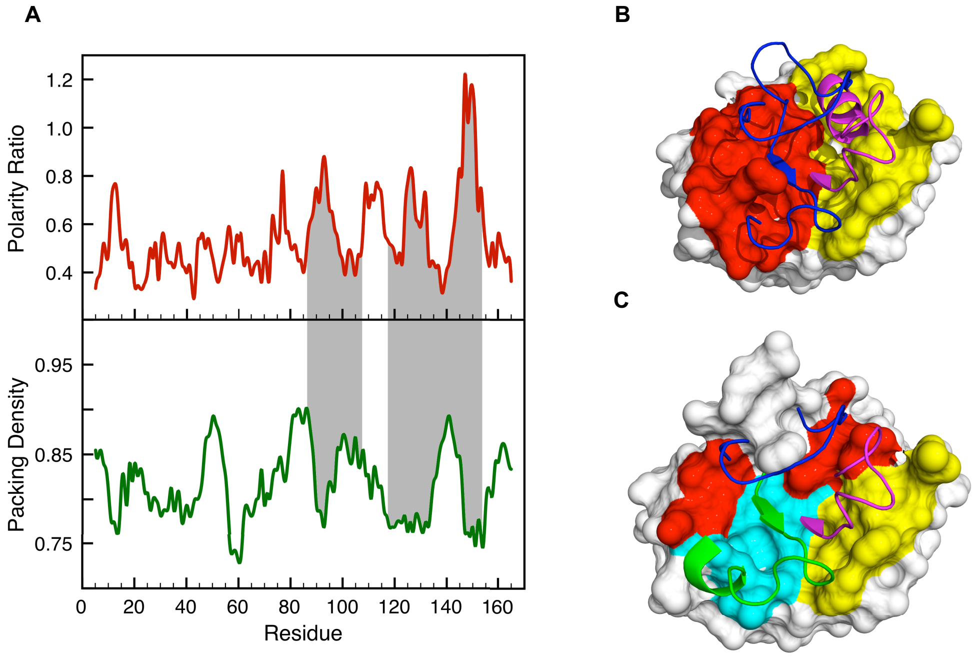


Electrostatic surface of wild-type Apoflavodoxin (PDB id: 1FTG). In the same representation the structure was divided first into two halves corresponding, respectively, to the disordered (in yellow) and well-folded (in green–cyan) regions in the thermal intermediate. These two halves are rotated about $90°$, as indicated, to display front views of the buried electrostatic surfaces between disordered and ordered regions. The half comprising the well-folded region in the thermal intermediate was further subdivided into two sections: one comprising helices $\alpha 1$ and $\alpha 5$ (in green) and the other one comprising the β-sheet and the ordered segments of helices $\alpha 2$, $\alpha 3$, and $\alpha 4$ (in cyan). These two sections are rotated approximately $90°$, as indicated, to exhibit front views of the buried electrostatic surfaces illustrating the low polarity characteristic of the buried interfaces in protein cores

The corresponding computed values are assigned to the central residue of the probe. When the scanning is completed, property sequence profiles are built. The profiles so obtained are not very sensitive to probe length –*i.e.* 7-to-9-long probes give rise to almost identical profiles. However, short probes tend to make the profiles noisier while longer probes tend to average the properties of distant regions that may include both unstable and stable interfaces. We have thus set probe length to eight residues in all the cases reported. We have additionally tested whether the resolution of the structures could affect the outcomes of our method. Basically, the polarity profiles do not change significantly in the 1.2–2.8 Å resolution range, while the packing density profiles retain their shape –*i.e.* the position of maxima and minima– and exhibit slightly lower packing densities in general as the crystal resolution decreases, which is in agreement with previous reports relating lower structural resolution with lower computed packing[85].

Altogether, our predictions based in the polarity and packing profiles are not sensitive to structure resolution.

FIGURE 3.3: Identification and Structural Characterization of LIPs in Apoflavodoxin



A) Stacked-aligned profiles for polarity ratio and packing density in *Anabaena* PCC 71191 Apoflavodoxin (PDB id: 1FTG, Resolution = 2.0 Å). The property values, defined in Equations 3.1 and 3.2, are plotted against the position of the fourth residue of an eight-residue probe fragment. The segments encompassing residues 87–108 and 118–152, which have been found to be unstructured in the equilibrium intermediate of this protein[84], are highlighted in grey. **B**) Surface representation of buried atoms at interfaces 87–107 (yellow) and 118–152 (red) and the associated interacting fragments (in cartoon representation) colored purple and blue, respectively. **C**) Surface representation of the buried atoms according to our characterization of polar light interfaces (LIPs). The LIPs 87–99 (yellow), 120–133 (red) and 140–155 (cyan) are shown and the associated interacting fragments are colored purple, blue and green and are depicted in cartoon representation

TABLE 3.1: Estimations of Solvent Accessibilities for a Group of Proteins

		Folded State		**Core Folded State**		**Unfolded State**	
PDB	$\#aa_s$	ASA_p (Å^2)	ASA_{ap} (Å^2)	ASA_p (Å^2)	ASA_{ap} (Å^2)	ASA_p (Å^2)	ASA_{ap} (Å^2)
1LN4	98	2520.5	3301.0	1647.7	4356.5	4168.2	7657.5
1T1D	100	2763.3	3150.6	2282.2	4348.0	5045.5	7498.6
1BKR	109	2619.5	3445.5	2016.1	5112.1	4635.6	8557.6
1BGF	124	3434.0	4442.9	2240.4	4928.2	5674.4	9371.1
1JB3	131	3231.2	4402.5	2684.9	5431.1	5916.1	9833.6

Continued on next page...

TABLE 3.1: (continued)

		Folded State		**Core Folded State**		**Unfolded State**	
2LIS	136	3197.6	5460.4	2697.5	5536.1	5895.1	10996.5
1QGV	142	3035.7	4085.1	3062.7	7033.9	6098.4	11119.0
1EY4	149	3077.6	4473.3	3540.8	6488.4	6618.4	10961.7
1EP0	185	4700.1	4940.4	3540.8	8802.2	8240.9	13742.6
1L3K	196	4360.3	4956.8	4671.5	8871.2	9031.8	13828.0
1BYI	224	4370.4	6409.4	4053.8	10068.8	8424.2	16478.2
1ES9	232	4151.5	5623.8	5568.9	11141.4	9720.4	16765.2
1II5	233	4039.3	6136.4	5304.8	10982.8	9344.1	17119.2
1WER	334	6882.8	9796.1	7014.2	15632.2	13897.0	25428.3
1FO9	348	6632.8	8261.1	8112.1	17889.1	14744.9	26150.2
1FCQ	350	6425.9	8197.8	9169.0	17305.2	15594.9	25503.0
1E5M	416	6490.2	9165.3	9303.4	19976.4	15793.6	29141.7
1GSO	431	8321.0	10532.7	8111.9	20189.2	16432.9	30721.9
2BCE	579	8673.8	12045.0	13471.8	29697.2	22145.6	41742.2
$\frac{ASA_p}{ASA_{ap}} \pm (SD)$		0.75 ± 0.08		0.46 ± 0.05		0.57 ± 0.04	

From a set of 19 proteins from different folding families, different sizes and sharing less than 20% of sequence similarity we obtained the polar and apolar solvent exposed areas in the folded state (columns 3–4) and in the unfolded ensemble (columns 7–8). We also estimated the polar and apolar buried surface in the core of the folded state (columns 5–6). The averages $\pm SD$ of the ratio of polar and apolar areas for the three states of each protein are indicated in the bottom line of the table

In Figure 3.3, panel A we show the polarity ratio and the packing density profiles corresponding to a representative α/β protein: the Apoflavodoxin from *Anabaena* PCC 7119. The polarity profile represents the ratio of polar over apolar surface area buried at the interface. A baseline with a polarity ratio of around 0.5 can be observed by visual inspection of the profile. Such a baseline is present in all polarity profiles we have built (see Appendix Figures D.1 and D.2, for profiles of other representative proteins) so it appears to be characteristic of protein cores. To confirm that this is the case, we have computed using the ProtSA[*] server[86] the polar and apolar solvent exposed areas in the folded state and in the unfolded ensemble[87], of a representative database[87,88] composed by 19 proteins from different folding families and sharing less than 20% sequence identity. From these data (Table 3.1), we have calculated the polarity ratio characteristic of protein cores at 0.46 ± 0.05. This value indicates that protein interfaces tend to bring

[*] Available at: http://oldwebapps.bifi.es/protSA/

into contact twice as much apolar atom surface than polar atom surface. Importantly, the Apoflavodoxin polarity profile reveals protein segments that form interfaces of a much higher polarity than that of the baseline, one extreme being the interface centered at residue 150, where the contribution of polar atoms to the buried area is even larger than that of apolar atoms –*i.e.* polarity ratio > 1. The packing density profile of the same protein (Figure 3.3, panel A bottom plot) also reveals significant local variations, with packing density minima centered at residues 13, 60, 92, 130 and 153.

It is possible to compare the predictive results obtained with our method with recent experimental data from our laboratory for the unstable regions of the Apoflavodoxin from *Anabaena* PCC 7119 that experience local unfolding at mild temperatures, giving rise to an equilibrium intermediate[54,59,84]. The unstable regions correspond to residues 87–108 and 118–152, while the rest of the protein retains the native conformation in the intermediate, see Figures 3.1 and 3.3, panel B for details. The two unstable regions of the protein are shadowed in grey in Figure 3.3, panel A. These regions include the three peaks with higher polarity ratios. Noticeably, each of those peaks is mirrored by a minimum in packing density and thus represents a low-density –*i.e.* light– interface of high polarity. In the polarity profile, three additional, albeit lower and/or narrower, peaks of high polarity appear centered at residues 12, 77 and 102. It is clear that the peaks at 77 and 102 are not at packing density minima and cannot be defined as 'light polar interfaces'. However, the one centered at residue 12 is at a packing minimum and represents an additional light polar interface. Indeed, this region, while ordered in the X-ray structure due to its association to a phosphate anion, appears disordered in solution even in the native conformation[84,89].

It is important to define light, polar interfaces in a quantitative manner so that the unstable regions of proteins can be predicted in an unbiased way. Since the properties calculated –*i.e.* polarity ratio and packing density– do not provide a value for the local unfolding free energy of the probe sequence, a threshold value must be defined to identify the unstable regions. Analysis of the solvent exposed area of the protein database[87,88] used to compute the polarity baseline described above (Table 3.1), indicates that the polarity ratio of protein surfaces is of 0.75 ± 0.08. This means that protein interfaces exhibiting polarity ratios of 0.8 or greater (0.8 rather than 0.75 is selected for simplicity) are more similar to exposed surface regions than to protein cores. Based on this fact, 0.8 constitutes an appropriate threshold in the polarity profile and we propose that 'light polar interfaces' can be identified as those organized around peaks exhibiting polarity ratios greater than this value. Each of these interfaces is considered to extend on either side of its polarity maximum to include the peak residues with polarity ratios above the 0.5 baseline (0.5 rather than the calculated average value of 0.46 is selected for simplicity). In all cases of proteins studied in this report, these buried interfaces appear located in packing density minima and, in fact, an anti-correlation between polarity ratios and packing densities is observed. While polarity ratios exhibited by protein interfaces or surfaces are expected to depend on general physicochemical properties of the amino acid residues involved and on their relative abundances, there is evidence indicating that packing densities may be related to the specific fold and size of the protein[90,91]. For the protein examples that will be discussed here, the mean values of their interfaces packing densities are different. Thus, we define the regions of low packing density of a given protein as those with values below the mean of its specific density distribution minus two standard deviations.

For Apoflavodoxin, only three peaks above the polarity threshold of 0.8 and located in low density regions are identified in this manner, the corresponding unstable segments being 87–99, 120–133, and 140–155. Since the experimentally determined unstable regions of Apoflavodoxin[84] are 87–108 plus 118–152 (see Figure 3.1), the correlation between light polar interfaces and locally unstable regions is excellent for this particular protein. Figure 3.3, panel B shows that although the unstable regions of Apoflavodoxin are separated in the primary structure, they cluster together in the 3D structure and define a continuous unstable region. Comparison of Figures 3.3, panel B and C allows noticing the structural correlation between the 'light, polar interfaces' of the protein and the experimentally determined unstable regions.

3.2.2 Occurrence of LIPs in All Major Protein Classes

To refer to protein buried interfaces exhibiting a high polarity ratio and a low packing density we have coined the term LIP: **L**ight **I**nterfaces of high **P**olarity. The conservation of LIPs within structurally related proteins can be assessed using multiple sequence alignments to compare property profiles. Superimposition of the polarity ratio profiles corresponding to Flavodoxins of known structure (Appendix Figure D.1, top chart) indicates that they are very similar. The three key peaks characteristic of the Flavodoxin from *Anabaena* PCC 7119 are present in the other Flavodoxins. Similarly, comparison of the packing density profiles (Appendix Figure D.1, bottom chart) indicates that the distribution of packing density heterogeneity in Flavodoxin interfaces is also conserved, which means that light polar interfaces are conserved among Flavodoxins.

Polarity and density analysis of interfaces present in a variety of proteins of known three-dimensional structure indicates that LIPs are present in all protein classes. Examples of conservation of polarity profiles in related proteins of classes α/β (folding TIM α/β barrel), α/β (folding Lysozyme-like), all α (folding Cytochrome c) and all β (folding

Immunoglobulin-like β-sandwich) can be visually assessed in Appendix Figure D.2. Conservation of the corresponding packing density profiles is similarly good in these folds (see Appendix Figure D.3), which indicates that related proteins of a given fold share specific, conserved patterns of LIPs. We notice, however, that more distant proteins with the same fold can display different LIPs patterns, as it is the case for Indole-3-glycerol phosphate synthase, see the corresponding Figure in the sections bellow, and that of Triosephosphate isomerase (Appendix Figure D.2, panel A). Our results show that the polarity and packing density profiles are different for these two enzymes of similar sizes but belonging to two different superfamilies within the TIM α/β barrel fold, and catalyzing rather disparate reactions such as decarboxylation and isomerization respectively.

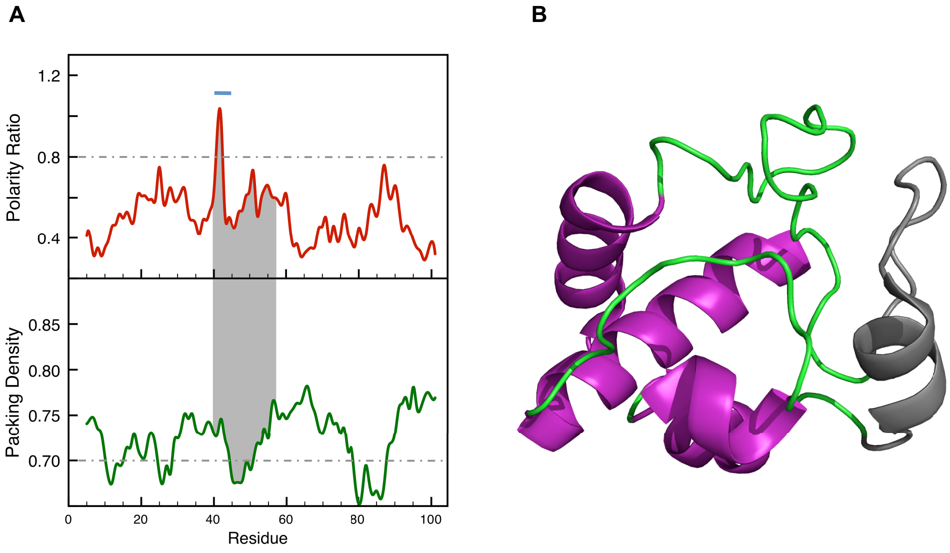
3.2.3 LIPs and Intermediates at the Native State Basin and Beyond

Tight packing, high hydrophobic burial and good pairing of buried polar groups are key ingredients of protein stability[92,93]. LIPs are bound to display high local instability due to their poor packing and low hydrophobicity, which, at least in Apoflavodoxin, is associated to an abundance of buried polar groups not forming hydrogen bonds. Thus, LIPs are expected to experience transient local unfolding events from the native conformation more frequently than other regions, and therefore to contain fast exchanging protons defining unstable foldons.

The correlation between LIPs and unstable foldons identified by their fast proton exchange rates can be illustrated for Cytochrome *c*. The native basin of this protein has been characterized in detail by equilibrium proton exchange[95,96]. Cytochrome *c* contains five foldons, or regions that can experience local unfolding uncoupled from that of the rest of the protein, that have been well defined at residue level. The more unstable one, so-called infrared foldon, comprises residues 40–57[94]. The polarity and packing profiles calculated with our methodology for Cytochrome *c* are shown in Figure 3.4. There is a single peak with polarity ratio higher than the 0.8 threshold, which defines a LIP spanning residues 40–45. Although not at the minimum center, this segment belongs to the wall of a deep packing density minimum including residues 40–55. Thus, the more unstable foldon in Cytochrome *c*, with an unfolding free energy of $4\ kcal/mol$, contains the only LIP present in the protein.

Due to their low stability, LIPs are expected to become unfolded in solution conditions that are nevertheless compatible with the rest of the protein retaining the native conformation. LIPs should therefore correlate with the unfolded regions of equilibrium intermediates. These partly unfolded conformations tend to accumulate at moderately high temperatures or denaturant concentrations, or at extreme pH values, usually low pH.

FIGURE 3.4: LIPs and the Lowest Stability Foldon in Cytochrome *c*

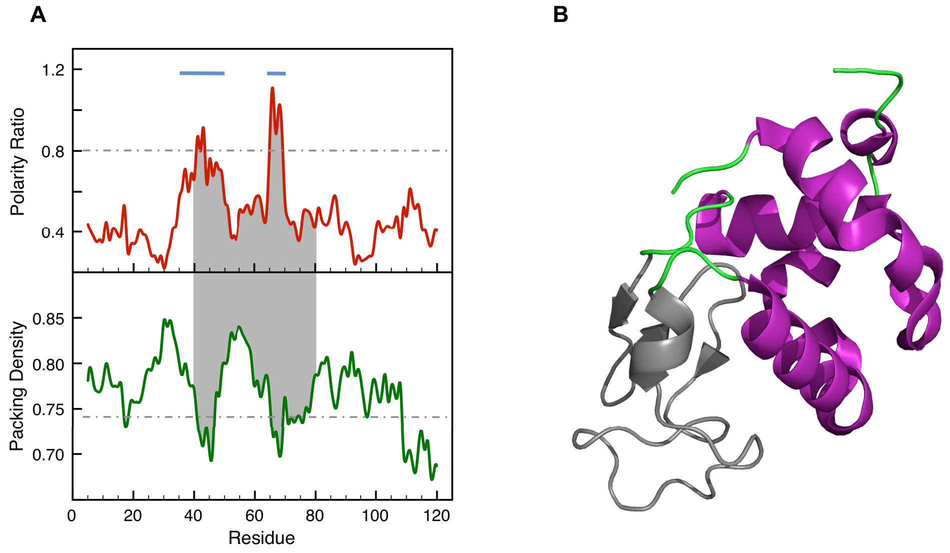


A) Polarity ratio and packing density profiles of Cytochrome *c* (PDB id: 1HRC, Resolution = 1.9 Å). The segment shadowed in grey corresponds to the lowest stability region of the protein (infrared foldon: residues 40–57) according to equilibrium and kinetic H-exchange NMR experiments[94]. The light blue bar indicates the only LIP in Cytochrome *c*, which includes residues 40–45 and is located in the unstable foldon. **B**) Ribbons representation showing the unstable foldon in grey. In the charts, the polarity and packing cutoffs are indicated as grey dashed lines

The Apoflavodoxin intermediate discussed above is a fine example of the autonomous unfolding of LIPs in a thermal intermediate. The free energy difference between this intermediate and the native state is of just 1.5–2.0 $kcal/mol$[84], and the intermediate clearly belongs to the native basin. Not surprisingly, the LIPs in Apoflavodoxin appear located in the functional regions involved in the binding of the FMN cofactor and of partner proteins[97].

A second common type of equilibrium intermediate is the molten globule[99]. Molten globules are typically observed after partial denaturation of certain proteins at low pH, although they have also been described in truncated proteins and in certain apoproteins at neutral pH. Molten globules have attracted attention because they bear similarity with kinetic folding intermediates and because they have been involved in physiological processes, such as membrane translocation. Structural characterization of molten globules is particularly difficult. One of the best-characterized molten globules is that of α-Lactalbumine[98,100], an $\alpha + \beta$ protein organized in two domains. Its molten globule retains a native-like secondary structure at the α domain, but not at the β domain, encompassing residues 40–81. Inspection of the polarity and packing profiles of α-Lactalbumine in Figure 3.5 clearly shows the presence of two LIPs centered at residues 43 and 66, and

FIGURE 3.5: LIPs and the Unfolded Domain of the α-Lactalbumine Molten Globule

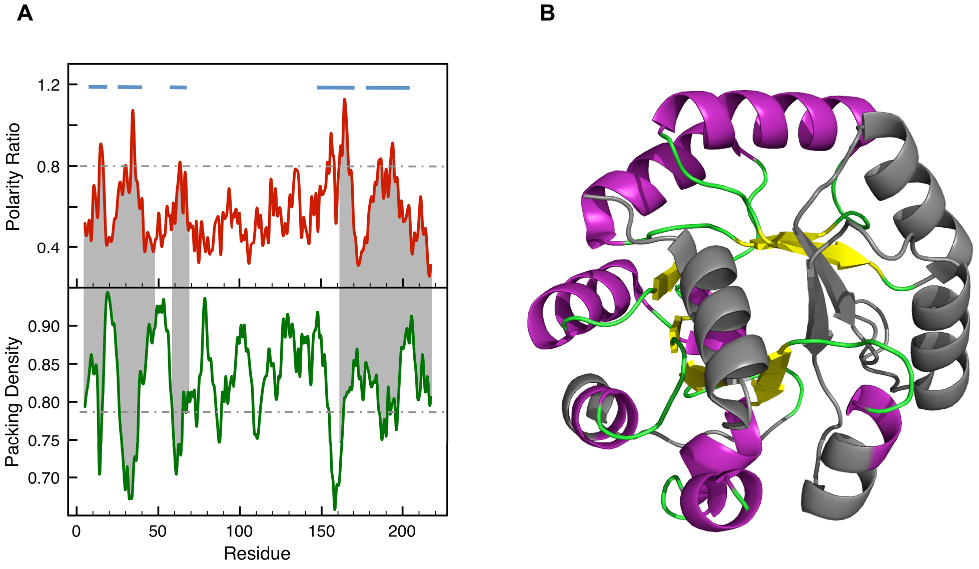


A) Polarity ratio and packing density profiles of α-Lactalbumine (PDB id: 1HML, Resolution = 1.7 Å). The segment shadowed in grey corresponds to the β-domain (residues 40–81), the one that lacks secondary structure in the molten globule intermediate[98]. The light blue bars indicate the two LIPs in α-Lactalbumine, encompassing residues 35–51 and 64–70, and essentially defining the β-domain. **B**) Ribbons representation showing the unstable β-domain in grey. In the charts, the polarity and packing cutoffs are indicated as grey dashed lines

including residues 35–51 and 64–70. These LIPs make the β domain to be the more unstable one and contribute to define the residual structure of the molten globule.

A third common type of equilibrium intermediates is that found in chemical unfolding. The role of LIPs in chemical intermediates can be exemplified by the equilibrium intermediate accumulating in the urea unfolding of **I**ndole-3-**G**lycerol **P**hosphate **S**ynthase (IGPS)[101]. The equilibrium unfolding and the refolding kinetics of this protein have been extensively investigated by hydrogen exchange mass spectroscopy. The equilibrium intermediate accumulates at 5 M urea and consists of two conformations termed I_a and I_b. The more unstable specie (I_a) is folded in the central segment (residues 48–161) and shows little or no protection in the 1–47 and 162–220 segments[101,102]. In addition, the 59–68 loop appears disordered both in the native state and in the intermediate. The limits reported for the central folded and the N- and C-terminal unfolded regions are approximate, as they have been deduced from analysis of peptide fragments. The polarity and packing profiles of IGPS are shown in Figure 3.6. IGPS presents several peaks with polarity ratios higher than the 0.8 threshold that are located in packing density minima. Those centered at residues 15 and 34 define two contiguous LIPs spanning residues 7–18 and 23–40, which nicely correspond to the N-terminal unfolded region (1–47). The next

FIGURE 3.6: LIPs and the Unfolded Regions of the Equilibrium (and Kinetic) Intermediate of Indole-3-Glycerol Phosphate Synthase



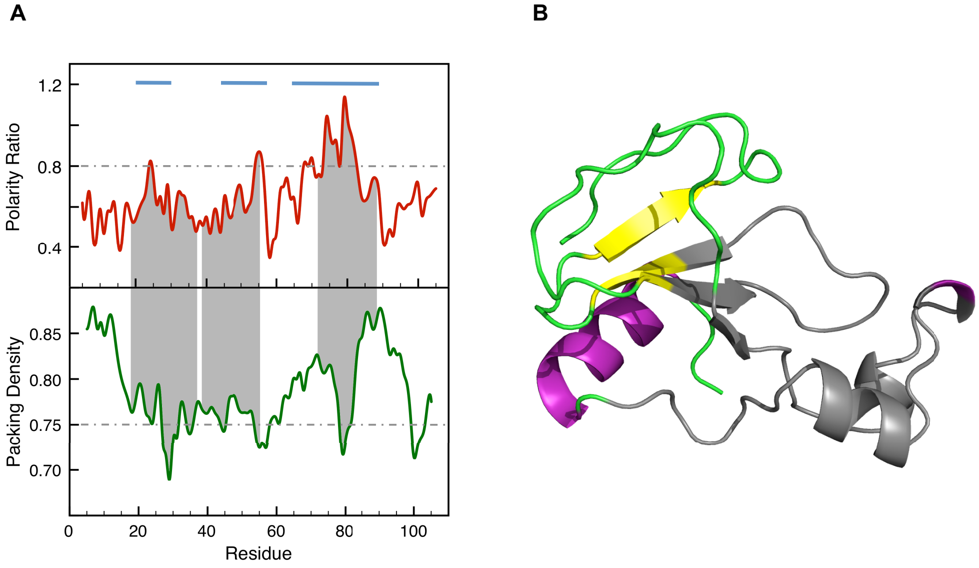
A) Polarity ratio and packing density profiles of Indole-3-Glycerol Phosphate Synthase (PDB id: 2C3Z, Resolution = 2.8 Å). The segments shadowed in grey correspond to the unfolded regions of the equilibrium intermediate of chemical unfolding (intermediate I_a), which coincides with the on-pathway kinetic folding intermediate[101]. The light blue bars indicate the five LIPs in Indole-3-Glycerol Phosphate Synthase. LIPs 7–18 and 23–40 map onto the N-terminal unfolded region of the protein (1–47). The next LIP, 58–68, defines the loop that is unfolded even in the native state (59–68). Finally, LIPs 148–170 and 178–205 are located at the C-terminal unfolded segment of the protein (162–220). **B**) Ribbons representation showing the unfolded regions of the intermediate in grey. In the charts, the polarity and packing cutoffs are indicated as grey dashed lines

LIP (towards the C-terminus) appears at residue 63 and extends on 58–68, in good correspondence with the loop that is unfolded in both the native and intermediate states (59–68). Finally, peaks at 156 and 164, define a single LIP at residues 148–170, while peaks at 186 and 194 define an additional LIP spanning residues 178–205. These two C-terminal LIPs (residues 148–170 and 178–205) are in reasonable agreement with the C-terminal disordered segment of the protein defined from 162–220[101,102]. The structure of the IGPS equilibrium intermediate seems to arise as a consequence of the unfolding of the LIPs present in the native protein.

3.2.4 LIPs and the Protein Folding Reaction

The free energy difference between the IGPS equilibrium intermediate and the native conformation is of $8.5\ kcal/mol$[103]. This intermediate can hardly be considered to be within the native basin or be expected to display a functional role under native conditions. Interestingly, kinetic analysis of IGPS indicates that the structure of this equilibrium intermediate precisely corresponds with that of the on-pathway intermediate of the IGPS folding reaction (intermediate I_a)[101]. On the other hand, the infrared foldon of Cytochrome *c* is also the latest folding region of the protein. Although LIPs have been defined as protein interfaces of the native conformation displaying low stability, and therefore prone to experience local unfolding, it is possible that they also constitute late folding regions of proteins. Both the IGPS and the Cytochrome *c* data point into this direction.

FIGURE 3.7: LIPs in the Late Transition State Ensemble of Barnase Folding



A) Polarity ratio and packing density profiles of Barnase (PDB id: 1A2P, Resolution = 1.5 Å). The segments shadowed in grey correspond to the regions displaying Φ-values equal to or lower than 0.5 in the late transition state of barnase folding (TS2)[104]. The light blue bars indicate the three LIPs in Barnase: 20–30, 44–57 and 65–89. They closely correspond to the segments exhibiting low Φ-values in the transition state (19–37, 39–55 y 72–88). **B**) Ribbons representation showing in grey the transition state regions with low Φ-values. In the charts, the polarity and packing cutoffs are indicated as grey dashed lines

In addition to kinetic intermediates, a key species in protein folding reactions are transition states of folding. These ephemeral conformations are of high energy and can only be characterized by a combination of protein engineering and fast kinetics[55], or by computer simulations. Despite the large energy gap between transition state and native conformations, the available experimental information indicates that the differences are not so large at the structural level. We have thus investigated whether the not-yet folded regions of transitions states could correspond to the LIPs of the native structure. One of the best-characterized transition states of protein folding is that of Barnase. Recently, a combination of Φ-analysis[55] and computer simulation was used to provide a structure of the transition state at the residue level[104]. The nativeness of the structure of a transition state around a given residue is described by its Φ-value. Residues in a fully native or a fully unfolded environment in the transition state will show Φ-values of 1.0 and 0.0, respectively. Barnase folds via a high energy intermediate and therefore two transition states appear in the reaction. The second transition state, connecting the high energy intermediate with the native state, is the one expected to be structurally closer to the native state, and will be compared to the LIPs in the native structure. The polarity and packing profiles of Barnase are shown in Figure 3.7. The segments of the protein exhibiting Φ-values below 0.5 in the transition state (an arbitrary threshold selected to represent the more unfolded regions) are 19–37, 39–55 and 72–88, which are shadowed in grey in Figure 3.7. The barnase LIPs encompass segments 20–30, 44–57 and 65–89, which quite closely correspond to the regions with low Φ-values[104].

3.2.5 Assessing the Statistical Significance of Property Profiles

An important issue to take into account is trying to estimate the statistical significance of the observations reported in this work regarding the special characteristics of buried interfaces related to unstable protein regions. Although the polarity ratio profiles included in this study (Figures 3.3, 3.4, 3.5, 3.6, 3.7) visually show a clear correlation between the LIPs and the conformationally unstable regions at the sequence level, it would be interesting to supply statistical evidence of the differences between the values of polarity ratios obtained for those regions when compared to stable protein segments. We show in Table 3.2 the results obtained for a Mann-Withney-Wilcoxon rank-sum test for comparing the polarities of the buried interfaces of stable regions versus those of unstable ones and versus LIPs. These results demonstrate that the polarity of buried interfaces of unstable regions are statistically different from those calculated for stable ones in all the proteins analyzed. Table 3.2 also shows the expected fact that LIPs, as quantitatively defined above, display a significantly higher polarity than non LIP regions. As can be inferred from the low $\wp - values$ for the comparison of the distributions returned by the test, the alternative hypothesis, determining statistical significant differences between the two distributions, should be accepted in all cases.

We also tried to compare our results with those obtained using a well-established software such as COREX[71,105], in order to test the performance of the two methodologies when processing the same set of proteins. In Figure 3.8, the residue specific stability estimations calculated for the proteins included in this work using COREX is presented. A visual inspection indicates that for these kinds of intermediates, the predictions made by COREX for unstable regions do not correspond in some cases with

TABLE 3.2: Statistical Significance of Interfacial Polarity and of COREX Stability Estimates

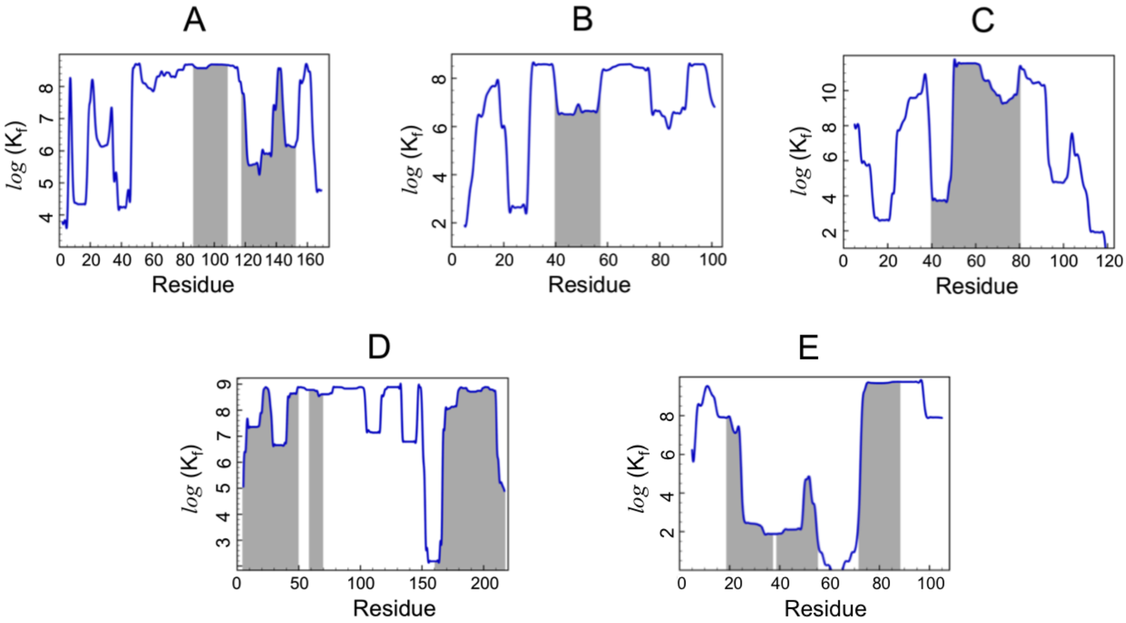
PDB	**Protein**	**Buried Interface Polarity Ratios**	
		Unstable Regions	**LIPs**
		$(\wp - value)$	$(\wp - value)$
1FTG	Flavodoxin	8.585×10^{-5}	6.051×10^{-11}
1HRC	Cytochrome *c*	6.285×10^{-5}	2.397×10^{-2}
1HML	α-Lactalbumin	3.173×10^{-9}	2.383×10^{-13}
2CZ3	IGPS	2.073×10^{-3}	2.200×10^{-16}
1A2P	Barnase	6.761×10^{-3}	2.160×10^{-8}

PDB	**Protein**	**COREX Stability Estimates**	
		Unstable Regions	**LIPs**
		$(\wp - value)$	$(\wp - value)$
1FTG	Flavodoxin	9.266×10^{-1}	4.279×10^{-1}
1HRC	Cytochrome *c*	3.628×10^{-2}	8.073×10^{-2}
1HML	α-Lactalbumin	$1.000 \times 10^{+0}$	6.665×10^{-1}
2CZ3	IGPS	2.839×10^{-2}	5.235×10^{-5}
1A2P	Barnase	5.581×10^{-1}	5.014×10^{-1}

We show the results of a one-sided Mann-Withney-Wilcoxon test performed on the interfacial polarity and the residue specific stability profiles obtained with our methodology and COREX[71,105], respectively. In the first case the alternative hypothesis $H1$ tests the significance of obtaining higher polarities in unstable segments determined experimentally (column: **Unstable regions**) and in the segments corresponding to our definition of LIPs (column: **LIPs**). For the stability estimates obtained with COREX the alternative hypothesis $H1$ tests the significance of obtaining lower stability values in the same protein segments described above. The confidence interval was set to $\wp < 0.05$ in all cases

the regions reported experimentally to be unstable. We repeated the statistical analysis described above to test whether the stabilities calculated by COREX for experimentally unstable regions, were significantly lower than those corresponding to the stable regions of the proteins. The results from this test are included in Table 3.2 and prove that the alternative hypothesis determining significant differences holds only for Cytochrome *c* and Indole-3-Glycerol-P synthase. In these cases, the $\wp - values$ obtained prove that the residue stability values of unstable regions are lower than those of stable regions. However, for the three other proteins there are no statistically significant differences among the distributions. In all cases, the $\wp - values$ obtained are higher than those obtained using our methodology. Not surprisingly, in only one of the five proteins tested (Indole-3-glycerol-P synthase) there is a statistical correlation between COREX predicted unstable regions and LIPs.

FIGURE 3.8: Residue Stability Estimations using COREX



For each protein studied in this report using our methodology we also calculate the local stability using this alternative procedure. As described in the Methodology section, for each protein we first generated the structure ensemble used in the calculations, then determined the entropy-weighting factor before obtaining the corresponding stability constants ($log(K_f)$ in the ordinate axis in each chart). The residue stability obtained are plotted in this figure in the following order: **A**) Apoflavodoxin, **B**) Cytochrome *c*, **C**) α-Lactalbumine, **D**) Indole-3-glycerol-P Synthase and **E**) Barnase. In each case the unstable regions determined experimentally are colored in light grey

3.3 Discussion

3.3.1 Using Buried Interface Physicochemical and Geometric Properties to Define Protein Conformationally Unstable Regions

On the hypothesis that the intrinsic dynamics of protein domains could be related to the presence of buried interfaces of low stability, we have devised a tool that allows to scan protein interfaces and to compute relevant physicochemical properties of the interfaces. Two key properties have been selected as indicative of low stability: the ratio of polar over apolar surface buried in the interface and the packing density at the interface. For a 200-residue protein it takes less than 2 minutes of CPU time in an average personal computer to calculate the polarity ratio and packing density profiles. Therefore, calculation of protein interface properties in a proteome scale is feasible. Our analysis indicates that protein buried interfaces display significant heterogeneity in polarity ratio and packing density. The protein examples discussed in this work contain interfaces, established by contiguous segments of 8 residues, whose polarity ratios vary from 0.3 to 1.2. In all proteins analyzed, a polarity baseline can be observed around 0.5, which appears to be the typical average polarity ratio of protein cores (Table 3.1). Above this baseline, peaks of higher polarity are observed. Since the average polarity ratio of protein solvent exposed surfaces is of approximately 0.8, the peaks with polarity ratios of 0.8 or greater identify the buried interfaces that are more polar than surface exposed regions (Table 3.1).

On the other hand, the packing densities vary from 0.65 to 0.9, with local minima along the profiles, but no obvious baseline value shared by different proteins. Nevertheless, it is clear that most interfaces of high polarity ratio appear at packing density minima below the cutoff established. This can be quantitatively assessed by calculating for each particular protein its average packing density and then determining whether the peaks of high polarity display packing densities below that average minus two standard deviations. Such is the case of 13 out of 14 polar interfaces discussed in this work, the only exception being the red foldon of Cytochrome *c*, which, as explained in the Results section, appears at the wall of a deep minimum.

Proteins thus contain interfaces of high polarity and low packing density. We have termed them LIPs (**L**ight, **I**nterfaces of high **P**olarity), they are expected to exhibit low local stability and they can be easily identified. To test the hypothesis that LIPs are related to the structure of protein folding intermediates, we have defined LIPs in a quantitative manner as buried interfaces including at least one window with polarity ratio greater than 0.8 and extending to those flanking residues with polarity ratios greater than 0.5. In addition, the potential LIP should contain a clear minimum, defined as above, in the packing density profile.

3.3.2 LIPs can be Successfully Used to Predict Local Instability in Different Types of Folding Intermediates

The correspondence between LIPs, so defined, and the unfolded regions in protein equilibrium intermediates of different kinds is excellent. Figure 3.3 illustrates the correspondence of the LIPs in Apoflavodoxin with the unfolded regions of the equilibrium intermediate that accumulates in the thermal unfolding. Figure 3.6 shows the fine correspondence between the LIPs in IGPS and the unstructured segment of the equilibrium intermediate of its chemical unfolding. In Figure 3.5 we show the location of the α-Lactalbumine LIPs in the β domain, the one deprived from secondary structure in the molten globule. LIPs also appear to correlate with unstable foldons exhibiting fast proton exchange from the native state and being late folding regions, as is the case of the infrared foldon of Cytochrome *c*, see Figure 3.4. On the other hand, the LIPs in IGPS also correspond to the not-yet folded regions of its on-pathway folding intermediate, as can be seen in Figure 3.6. It is thus possible that, due to their instability, LIPs can only form on the scaffold provided by the rest of the protein. If this is the case, transition states of protein folding should also display conformations where the LIPs would still be essentially unfolded. The late transition state of Barnase folding (Figure 3.7) illustrates this fact.

Altogether, our analysis reveals that protein domain cores contain interfaces of high polarity and low packing density that appear to be involved in protein dynamics, as they correlate with late folding events and with local instability in the native state, that can lead to alternative partly unfolded conformations. Some of these conformations will be energetically distant form the native state, while others will be close in energy. The latter are expected to populate under native conditions and to get involved in function more easily.

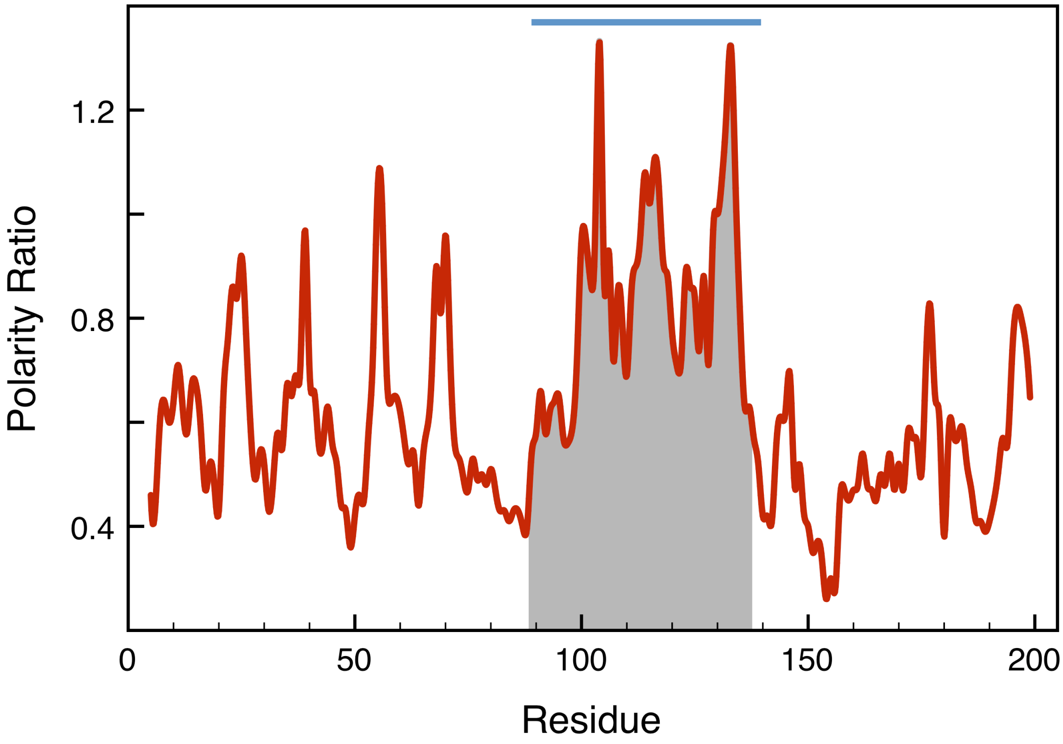
As can be seen in Table 3.2, there is a strong statistical support indicating that the interfaces of unstable regions have a higher polarity than those of stable ones, which confirms that the physicochemical characteristics of buried interfaces can be suitably used to identify conformationally unstable regions with a rather low error rate. The analysis of the polarity profiles in comparison with packing profiles indicates that the latest are less informative, as the fluctuations observed for the packing values are lower in comparison to those observed for interface polarity. This is why we used polarity as our primary source of discrimination. However, as can be seen from our results (Figures 3.3, 3.4, 3.5, 3.6, 3.7), the LIPs correlate in all cases with packing density minima, which is an interesting outcome in agreement with previous reports which had pointed to the relation of packing efficiency with local conformational changes and disorder[16,106]. The reason why the statistical correlation between unstable regions and poorly packed ones is lower is due to the fact that although unstable regions are indeed poorly packed, there are other poorly packed regions that are not particularly unstable, –*e.g.* those exhibiting the characteristic low polarity of protein cores cavities.

Importantly, the computational methodology developed here to identify these proteins' dynamic *loci* is simple and fast. A brief comparison with the COREX algorithm[70,71] seems appropriate because both our structural method and COREX try to capture local differences in protein stability. COREX uses a more complex approach based in constructing an ensemble representation of the protein, which contains a large number of microstates. Since the method has to deal with a huge exponential search space, heuristic strategies are used to simplify the conformational search. Then, the stability of each residue is estimated by computing its free energy of folding from a parameterization of thermodynamic quantities as functions of the surface areas involved[107]. The method has proven to find correlation between calculated stabilities and hydrogen exchange rates[72]. In contrast, our method, does not attempt to provide stability evaluation at a residue level. We only try to identify the segments of the protein whose interaction with the rest of the protein is far from optimal. To achieve this goal we do not calculate free energy values, a very difficult task even if using careful parameterizations because, as it is known, free energy values are typically small and arise from compensation of much larger numbers involving enthalpic and entropic contributions. Instead, we compute simple physicochemical and geometric properties to produce sequence profiles that help to highlight the regions of proteins displaying low local stability. We do not attempt either to give numbers for those stabilities. In this way, our approach is greatly simplified since our ensemble is linear with the number of residues in a given protein. Analysis of a 200-residue protein that takes less than 2 minutes with our method may take one day using the COREX server. One clear limitation of our method is that it does not provide stability values at residue level, however its performance identifying unstable regions related to experimentally characterized equilibrium and kinetic folding intermediates seems good (Figures 3.3, 3.4, 3.5, 3.6, 3.7). To evaluate the performance of COREX towards the same prediction targets we have used the COREX/BEST server[†]. The stability plots calculated for Apoflavodoxin, Cytochrome *c*, α-Lactalbumine, Indole-3-Glycerol Phosphate Synthase and Barnase are shown in Figure 3.8, where the experimentally determined unstable regions are shadowed in grey. Inspection of the figure indicates that for these particular types of intermediates COREX tends to provide a significant number of false positives (regions predicted to be unstable by COREX and not found to be unstable experimentally) together with some false negatives (experimentally unstable regions not predicted as unstable by COREX). The results included in Table 3.2, also prove that the stabilities calculated by COREX for unstable regions were not significantly lower than those corresponding to stable regions in three out of five of the proteins analyzed. In addition, this statistical test shows that, in as much as it can be approximated by the $\wp - values$ returned by the assay, the performance of COREX in distinguishing stable from unstable regions is lower than that of our method.

This approach can also be of help as a complement in the process of 3D-structure solution, when used in conjunction with mutational studies and comparative homology modeling. In the case of the solution of the structure of the oxygen-insensitive nitroreductase RdxA from *Helicobacter pylori*[108], we were able to provide a plausible explanation to the fact that a given region of the protein is missing in the crystal, please see Figure 3.9. The absence of electron density for residues 97–128 (chain A) and 90–133 (chain B) is very likely a result of proteolysis during purification. Mass Spectrometry data corresponding to dissolved RdxA crystals show a main peak at 12174 kDa, with

[†] Available at: http://best.bio.jhu.edu/BEST/index.php

FIGURE 3.9: Detection of Conformationally Unstable Segments in *E. coli* NTR by Interfacial Polarity Analysis



The profile is built by assigning to each residue the ratio of the polar/apolar buried surface area associated with the interaction of an eight-residue flanking peptide with the rest of the protein. The sequence segment encompassing the peaks of highest ratios over the 0.5 polarity ratio baseline is indicated by the blue bar. The packing of this segment with the rest of the protein is predicted to be unstable. The sequence shadowed in grey represents the *E. coli* NTR segment structurally equivalent to the missing helices in the structure of RdxA

two additional components at 12370 and 12530 kDa. Possible H^{+2} peaks appear at 6086 and 6182 kDa. The main peak thus corresponds to approximately half the mass of the RdxA protein (24067 kDa according to the sequence). This indicates that proteolysis of the protein sample has taken place. However, a sample of the same protein preparation that was incubated with NADP+ (and did not crystallize under the same conditions) was shown by SDS/PAGE to have a mass of 26313 ± 10 Da, which is close to the theoretical mass of the protein. The missing segments are occupied by helices F and G in the homologous protein used as search model: the NTR of *E. coli*[109]. In this NTR, helix F is part of a solvent-exposed channel at the dimer interface where FMN lies and helix G is assumed to convey substrate specificity. Helices F and G exhibit high B-values in the structure, and it has been proposed that the mobility of helix F might be important for optimal binding and catalysis[110]. The polarity profile for the NTR of *E. coli* is displayed in Figure 3.9. The more unstable region of this homologous nitroreductase, characterized by its high content of buried polar area, encompasses residues A89–M139, corresponding to residues P91 and I136 of *H. pylori* RdxA, which mimics the missing segment of the structure. This suggests that the missing region in the structure of RdxA will also be locally unstable and more exposed to proteolysis. The proposal that helix F of *E. coli* NTR may exhibit high functional mobility appears to be extended to the equivalent helix of RdxA and also to helix G. Our observation that NADP+ protects RdxA against proteolytic removal of the missing segment agrees with the role assigned to helix F of *E. coli* NTR with respect to binding this cofactor[110].

3.3.3 The Physicochemical and Geometric Characteristics of LIPs are Conserved in Protein Families

Although an analysis of the evolutionary significance of protein LIPs is clearly beyond the scope of this work, we would like to note some features of those buried interfaces that might turn out to be relevant. As we prove in this work, the methodology presented here is useful to identify unstable regions in proteins by means of our definition of LIPs, which in some cases match fairly well the location of unstable regions in proteins. Because those interfaces are simply characterized by displaying outlying values for averaged properties, their evolutionary conservation may not require a high conservation of the sequences involved. To illustrate this fact, structural multiple alignments of Flavodoxin, Cytochrome *c* and α-lactalbumine protein families are shown in Appendix Figure E.1. The average protein identity percentage of these alignments ranges from 34% for Flavodoxins to 50% for α-lactalbumin. Comparison of the alignments with the corresponding polarity ratio and packing density profiles obtained for members of those families (Appendix Figures D.1, D.2 and D.3) shows that the profiles are conserved despite the sequence variation observed. On the other hand, the unstable regions of proteins studied in this work that superposed with LIPs often include protein segments with one side located at the interface and the other side exposed to solvent. Therefore, if the conservation at solvent exposed positions would tend to be lower than at buried ones, the solvent exposed backs of those interfaces could be suited to evolve new functions – *i.e.* recognition of new partners– because they could be mutationally tailored without seriously compromising the intrinsic dynamic nature of the interface. Actually, for the experimentally determined unstable regions and LIPs in Appendix Figure E.1, the column averaged conservation scores estimated using the values obtained with CLUSTAL[111] for solvent exposed residues, are roughly half the averages corresponding to buried residues (see legend of Appendix Figure E.1). This means that in these regions buried residues presented to the interface and responsible for shaping its geometry and determining its physicochemical characteristics are more conserved in average than residues in contact to solvent. These results demonstrate that the trends observed for the conservation of buried and exposed residues obtained for LIPs, which are a methodological definition to identify unstable regions, are shared by the stretches found to be unstructured experimentally. Finally, our preliminary analysis indicates that, as can be inferred from the comparison of the profiles in Figure 3.6 and Appendix Figure D.2 panel A, within a given fold, less closely related proteins corresponding to different functional protein families could contain very different LIPs. It is thus possible that the different distribution of LIPs among distant homologues could help predict variations in dynamics, local stability and folding mechanism.

3.4 Conclusions

Here we have developed a straightforward and computationally inexpensive methodological formalism based on the physicochemical and geometric properties of protein buried interfaces that apparently captures the underlying structural bases of protein local instability. With this method, we have analyzed the 3D-structures of an ample group of proteins from different functional and structural categories, and we have been able to find LIPs in all cases, an indicative that they may be characteristic of protein folds. We have also compared our predictions of locally unstable regions with data from ours and other groups obtained using diverse experimental techniques, and in all cases our method was able to correctly map the regions described as conformational unstable experimentally. This comparison encompassed different kinds of equilibrium and kinetic intermediates, as well as molten globules and the transition state ensemble, which correspond to almost all the possible folding intermediates described so far. LIPs characteristics also appear to be conserved in protein families despite the sequence variation observed for the amino acids presented at the interfaces, which might be related with a mechanism for proteins evolving new functionalities by mutational tailoring, while conserving the essential characteristics of the buried interfaces, that determine the intrinsic dynamics of protein *loci*. The extension of this kind of studies for a wide group of proteins, or even at a genomic scale, could be of great help to study the variations in dynamics, local stability and folding mechanisms of close and distant homologous proteins, and also to obtain a better insight into the evolutionary constraints of proteins and the way protein intrinsic dynamics are encoded.

3.5 Methodology

3.5.1 Estimation of Buried Interfaces' Surface Polarity

We have developed a set of *ad hoc* Perl and Tcl scripts to estimate, from the 3D structure of a given protein, the ratio of polar/apolar surface area of its buried interfaces, for a piece of the code of the main script please see Appendix Script C.2. The input coordinates are used to extract a fragment of variable length –eight residues was the window size used– then the cropped protein and the extracted fragment are processed using NACCESS[112] –with a Probe Size = 1.40– to estimate the surface area of the atoms buried by interaction of the two parts. The polar or apolar character of each atom type is set by the NACCESS library, and it is attributed to its surface. Using this information we defined the polarity ratio (Pr) as follows:

$$Pr_{interface} = \sum_{i=1}^{m} SASA_{(polar)_i} \Bigg/ \sum_{j=1}^{n} SASA_{(hydrophobic)_j} \tag{3.1}$$

in which the total area of polar atoms as defined by NACCESS is divided by the total area of apolar atoms at the interface. This procedure is repeated by means of a sliding-window approach that permits the generation of a $Pr_{interface}$ profile of all interfaces along the structure of a protein. In these profiles, the Pr of each interface appears assigned to the fourth residue of the 8-residue probe. 7-residue or 9-residue probes give rise to close to identical profiles.

3.5.2 Buried Interface Packing Density

We have computed the packing density of buried interfaces $(\rho_{interface})$ using the following expression:

$$\rho_{interface} = \sum_{i=0}^{N} V_i^{\circ} \Bigg/ \sum_{i=0}^{N} V_i \tag{3.2}$$

in which the numerator corresponds to Voronoi standard atomic volumes and the denominator to the real Voronoi atomic volumes of the atoms found at the interface. The standard volumes were derived in a recent report from an extensive study of the intramolecular contacts made by atomic groups in small molecule crystals[83]. The actual Voronoi volumes of the atoms at the interface were calculated using the program CALC-VOL[113]. Packing density values close to 1 correspond to tightly packed interfaces. The iteration of this calculation along the structure of a protein generates the packing density profiles presented in this study. In these profiles, the packing density of each interface appears assigned to the fourth residue of the 8-residue probe. As in the case of the polarity profiles, there are no significant differences in the packing density profiles obtained for 7- and 9-residue windows.

3.5.3 Structural Multiple Alignments

The structural alignment of members of different protein families were constructed with the Multiseq package of VMD[114], based on the STAMP algorithm[115]. We aligned all the structures available for the Flavodoxin, Cytochrome *c* and α-Lactalbumine and the resulting structural alignments were processed with JOY[116] and CLUSTAL[111] to include information concerning residue surface exposure (see Appendix Figure E.1). For the other protein families studied in this work the small number of members with solved structure precluded building an informative enough structural alignment.

3.5.4 COREX Local Stability Calculations

The structures of the five proteins analyzed in this work (Apoflavodoxin, Cytochrome *c*, α-Lactalbumine, Indole-3-Glycerol Phosphate Synthase and Barnase) corresponding to different representative folds were processed with the software COREX[71,105‡] to estimate the local stability of protein regions. For each protein, a structure ensemble was first generated, which is used by the program in subsequent calculations. Based in the ensemble of structures generated, entropy-weighting factors were determined and stability constants calculated. The results obtained are represented in Figure 3.8.

3.5.5 Assessing the Statistical Significance of the Profiles

In order to evaluate the significance of our results we performed a non-parametric test to compare the polarity ratio of buried interfaces adjacent to structurally unstable regions, and of our predicted LIPs with conformationally stable protein segments or with non LIP regions, respectively. For each polarity profile, we performed a one-sided Mann-Withney-Wilcoxon rank-sum test with a confidence interval of $\wp - value < 0.05$ to test the significance of obtaining higher buried interface polarities in unstable segments confirmed experimentally or in LIPs, when compared to the polarities of stable regions, see upper half of Table 3.2. In order to make a quantitative comparison of our methodology with COREX[71,105], we performed the same statistical test for the residue specific stability profiles obtained with this program in the same set of proteins analyzed in this

[‡] Available at: `http://best.bio.jhu.edu/BEST/index.php`

study, depicted in Figure 3.8. In this case we aim to test the significance of obtaining lower stabilities for experimentally determined unstable regions or for LIPs in comparison to stable protein regions, respectively, see bottom half of Table 3.2. All the statistical calculations were implemented using the R statistical package[117].

3.6 Bibliography

[1] Oliver F Lange et al. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**: 1471–5, 2008. (see p. 122)

[2] B K Muralidhara et al. Thermodynamic fidelity of the mammalian cytochrome P450 2B4 active site in binding substrates and inhibitors. *Journal of Molecular Biology*, **377**: 232–45, 2008. (see p. 122)

[3] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci*, **33**: 2–8, 2008. (see p. 122)

[4] Robbyn L Tuinstra et al. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, **105**: 5057–62, 2008. (see p. 122)

[5] Marina Mapelli et al. The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint. *Cell*, **131**: 730–43, 2007. (see p. 122)

[6] Christopher M Dobson. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol*, **15**: 3–16, 2004. (see p. 122)

[7] Ivet Bahar et al. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys*, **39**: 23–42, 2010. (see pp. 122, 123)

[8] Harold A Scheraga, Mey Khalili, and Adam Liwo. Protein-folding dynamics: overview of molecular simulation techniques. *Annu Rev Phys Chem*, **58**: 57–83, 2007. (see p. 122)

[9] Constance J Jeffery. Moonlighting proteins–an update. *Mol Biosyst*, **5**: 345–50, 2009. (see p. 122)

[10] Peter Tompa, Csilla Szász, and László Buday. Structural disorder throws new light on moonlighting. *Trends Biochem Sci*, **30**: 484–9, 2005. (see p. 122)

[11] Constance J Jeffery. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current opinion in structural biology*, **14**: 663–8, 2004. (see p. 122)

[12] Constance J Jeffery. Multifunctional proteins: examples of gene sharing. *Ann Med*, **35**: 28–35, 2003. (see p. 122)

[13] C J Jeffery. Moonlighting proteins. *Trends Biochem Sci*, **24**: 8–11, 1999. (see p. 122)

[14] Shelley D Copley. Moonlighting is mainstream: paradigm adjustment required. *Bioessays*, **34**: 578–88, 2012. (see p. 122)

[15] Daphne H E W Huberts and Ida J van der Klei. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta*, **1803**: 520–5, 2010. (see p. 122)

[16] Nobuhiko Tokuriki and Dan S Tawfik. Protein dynamism and evolvability. *Science*, **324**: 203–7, 2009. (see pp. 122, 140)

[17] Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*, **19**: 596–604, 2009. (see p. 122)

[18] Shimon Bershtein, Korina Goldin, and Dan S Tawfik. Intense neutral drifts yield robust and evolvable consensus proteins. *Journal of Molecular Biology*, **379**: 1029–44, 2008. (see p. 122)

[19] Nobuhiko Tokuriki et al. How protein stability and new functions trade off. *PLoS Comput Biol*, **4**: e1000002, 2008. (see p. 122)

[20] Misha Soskine and Dan S Tawfik. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*, **11**: 572–82, 2010. (see p. 122)

[21] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual review of biophysics*, **37**: 215–46, 2008. (see p. 122)

[22] Javier F Cáceres and Alberto R Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*, **18**: 186–93, 2002. (see p. 122)

[23] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nat Genet*, **30**: 13–9, 2002. (see p. 122)

[24] B R Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, **17**: 100–7, 2001. (see p. 122)

[25] Sudhakaran Prabakaran et al. Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip Rev Syst Biol Med*, **4**: 565–83, 2012. (see p. 122)

[26] George A Khoury, Richard C Baliban, and Christodoulos A Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*, **1**: 2011. (see p. 122)

[27] Chunaram Choudhary et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**: 834–40, 2009. (see p. 122)

[28] Damian F Brennan and David Barford. Eliminylation: a post-translational modification catalyzed by phosphothreonine lyases. *Trends Biochem Sci*, **34**: 108–14, 2009. (see p. 122)

[29] Leo C James and Dan S Tawfik. Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends Biochem Sci*, **28**: 361–8, 2003. (see p. 122)

[30] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, **450**: 964–72, 2007. (see p. 123)

[31] Yawen Bai. Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem Rev*, **106**: 1757–68, 2006. (see p. 123)

[32] Anthony Mittermaier and Lewis E Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, **312**: 224–8, 2006. (see p. 123)

[33] Joanna F Swain and Lila M Gierasch. The changing landscape of protein allostery. *Curr Opin Struct Biol*, **16**: 102–8, 2006. (see p. 123)

[34] Ryo Kitahara, Shigeyuki Yokoyama, and Kazuyuki Akasaka. NMR snapshots of a fluctuating protein structure: ubiquitin at 30 bar-3 kbar. *Journal of Molecular Biology*, **347**: 277–85, 2005. (see p. 123)

[35] Joan J Englander et al. Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proc Natl Acad Sci USA*, **100**: 7057–62, 2003. (see p. 123)

[36] Haripada Maity et al. Protein hydrogen exchange mechanism: local fluctuations. *Protein Sci*, **12**: 153–60, 2003. (see p. 123)

[37] S W Englander. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu Rev Biophys Biomol Struct*, **29**: 213–38, 2000. (see p. 123)

[38] R Ishima and D A Torchia. Protein dynamics from NMR. *Nat Struct Biol*, **7**: 740–3, 2000. (see p. 123)

[39] Andrew Aquila et al. Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Opt Express*, **20**: 2706–16, 2012. (see p. 123)

[40] Marius Schmidt et al. Five-dimensional crystallography. *Acta Crystallogr, A, Found Crystallogr*, **66**: 198–206, 2010. (see p. 123)

[41] Lin X Chen. Probing transient molecular structures in photochemical processes using laser-initiated time-resolved X-ray absorption spectroscopy. *Annu Rev Phys Chem*, **56**: 221–54, 2005. (see p. 123)

[42] Fritz G Parak. Proteins in action: the physics of structural fluctuations and conformational changes. *Curr Opin Struct Biol*, **13**: 552–7, 2003. (see p. 123)

[43] V Srajer et al. Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochemistry*, **40**: 13802–15, 2001. (see p. 123)

[44] U K Genick et al. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science*, **275**: 1471–5, 1997. (see p. 123)

[45] V Srajer et al. Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. *Science*, **274**: 1726–9, 1996. (see p. 123)

[46] Veronique Receveur-Brechot and Dominique Durand. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci*, **13**: 55–75, 2012. (see p. 123)

[47] Pau Bernadó and Dmitri I Svergun. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst*, **8**: 151–67, 2012. (see p. 123)

[48] Alexander V Shkumatov et al. Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering. *Proteins*, **79**: 2122–31, 2011. (see p. 123)

[49] Peter Tompa. Unstructural biology coming of age. *Curr Opin Struct Biol*, **21**: 419–25, 2011. (see p. 123)

[50] Pau Bernadó et al. Structure and Dynamics of Ribosomal Protein L12: An Ensemble Model Based on SAXS and NMR Relaxation. *Biophys J*, **98**: 2374–82, 2010. (see p. 123)

[51] Lee Makowski. Characterization of proteins with wide-angle X-ray solution scattering (WAXS). *J Struct Funct Genomics*, **11**: 9–19, 2010. (see p. 123)

[52] Marco Cammarata et al. Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat Methods*, **5**: 881–6, 2008. (see p. 123)

[53] R F Fischetti et al. Wide-angle X-ray solution scattering as a probe of ligand-induced conformational changes in proteins. *Chem Biol*, **11**: 1431–43, 2004. (see p. 123)

[54] Luis A Campos et al. Structure of stable protein folding intermediates by equilibrium phi-analysis: the apoflavodoxin thermal intermediate. *Journal of Molecular Biology*, **344**: 239–55, 2004. (see pp. 123, 124, 129)

[55] A R Fersht, A Matouschek, and L Serrano. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *Journal of Molecular Biology*, **224**: 771–82, 1992. (see pp. 123, 135, 136)

[56] Bruno Rizzuti and Valerie Daggett. Using simulations to provide the framework for experimental protein folding studies. *Arch Biochem Biophys*, **531**: 128–35, 2013. (see p. 123)

[57] Ron O Dror et al. Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, **41**: 429–52, 2012. (see p. 123)

[58] R Bryn Fenwick, Santi Esteban-Martín, and Xavier Salvatella. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J*, **40**: 1339–55, 2011. (see p. 123)

[59] Sara Ayuso-Tejedor et al. Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. *Journal of Molecular Biology*, **406**: 604–19, 2011. (see pp. 123, 124, 129)

[60] Christopher D Snow et al. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, **420**: 102–6, 2002. (see p. 123)

[61] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, **9**: 646–52, 2002. (see p. 123)

[62] H J Berendsen and S Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, **10**: 165–9, 2000. (see p. 123)

[63] T Lazaridis and M Karplus. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**: 1928–31, 1997. (see p. 123)

[64] Bosco K Ho, David Perahia, and Ashley M Buckle. Hybrid approaches to molecular simulation. *Current opinion in structural biology*, **22**: 386–93, 2012. (see p. 123)

[65] Olivier Fisette et al. Synergistic applications of MD and NMR for the study of biological systems. *J Biomed Biotechnol*, **2012**: 254208, 2012. (see p. 123)

[66] Nicoletta Calosci et al. Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc Natl Acad Sci USA*, **105**: 19241–6, 2008. (see p. 123)

[67] Christopher J Francis et al. Characterization of the residual structure in the unfolded state of the Delta131Delta fragment of staphylococcal nuclease. *Proteins*, **65**: 145–52, 2006. (see p. 123)

[68] Emanuele Paci et al. Comparison of the transition state ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with phi value restraints. *Proteins*, **54**: 513–25, 2004. (see p. 123)

[69] Michele Vendruscolo et al. Structures and relative free energies of partially folded states of proteins. *Proc Natl Acad Sci USA*, **100**: 14817–21, 2003. (see p. 123)

[70] Vincent J Hilser et al. A statistical thermodynamic model of the protein ensemble. *Chem Rev*, **106**: 1545–58, 2006. (see pp. 123, 140)

[71] V J Hilser and E Freire. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *Journal of Molecular Biology*, **262**: 756–72, 1996. (see pp. 123, 136, 137, 140, 146)

[72] Tong Liu et al. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *J Am Soc Mass Spectrom*, **23**: 43–56, 2012. (see pp. 123, 141)

[73] Andrew D Robertson and Kenneth P Murphy. Protein Structure and the Energetics of Protein Stability. *Chem Rev*, **97**: 1251–1268, 1997. (see p. 123)

[74] Ivet Bahar, Chakra Chennubhotla, and Dror Tobi. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr Opin Struct Biol*, **17**: 633–40, 2007. (see p. 123)

[75] Ivet Bahar and A J Rader. Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol*, **15**: 586–92, 2005. (see p. 123)

[76] Paul Maragakis and Martin Karplus. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *Journal of Molecular Biology*, **352**: 807–22, 2005. (see p. 123)

[77] O Miyashita, J N Onuchic, and P G Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA*, **100**: 12570–5, 2003. (see p. 123)

[78] Aqeel Ahmed, Saskia Villinger, and Holger Gohlke. Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins*, **78**: 3341–52, 2010. (see p. 123)

[79] Ivet Bahar et al. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem Rev*, **110**: 1463–97, 2010. (see p. 123)

[80] Lei Yang, Guang Song, and Robert L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys J*, **93**: 920–9, 2007. (see p. 123)

[81] Manuel Rueda, Pablo Chacón, and Modesto Orozco. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**: 565–75, 2007. (see p. 123)

[82] David Schell et al. Hydrogen bonding increases packing density in the protein interior. *Proteins*, **63**: 278–82, 2006. (see p. 123)

[83] J Tsai et al. The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology*, **290**: 253–66, 1999. (see pp. 123, 145)

[84] Sara Ayuso-Tejedor et al. Design and structure of an equilibrium protein folding intermediate: a hint into dynamical regions of proteins. *Journal of Molecular Biology*, **400**: 922–34, 2010. (see pp. 124, 127, 129, 130, 132)

[85] Daniel Seeliger and Bert L de Groot. Atomic contacts in protein structures. A detailed analysis of atomic radii, packing, and overlaps. *Proteins*, **68**: 595–601, 2007. (see p. 126)

[86] Jorge Estrada et al. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. *BMC Bioinformatics*, **10**: 104, 2009. (see p. 128)

[87] Pau Bernadó, Martin Blackledge, and Javier Sancho. Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. *Biophys J*, **91**: 4536–43, 2006. (see pp. 128, 129)

[88] Eran Eyal et al. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J Comput Chem*, **25**: 712–24, 2004. (see pp. 128, 129)

[89] G M Langdon et al. Anabaena apoflavodoxin hydrogen exchange: on the stable exchange core of the alpha/beta(21345) flavodoxin-like family. *Proteins*, **43**: 476–88, 2001. (see p. 129)

[90] Kristian Rother et al. Voronoia: analyzing packing in protein structures. *Nucleic Acids Research*, **37**: D393–5, 2009. (see p. 130)

[91] Shrihari Sonavane and Pinak Chakrabarti. Cavities and atomic packing in protein structures and interfaces. *PLoS Comput Biol*, **4**: e1000188, 2008. (see p. 130)

[92] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, **338**: 1042–6, 2012. (see p. 131)

[93] Ken A Dill et al. The protein folding problem. *Annual review of biophysics*, **37**: 289–316, 2008. (see p. 131)

[94] Mallela M G Krishna et al. Branching in the sequential folding pathway of cytochrome c. *Protein Sci*, **16**: 1946–56, 2007. (see pp. 131, 132)

[95] Haripada Maity et al. Protein folding: the stepwise assembly of foldon units. *Proc Natl Acad Sci USA*, **102**: 4741–6, 2005. (see p. 131)

[96] Y Bai et al. Protein folding intermediates: native-state hydrogen exchange. *Science*, **269**: 192–7, 1995. (see p. 131)

[97] Jon López-Llano et al. The long and short flavodoxins: I. The role of the differentiating loop in apoflavodoxin structure and FMN binding. *J Biol Chem*, **279**: 47177–83, 2004. (see p. 132)

[98] B A Schulman et al. A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nat Struct Biol*, **4**: 630–4, 1997. (see pp. 132, 133)

[99] M Ohgushi and A Wada. 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett*, **164**: 21–4, 1983. (see p. 132)

[100] K Kuwajima. The molten globule state of alpha-lactalbumin. *FASEB J*, **10**: 102–9, 1996. (see p. 132)

[101] Zhenyu Gu et al. Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Gō model simulation. *Journal of Molecular Biology*, **374**: 528–46, 2007. (see pp. 133–135)

[102] Zhenyu Gu, Jill A Zitzewitz, and C Robert Matthews. Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: the roles of motif and sequence for the indole-3-glycerol phosphate synthase from Sulfolobus solfataricus. *Journal of Molecular Biology*, **368**: 582–94, 2007. (see pp. 133, 134)

[103] William R Forsyth and C Robert Matthews. Folding mechanism of indole-3-glycerol phosphate synthase from Sulfolobus solfataricus: a test of the conservation of folding mechanisms hypothesis in (beta(alpha))(8) barrels. *Journal of Molecular Biology*, **320**: 1119–33, 2002. (see p. 134)

[104] Xavier Salvatella et al. Determination of the folding transition states of barnase by using PhiI-value-restrained simulations validated by double mutant PhiIJ-values. *Proc Natl Acad Sci USA*, **102**: 12389–94, 2005. (see pp. 135, 136)

[105] Jason Vertrees et al. COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures. *Bioinformatics*, **21**: 3318–9, 2005. (see pp. 136, 137, 146)

[106] Nitin Bhardwaj and Mark Gerstein. Relating protein conformational changes to packing efficiency and disorder. *Protein Sci*, **18**: 1230–40, 2009. (see p. 140)

[107] Ernest Freire. Thermodynamics of protein folding and molecular recognition. *Pure and Applied Chemistry*, **69**: 2253–2261, 1997. (see p. 140)

[108] Marta Martínez-Júlvez et al. Structure of RdxA–an oxygen-insensitive nitroreductase essential for metronidazole activation in Helicobacter pylori. *FEBS J*, **279**: 4306–17, 2012. (see p. 141)

[109] G N Parkinson, J V Skelly, and S Neidle. Crystal structure of FMN-dependent nitroreductase from Escherichia coli B: a prodrug-activating enzyme. *J Med Chem*, **43**: 3624–31, 2000. (see p. 142)

[110] A L Lovering et al. The structure of Escherichia coli nitroreductase complexed with nicotinic acid: three crystal forms at 1.7 A, 1.8 A and 2.4 A resolution. *Journal of Molecular Biology*, **309**: 203–13, 2001. (see pp. 142, 143)

[111] M A Larkin et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**: 2947–8, 2007. (see pp. 143, 146)

[112] S Hubbard and J Thornton. NACCESS Computer Program Department of Biochemistry and Molecular Biology. University College London, London, UK. *NACCESS Computer Program Department of Biochemistry and Molecular Biology. University College London, London, UK*, 1993. (see p. 145)

[113] N R Voss and M Gerstein. Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. *Journal of Molecular Biology*, **346**: 477–92, 2005. (see p. 145)

[114] W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *J Mol Graph*, **14**: 33–8, 27–8, 1996. (see p. 146)

[115] R B Russell and G J Barton. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**: 309–23, 1992. (see p. 146)

[116] K Mizuguchi et al. JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**: 617–23, 1998. (see p. 146)

[117] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0 R Foundation for Statistical Computing, Vienna, Austria, 2012. (see p. 147)

CHAPTER 4

Linking SNPs with Abnormal Phenotypes in Familial Hypercholesterolemia using Molecular Dynamics

Contents

THE RESULTS PRESENTED IN THIS CHAPTER ARE A COMPENDIUM OF THE FOLLOWING RESEARCH ARTICLES:

- Angarica, V.E., Orozco, M. and Sancho, J. (2013). **Exploring the complete mutational space of the LDLr LA5 domain using molecular dynamics: linking SNPs with abnormal phenotypes in Familial Hypercholesterolemia**. to be submitted

4.1 Introduction

Familial Hypercholesterolemia (FH) is a genetic health disorder associated to abnormally high levels of cholesterol-carrying LDL in the blood stream, which can cause a varied set of pathologic manifestations related to the accumulation under the skin –*i.e.* xanthelasmas, in tendinous ends –*i.e.* xanthomas, or in the cornea[1–4], of this excess of non-metabolized cholesterol. Probably, the most serious pathologic manifestation of this disease is the accumulation of cholesterol in the blood vessels, with the formation of cholesterol artery plaques –*i.e.* in a process termed atherosclerosis– that constitutes a significant risk factor for cardiovascular and cerebrovascular diseases[5–9]. Atherosclerosis is also often associated to cognitive impairments and central nervous system disorders[10–12] and in many cases to early death caused by strokes or myocardial infarcts[5,6,9], which combined are by far the most common reason of death by unnatural causes in the last decade, as reported by the World Health Organization[*]. This disease follows an autosomal dominant inheritance pattern, and is one of the most common genetically inherited diseases in the human population, with well-accepted estimates[13–15] reporting a prevalence in the most severe stage of the disease manifested in homozygosis of $1 : 10^6$ and as high as $1 : 500$ for heterozygous individuals displaying the more moderated forms of FH, though in some specific populations[16–22] the prevalence might be even higher due to the founder effect[15,23]. FH is a complex disease with incomplete penetrance[24] and caused by the defects in a diverse group of proteins linked to cholesterol internalization and metabolism in cells –*e.g.* ***Apo B-100***[25–27], ***PCSK9***[28–31], and the LDL receptor (LDL-r)[32,33]. The gravity of the phenotype is related to the specific defective protein –*e.g.* the penetrance of mutations in ***Apo B-100***[34,35] is considerably lower than that of alleles of ***PCSK9***[31] and LDL-r[36] (with penetrance higher than 90% in both cases)– and the phenotypes observed in persons with a defective ***Apo B-100*** gene are less severe than those caused by impairments in the ***PCSK9*** or LDL-r genes. The great majority of FH cases and the most severe are associated to mutations in the LDL-r[34], which is in accordance the most studied protein among the three listed above to underscore the

[*] WHO Fact Sheet for the Top Ten leading causes of death worldwide between 2001 and 2011: http://who.int/mediacentre/factsheets/fs310/en/

molecular causes of this illness. However, much still needs to be done to understand the relationship between genetic variations and the phenotype in FH, as the number of mutations directly related to the disease is scarce when compared to the total possible mutations expected for these proteins. Besides, little is known on the specific molecular mechanisms linking genetic variations and phenotype, in the sense of how mutations affect the structure and/or function of the proteins related to FH. On top of that, it has been reported that there exists a 'diagnostic gap' because many clinically diagnosed FH patients fail to show any mutation in the above mentioned genes in FH cohorts screened for mutations[34], which might imply that other proteins from the cholesterol metabolism pathway different from those studied so far, are also implicated in the development of the disease. Indeed, in spite of the high prevalence of this disease worldwide and the serious health problems related to these genetic defects, recent reports point to the fact that in most cases FH is under-diagnosed and under-treated[37,38].

The LDL receptor gene family is an ancient family of membrane receptors whose members play important roles in multicellular organisms and that appeared very early with the onset of the first metazoans. There are evidences of the existence of members of the LDL-r family in primitive nematodes such as *C. elegans*[39,40] where they are essential proteins, in *D. melanogaster* where there are various members of this family coded in its genome[41,42] and they are also very distributed in higher eukaryotes. All the integrants of this family –*e.g.* which are organized in the following subfamilies: LDL-r, VLDL-r, ApoER2, LRP1, LRP2 and LRP6[43]– have different sequence and structural composition, and also diverse domain stoichiometry, but are assembled from the same set of structural constituents, including **a**) a cytoplasmic region that embodies $NPxY$ and $PPPSP$ motifs, **b**) a single transmembrane segment anchoring the cytoplasmic and extracellular sections to the cell membrane, **c**) an extracellular region formed by an epidermal growth factor (EGF)-like domain composed by multiple EGF repeats and a β-propeller domain, and **d**) the ligand binding region, consisting on a variable number of small cysteine-rich domains (CRD)[43–46] –*i.e.* known as LDL-r type A domains (LA domains). Some members of the LDL-r, VLDL-r and ApoER2 subfamilies also contain additional O-linked sugar domains positioned at the beginning of the extracellular section, right close to the outer face cytoplasmic membrane. There is plenty of information regarding the function of the cytoplasmic domain that is subject of modification by regulatory proteins at conserved sites for protein kinases and adaptor proteins. This domain is also responsible for establishing direct or adaptor-mediated interactions with structural proteins for triggering the endocytosis of the receptor-ligand complex upon binding[47–53]. However, the specific functions of the domains in the extracellular section of the receptor and the structural details of the interactions with the multiple ligands that can bind the members of the LDL-r family, are quite less known[44,46,54]. Lipoprotein binding is the common function of most members of this family but in addition they are also implicated in a wide spectrum of important biological processes such as migration, calcium influx, transcytosis, pericellular proteolysis, signal transduction, antigen presentation and synaptic plasticity among others, through binding to a diverse set of partners such as proteases, protease inhibitors, signaling molecules, heat-shock proteins, vitamin carriers, toxins and antibiotics[44–46,54,55].

Specifically, the LDL receptor, the protein that gives the name to the family, is the founding member and structurally the simplest of all the integrants of the family. For obvious reasons, the ligand binding domain has attracted great interest for the study of the interaction with LDL and other ligands, and the mechanism of ligand release after endocytosis. These studies have contributed to the accumulation of structural data of this domain, and there are multiple reports of structures of different domains within the ligand binding region, in isolation, in tandem or of the complete extracellular region[56–61] and the low resolution structure of the LDL–LDL-r complex[62]. It has been previously described that domains LA1–7[63], and most importantly domains LA4–5[55,64], are the key to establishing a correct interaction with the complexes formed between ApoE and lipids in VLDL particles, and are also necessary for efficient LDL binding, through interactions with apolipoprotein B[65]. These LA domains are small domains –*i.e.* less than 40 amino acids– the tridimensional structure of which are stabilized by three disulfide bridges and a coordination of a calcium ion. Small disulfide-rich domains are very common in the protein universe and comprise a varied set of proteins related to diverse functions such as growth factors, toxins, enzyme inhibitors, and structural or ligand-binding domains within larger polypeptides[66,67]. Among small disulfide-rich domains, the LA repeats characteristic of the LDL-r and all members of this family, are one of the most common autonomously structured extracellular modules found in non-redundant protein sequence databases[68]. In these domains lacking extensive hydrophobic cores and with little secondary structure, the pattern of disulfide bridges formed and the binding of the coordinating atom are essential for folding and the stabilization of the structure[55,61,66,68–71]. The LA domains in the LDL-r binding region are connected through linkers that provide great flexibility to the region[59] which might be related to adopting the correct tridimensional arrangement to bind voluminous ligands such as LDL.

The human LDL-r is a transmembrane protein of 839 amino acids (approximately 160 kDa[72]) coded in the *locus* 19p13.2[†] in chromosome 19, composed of 18 exons that span a region of ≈ 44.5 $kBases$ that can be transcribed into approximately 14 splice variants, of which only 9 are translated to protein species of different lengths. In the gene, exon

[†] Ensemble Detailed Description of the LDL-r gene: **Genomic Coordinates** = `19:11,200,036 - 11,244,505`

1 codes for the signal peptide, exons 2–6 code for the ligand binding site LA domains 1–7, exons 7–14 code for the EGF precursor-like domain, exon 15 codes for the *O*-linked sugar domain, exons 16–17 code for the transmembrane segments and exon 18 codes for the cytoplasmic region. The LDL-r gene is highly polymorphic, and many different types of mutations have been found, comprising large rearrangements of coding and/or intron regions –*e.g.* insertions, deletions, duplications, inversions; substitutions leading to stop codons causing the transcription of incomplete templates (nonsense); substitutions –*e.g.* synonymous or non-synonymous– which in the later case generate single amino acid replacements (missense) and mutations in the regulatory regions or splicing sites. Most of these different types of mutations associated with FH has been annotated in the most extensive and up-to-date databases including genetic variations in this protein[33,73]. The exact number of mutations varies from one repository to other, but the total number of mutations known nowadays for the LDL-r is between 1741–1835 for current releases of the LDL receptor database[33] and the Professional version of the Human Gene Mutation Database (HGMD)[73] respectively. These figures constitute a two-fold increase of the approximately 800 mutations compiled in 2007[14]. Substitutions are by large the most frequent type of mutation (≈ 73% of all the mutations), among which missense mutations –*i.e.* including SNPs– are significantly more frequent than the rest (≈ 73% of all the substitutions)[33]. The distribution of mutations is not uniform throughout the coding regions in the LDL-r gene, with a fairly high accumulation of mutations in the exons coding for the ligand binding domain LA domains 1–7 (≈ 41% of all the mutations in the coding regions), with the highest value observed for exon 4, which concentrates ≈ 20% of all the mutations in the coding regions[33]. Significantly, exon 4 codes for the complete sequence of LA domains 3–5, which includes the most relevant domains for binding lipoproteins[55,64,65].

Since the recognition of the association of FH to genetic variations in the LDL-r and the discovery of the first mutations causing the disease[74–76], a lot of information has been gathered on different types of mutations in the LDL-r gene and deposited in sequence and genetic variations databases[33,73,77–79]. More recently, with the great development of high-scale DNA genotyping and sequencing methodologies[80–88] –*e.g.* usually the procedure followed is extraction of genomic DNA that is amplified using PCR and then processed with denaturing HPLC (dHPLC) or single-strand conformation polymorphism (SSCP), and finally coupled with automated sequencing or multiplex ligation-dependent probe amplification analysis (MLPA) to identify mutations– there has been an increase of cascade screening programs in partial populations or countries, and in high risk groups[83,85,87,89–95]. These programs are promising but have some instrumental problems related to the lack of knowledge of the fate of most mutations and how they affect the LDL-r protein, and thus standard procedures work by probing for sets of known mutations, and not finding at least one of them does not guarantee that the subject is FH negative[96]. Other important problem is the cost effectiveness tradeoff of these tests, as most of these methods are still very expensive to be successfully applied at large scales with affordable costs. In order to underscore the final causes of FH at the LDLr protein structural level which are originated by DNA mutations, *in vitro* and computational studies have been performed with different domains of the ligand binding region, the complete binding region or with mutant species of the LA domains[55–58,68–71,97–101]. This kind of studies are of great importance because they could provide key insights into the real gravity of mutations, inferred from the structural or stability impairments caused in the LDL-r, which would be invaluable for broadening the spectrum of mutations known to be associated to FH. This in turn would help to reformulate the cascade screening tests described above by restructuring the set of mutations to be searched for in patient's DNA samples, as well as to widening the set of mutations to be considered in the search.

Following preliminary results obtained in our lab[101] regarding the potentiality of using atomistic simulations to study the effects of mutations in the native structure of the LA5 LDL-r binding domain, we designed this study to assess the fate of all possible missense mutations in the structure of this domain. As commented in the preceding paragraph, for this protein only a limited set of mutations is known to be associated to FH, and in most cases there is a lack of knowledge regarding the exact structural or stability changes caused by genetic variations at the structural level. Besides, there are important experimental limitations to study how all the possible mutations affect the stability of the protein due to the high number of possible mutant variants. Thus, we started from the cDNA sequence coding for this independent-folding interaction domain, and generated all the possible mutants arising from non-synonymous SNPs (256 SNPs coding for 227 different mutants) to try to understand how mutations affect the conformational dynamics of the LA5 domain, which although operationally complex is attainable computationally. We used state-of-the-art Molecular Dynamics simulations to assess the distortions caused by single amino acid substitutions in the tridimensional structure of the LA5 domain along time. As a result of this study, we generated a large amount of MD trajectory data embodying the dynamical evolution of all the mutants. We applied combined Data Mining methodologies including PCA and clustering techniques and identified interesting singularities in the conformational behavior of different types of mutants, that could cause the destabilization of the LA5 domain, thus impairing recognition of LDL. We have obtained quantitative evidence for estimating the grade of structural distortions caused by mutations, and uncovered a new set of mutations that could cause a significant destabilization of LA5 domain. We hope that the data generated in this type of computational studies could help experimentalist to complement the study of conformational diseases in which mutations reduce the stability of the protein, by reducing the search space for putative pathogenic mutations. Besides, we also believe that the results and methodological propositions arising from this work could help in the study of the fate of mutations in the ligand binding of other membrane receptors bearing LA domains, and even in other unrelated protein domains, guiding in the development of new strategies to stabilize those molecules and to better understand conformational diseases from a structural perspective.

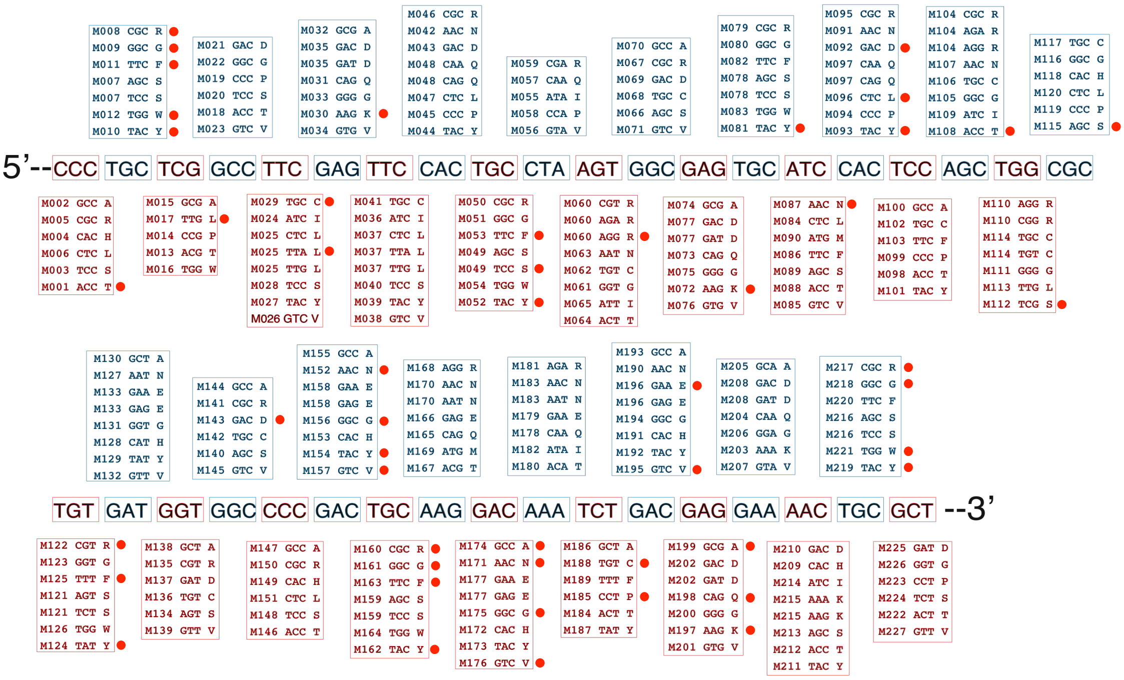
4.2 Results

4.2.1 The Complete SNP Mutational Space of the LA5 domain

Relating variations in the DNA sequence with phenotype is challenging, specially in polygenic diseases such as Familial Hypercholesterolemia. The identification of mutations that could be considered 'markers' of FH proceeds by the sequencing of the associated genes –*e.g.* ***Apo B-100***[25–27], ***PCSK9***[28–31], and the LDL receptor[32,33]– in diseased individuals to uncover the genetic variations that could explain the pathologic phenotype. However, the 'cause-effect' link established following this approach is indirect, because identifying a group of mutations in persons suffering this illness does not provide any clue regarding the affections at the post-transcriptional levels that are indeed determining the abnormal phenotype. Besides, as the cascade screening procedures used to identify FH genetic markers only probe for sets of known mutations in the associated genes[96], and because of the polygenic nature of FH in which not all the genes from the cholesterol metabolism route that are implicated have been identified, in many cases no known mutations have been found in people with FH[34,82,83]. Thus, the need for experimental or computational approaches that could shed light into the molecular mechanisms by which mutations in the DNA affect protein structure or function in genetic diseases.

There exists a many steps between the processes of gene transcription and translation, and also beyond the processes of protein folding, maturation and turnover in which mutations can cause a reduction of the quality or quantity of the protein available to exert its function, therefore leading to disease. Taking this into consideration, it is also known that a significant proportion of protein mutations affect protein stability and folding[102–105], an event that is very difficult to anticipate from the assessment of DNA sequence variations. There are different methods that try to tackle this problem based on different rationales and at different organizational levels –*e.g.* some work at the DNA level by mapping the location of SNPs for finding if the variation is found in a tentatively functionally important region[106], while others rely on mapping SNPs into the sequence[107–112] or the structure[104,113–118] of proteins, or combinations of structure and sequence information[119–121]. Of course, like most predictive methodologies, those mentioned have a more or less significant error rate, arising from the fact that their performance is dependent on the quality of the data used for training and the assumptions made in the predictive pipeline. Another possibility would be to try to assess the effect of mutations directly at the structural level without previous assumptions or prior evolutionary knowledge, for example by exploring the effects of mutations on the structure of a protein running molecular dynamic simulations[101,122,123].

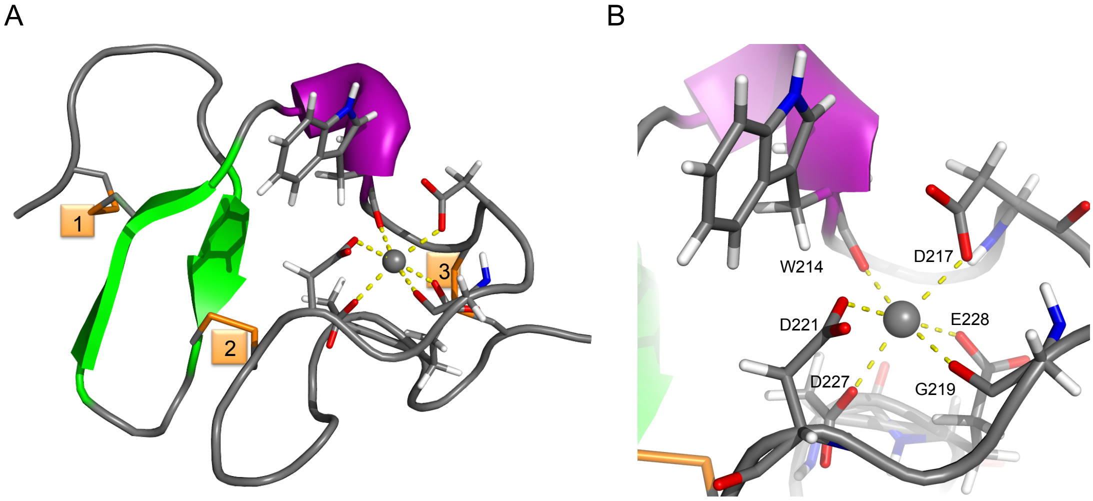
FIGURE 4.1: Complete SNP Mutational Space of the LA5 Domain



A summarized graphical representation of the cDNA coding for the LA5 domain and all the possible mutations arising from SNPs. In the chart, alternate codons are colored in red and blue, and for each codon we also include above or below a box with the same color than the corresponding wild-type codon, with the indexes used to identify the mutations, the mutated codons and the mutated amino acids. We tagged with a red dot all mutations that have been identified in persons with FH (as included in Appendix Table F.1). For a more detailed view of this data please see Appendix Table F.2

In the specific case of trying to relate genetic variations with phenotype in FH, due to its polygenic nature, there could be three possible outcomes for a given mutation found in a person affected by this disease: **a**) the mutation is related to FH by affecting the quality or quantity of the gene product, **b**) the mutation is not related to FH but there is(are) other unidentified mutation(s) –*e.g.* not found in the specific screening study or an unreported mutation– in any of the genes known to be related to the disease, or **c**) the mutation is not related to FH but there is(are) other mutation(s) in other genes from the cholesterol metabolism pathway not yet associated to FH that are causing the disease. Thus, the results of cascade screening studies and of any predictive method should be taken with a grain of salt, always pondering the possibility that the genetic variation could be a false positive. In our case, the approach is also a reductionism because we are studying only one domain of one of the proteins associated with FH –although specifically the protein that is related to the majority of cases of FH and the domain that is key to LDL-r function, coded in the exon that bears the highest number of genetic variations. However, we would able to directly appraise the extent of conformational instability caused by specific mutations, and also to examine all the possible mutations –*i.e.* reported and unknown LDL-r LA5 domain mutations– in the same experimental conditions. A summary of the initial dataset used in this study is shown in Figure 4.1, where we include the coding sequence for the LDL-r LA5 domain and all the possible SNPs, for a more detailed description please see Appendix Table F.2. From all the possible SNPs mutations –*i.e.* 256 non-synonymous SNPs coding for 227 different single amino acid substituted protein variants– only 22% have been found in persons with FH, red dot tagged mutants in Figure 4.1 and described in detail in Appendix Table F.1.

FIGURE 4.2: Tridimensional Structure of the LA5 Domain



The structure of the human LDL-r LA5 domain (PDB id: 1AJJ). **A)** The structure of the complete LA5 domain showing the different structural *loci*, including the calcium binding box and the three disulfide bridges highlighted with orange boxes with the corresponding index starting from the N-t end: **1)** between C**197**-C**209**, **2)** between C**216**-C**231** and **3)** between C**204**-C**222**. **B)** A close perspective of the calcium binding site showing the six residues forming the octahedral coordination box (with the notation: Residue Name-Residue Index: Atom Name): W**214**:O, D**217**:OD1, G**219**:O, D**221**:OD2, D**227**:OD2, E**228**:OE2

As can be seen in Figure 4.2, although LDL-r LA domains are short domains of less than 40 residues, they contain some important structural *loci* which are very important on determining their stabilization and correct folding[55,61,66,68–71]. The analysis of the predicted effect of mutations on some of the amino acids participating in the formation of these structural *loci* –*e.g.* the three disulfide bridges and the calcium binding site– can be extracted from Appendix Table F.1, for the assessment of the effect of mutations using four different methodologies like PMUT[109,115,124], and a consensus approach, CONDEL[119], integrating the predictions made using SIFT[108,125], Polyphen-2[126] and Mutation Assessor[127,128]. These results show that there exist discrepancies among the predictions of the deleteriousness of mutations according to different approaches due to the differences on the training of the methods and the statistical procedures followed by each to conclude in the predictions. For example, for some known mutations in cysteines forming the stabilizing disulfide bridges like C**197**{*176*}G, C**204**{*183*}S and C**231**{*210*}G. The same inconsistency among predictive methods is found for known mutations in amino acids from the calcium binding box, such as D**221**{*200*}N, D**221**{*200*}Y, D**221**{*200*}G, D**221**{*200*}V, D**227**{*206*}V, E**228**{*207*}K and E**228**{*207*}Q. And of course the same discrepancies are found for an ample number of other known (Appendix Table F.1) and unknown (Appendix Table F.2) mutations in residues all around the LA5 structure. Thus, depending on the specific predictive approach used, the conclusions draw would be completely disparate. The use of an approach such as CONDEL[119] is also interesting because a consensus approach like this permits to obtain better predictive results than the methodologies on which it is based can obtain independently. And here again, the results obtained for a significant number of mutations by SIFT[108,125], Polyphen-2[126] and Mutation Assessor[127,128], which are used by CONDEL[119] to conclude in its consensus deleteriousness score, show differences from model to model, so people using these methods independently would conclude on a different pattern of Neutral and Deleterious mutations in the LDL-r LA5 domain.

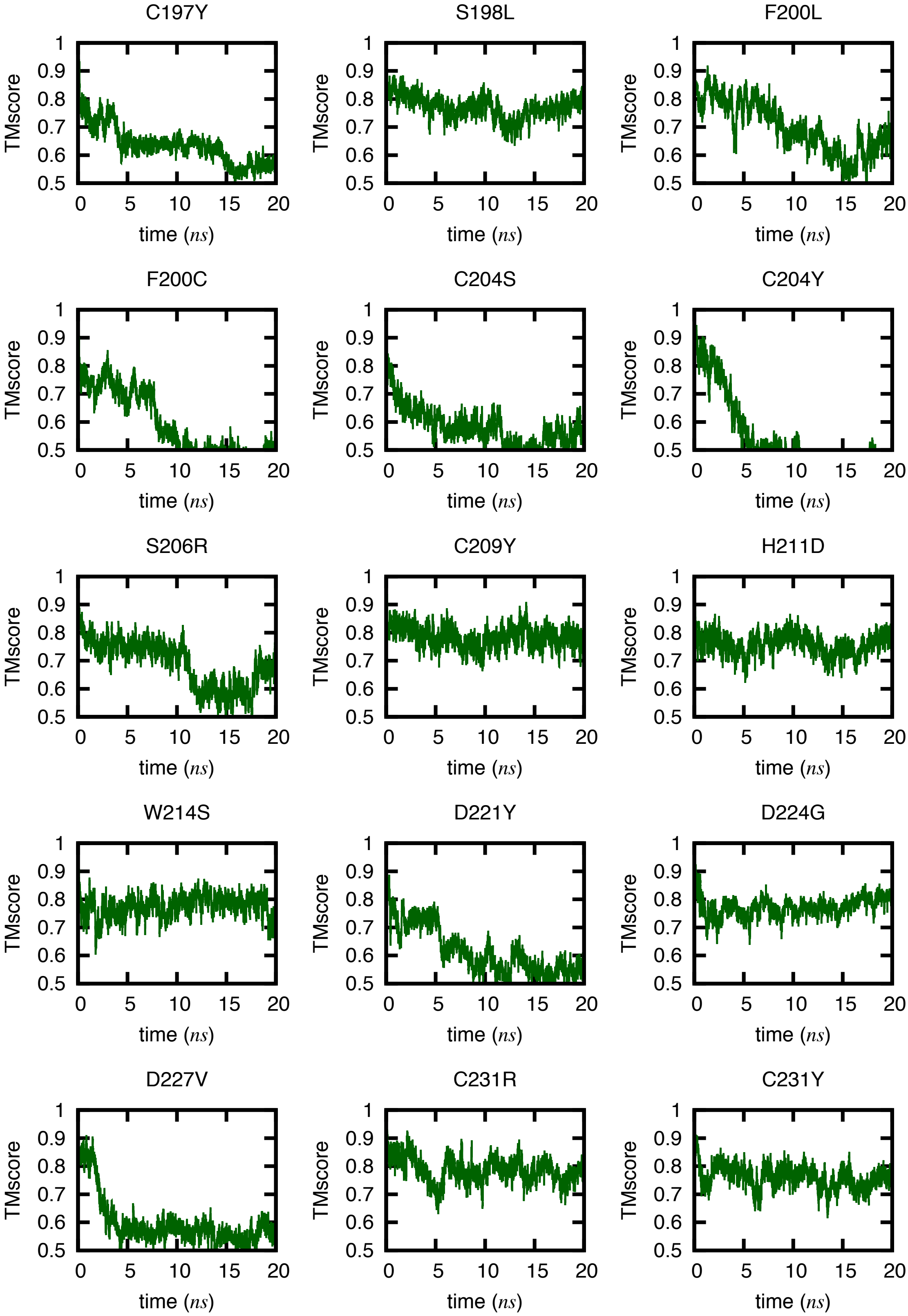
4.2.2 Conformational Instability of the LA5 Domain Mutant Variants

To study the relationship between mutations and the phenotype we followed a different approach by studying the direct effect of single amino acid substitutions in the structure of the LDL-r LA5 domain (Figure 4.2) using Molecular Dynamics. All the 227 mutants generated by SNPs in this domain (Figure 4.1 and Appendix Table F.2) were generated *in silico* at the structural level using SCWRL[129], which was also used to find the best conformations for the side chains of the substituted amino acids. Then, all mutants were minimized and equilibrated in explicit water and we run MD production trajectories totaling 6 μs, as described in the Methodology section. For each mutant we run 20 ns-long MD simulations to explore its conformational evolution, taking into consideration that in principle in this approximated timeframe it is possible to observe significant differences for unstable mutations[101]. These large amount of MD trajectory data can be followed using different structural metrics such as the Template Modelling score $(TM - score)$[130,131], which is a more accurate and a protein-length independent measure of the comparison among protein structures than more commonly used metrics such as the Root Mean Square Deviation (RMSD) or the Global Distance Test (GDT). Using this metric it is possible to follow multiple MD simulations to identify conformational instabilities, because in general the $TM - score$ values tend to fluctuate very close to 1 in equilibrium trajectories, which corresponds to the highest value obtained when comparing identical structures. On the other hand, values below 0.5 are obtained for proteins of different structural topologies or foldings[130], and this limit score can be used for identifying significant structural distortions in MD simulations in which the protein has been perturbed, as in our case, by mutating a single amino acid.

These results are shown in Figure 4.3 for a selection of trajectories out of the 227 total mutants, representing some examples of the conformational instability introduced by different mutations from those found in persons with FH. These results show some interesting facts regarding the effect of mutations at the structural level, as the extent of the conformational instability is clearly distinct for different known mutations. For example, the substitutions C**197**{*176*}Y, F**200**{*179*}L and C, C**204**{*183*}S and Y, S**206**{*185*}R, D**221**{*200*}Y and D**227**{*206*}V cause great conformational instability in this timescale, in almost all cases reaching to conformations containing significant structural distortions after a few nanoseconds of simulation. On the other hand, other known mutations such as S**198**{*177*}L, C**209**{*188*}Y, H**211**{*190*}D, W**214**{*193*}S, D**224**{*203*}G, C**231**{*210*}R and C**231**{*210*}Y cause mild or apparently none distortions to the structure of the LA5 domain during the simulations. Interestingly, in both groups there are mutations in residues involved in the three disulfide bridges or the calcium binding box, and as previously stated, not all substitutions in these structurally key residues tend to affect the global conformational stability of the domain.

We have assessed the degree of conformational distortion in all mutants as follows. First, we performed Principal Component Analysis (PCA) of all trajectories in order to both reduce the dimensionality of the large amount of conformational data in our MD simulations, and to allow their comparison in the same coordinate system. PCA is a statistical technique that has been customarily used to analyze MD data[132–136] to extract the dominant trends of motions in trajectories, and to identify the correlated movement of different protein regions, please see the Methodology section for details. A summary of the PCA in all the 227 trajectories corresponding to the different mutants is included in Appendix Figures G.1 and G.2. In the first of these two charts we include an overview

FIGURE 4.3: Protein $TM - score$ of some LDL-r LA5 Mutants along MD Simulations



The evolution of the $TM - score$ for the structure of some selected LDL-r LA5 mutant domains throughout the 20 ns-long MD simulations. In each case we include the description of the mutation as the subchart title. All these mutations have been found in persons with FH and are a subset of those detailed in Appendix Table F.1

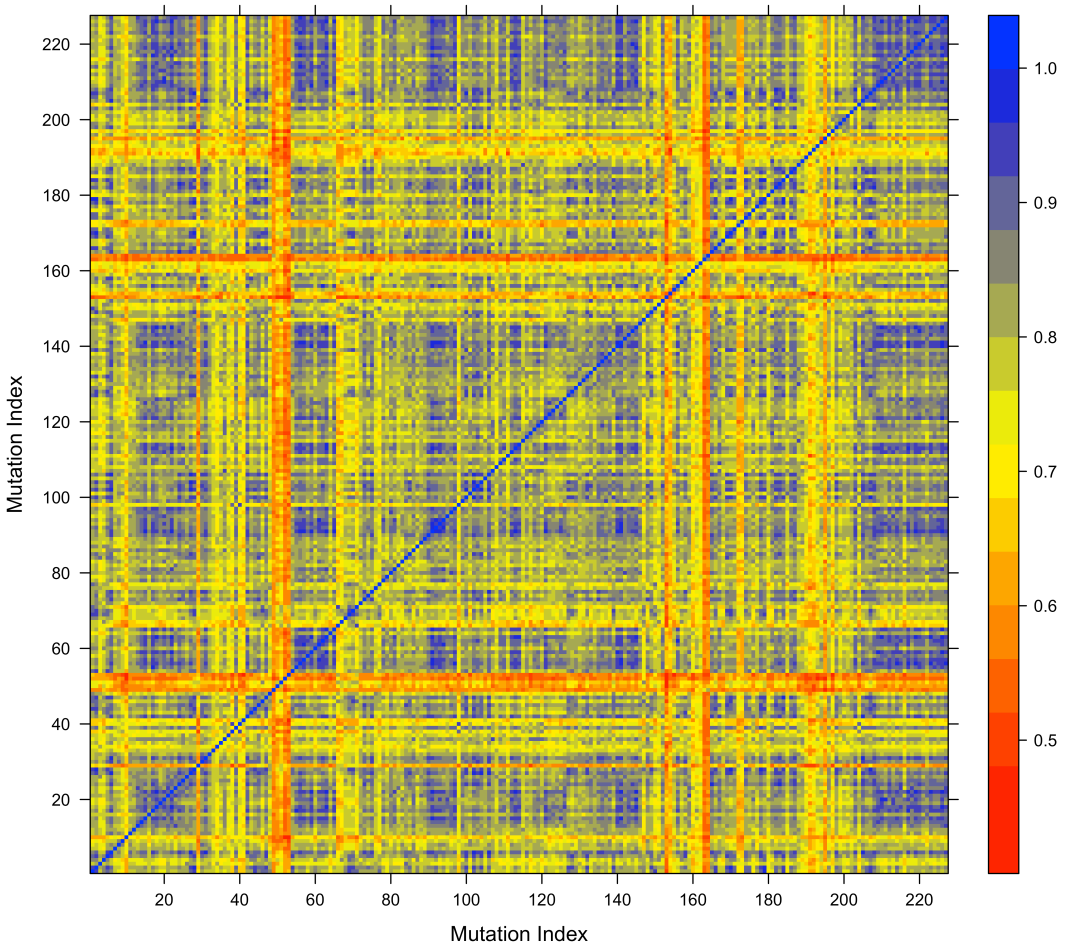
of the percentage of the variance in the original trajectories explained by the corresponding eigenvectors. As can be seen in the 'Scree Plot' in Appendix Figure G.1 for the 1^{st} to 30^{th} modes, in the MD simulations analyzed the first three modes describe on average 36%, 16% and 9% of the total variance. Thus, by just considering those three eigenvectors it would be possible to describe as much as 61% of the variance, with a considerable reduction of the degrees of freedom to be considered. In this chart it is also clear that below the fourth mode, the variance described by each eigenvector individually is always below 6%, which also justify the selection of the first three modes, as previously described[132,135,137,138], for graphical summarizing which has been termed as the 'essential dynamics' of the system[132,139–142]. In the second figure (Appendix Figure G.2) we describe a histogram of the number of modes needed for describing 95% of the total variance in the trajectories, and in most cases this number is around 19 eigenvectors.

After performing this analysis in all trajectories, we first compared the average structures, to get an idea of the structural differences among mutants and the conformational instabilities caused by the amino acid substitutions in the initial LDL-r LA5 domain. As can be seen in Figure 4.4, there are some groups of mutants that are structurally dissimilar to most other mutants, specifically trajectories M009 and M010 (C**197**{*176*}G and Y), trajectory M029 (F**200**{*179*}C), a cluster including trajectories M049–M053 (C**204**{*183*}S, R, G, Y), another cluster including trajectories M153–M155 (D**221**{*200*}H, Y, A), another cluster including trajectories M161–M164 (C**222**{*201*}G, Y, F, W) and another cluster including trajectories M191–M197 (D**227**{*206*}H, Y, A, G, V, E and E**228**{*207*}K). There exist other groups of mutants whose average structures are fairly similar to those obtained for other substitutions, like those for which the vertical and horizontal lines in the heatmap are colored in blue or in different tones of blue, please see Figure 4.4 and refer to the Appendix Table F.2 for the correspondence among trajectory indexes, codon change and amino acid substitutions.

To get an idea of the conformational evolution of mutants along the trajectories, we obtained the projections of each trajectory into the first three PCs. The results of this assay is depicted in Figure 4.5 and the associated videos[‡]. These charts depicts the projections of the conformations visited during a simulation into the space formed by the first three PCs generated from the PCA analysis of the trajectory, please see the Methodology section for a detailed description of the PCA methodology. The projections correspond to a statistical representation of the structures in each time step during a simulation, represented in the transformed space defined by the eigenvectors of the trajectory covariance matrix, taking into consideration the eigenvalues –*i.e.* the weights– of

[‡] While displaying the videos there are different quality options, thus for a better video experience it is better to select the 720p in the bottom right angle of the video panel, in the **Settings** button. In case you are reading the printed version of this document, please refer to Appendix Table F.3 to obtain the URL link addresses

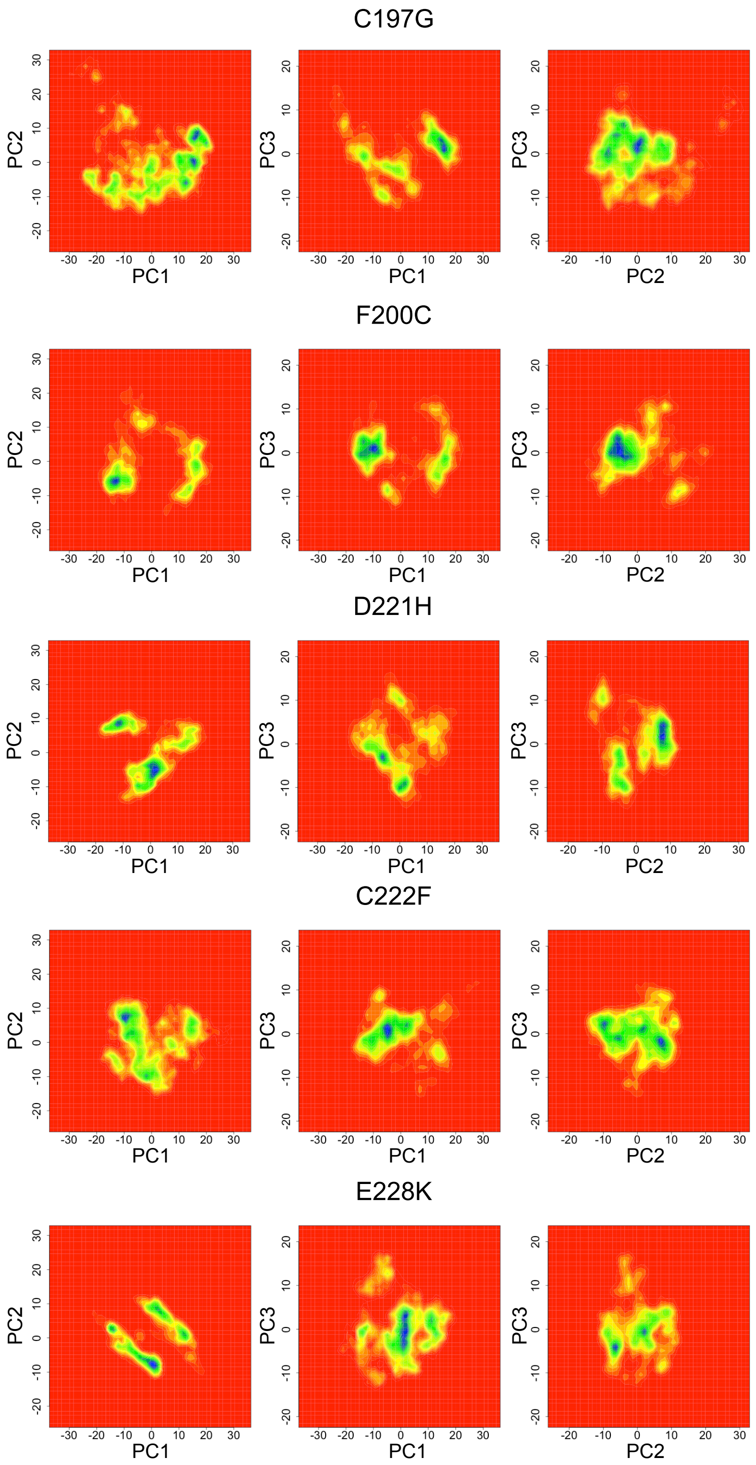
FIGURE 4.4: All-to-all Comparison of the Trajectories Average Structures



The average structures of all the 227 mutants were extracted from the trajectories after performing a PCA. The average structures were compared in pairs using the $TM - score$ metric. The chart is a heatmap of the comparison of all *vs* all, and the color in each cell corresponds to the $TM - score$ for the comparison of two structures whose indexes are found in the abscissa and ordinate axes. In the rightmost side of the chart we include the color legend for the $TM - score$, from red for dissimilar structures $(TM - score \approx 0.5)$ to blue for identical structures $(TM - score \approx 1)$, passing through various intermediate tones of orange and yellow

each eigenvector. The values of the projections correspond to a measure of the similarity of the structure at a given time step to the average structure of the simulation, located at the origin $(0_{1^{st}}, 0_{2^{nd}}, 0_{3^{rd}}, \ldots, 0_{N^{th}})$ formed by the N^{th} eigenvectors, and by association, a measure of the similarity of whichever two structures from the trajectory in the space formed by the same number of principal components. A Boltzmann-weighted ensemble of molecules that behave harmonically will show normal 'Gaussian' distributions of projections in each dimension, and accordingly, deviation from this expected behavior indicates that the ensemble of structures may not be correctly representing the predicted accessible space[132]. Thus, the expected graphical outcome when plotting

FIGURE 4.5: Dynamical Evolution of LA5 Mutants in the PCA Space (Destabilizing Mutations)


C197G

F200C

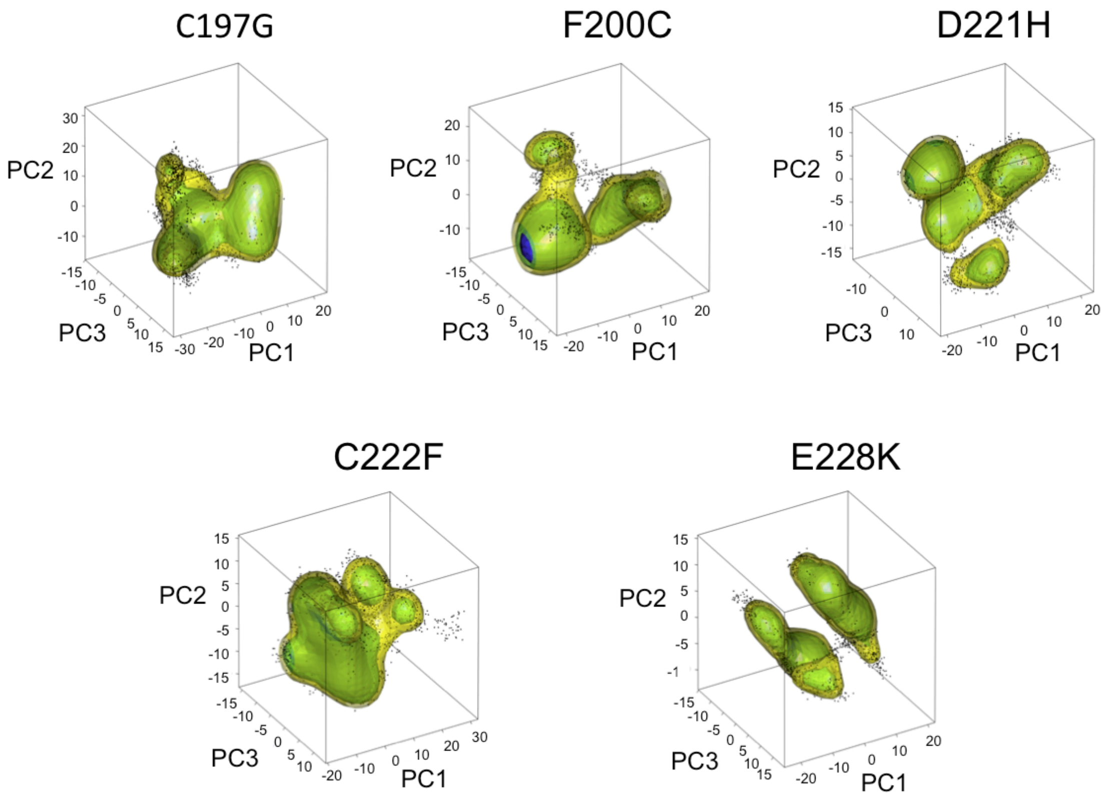
D221H

C222F

E228K

The MD trajectories are followed along time by projecting the structures at each time step into the space described by the first three PCs. Each subchart is a two-dimension density plot of the projections of the structures into PC1 *vs* PC2, PC1 *vs* PC3 and PC2 *vs* PC3. The color scale goes from red (no occupancy) to blue (high occupancy), passing through intermediate scales of yellow and green. For accessing the more descriptive animations please visit the following links for each example: `C197G`, `F200C`, `D221H`, `C222F`, `E228K` the projections of the conformations visited during a MD equilibrium simulation in the PC space, should be an ellipsoid around the coordinate axes origin, with the PCs corresponding to the axes of the plot and the dispersion of the ellipsoid length in the i^{th} PC being proportional to the i^{th} eigenvalue, please see the Methodology section for more details. Accordingly, a deviation from this expected behavior –*e.g.* due to perturbations caused by amino acid substitutions– could be taken as a qualitative and quantitative estimate of the conformational instability caused by the perturbation.

FIGURE 4.6: 3D Density Plots of LA5 Mutants in the PCA Space (Destabilizing Mutations)



We show the 3D density plots of the projections of the structural conformations visited throughout the MD simulations in the transformed space formed by the first three PCs in some destabilizing mutations. We include the same examples depicted on Figure 4.5 to provide a better view of the different states visited in the trajectories

In Figure 4.5 and the associated videos the significant conformational instability that some amino acid substitutions cause on the LDL-r LA5 structure is made clear, as in some cases the projections randomly visit different states corresponding to disparate structural arrangements, as in the cases of C**197**{*176*}G and F**200**{*179*}C. In other cases, there are different structural conformations which are significantly visited in the PCA space, such as for D**221**{*200*}H and E**228**{*207*}K. Probably a clearer picture can be obtained from the inspection of Figure 4.6 for the same examples in Figure 4.5, on which including in the same plot the three dimensions and removing the contribution of time, the different substates visited and the extent of the conformational instability are better appraised. The same analysis for other mutants, as those included in Figure 4.7 and the associated videos[§], and the composite three dimensional plot in Figure 4.8, shows the fairly stable behavior in other cases in which the amino acid substitutions have little or no effect on the conformation of the LDL-r LA5 domain, as can be inferred from the ellipsoidal shape of the projections plots in the first three PCs.

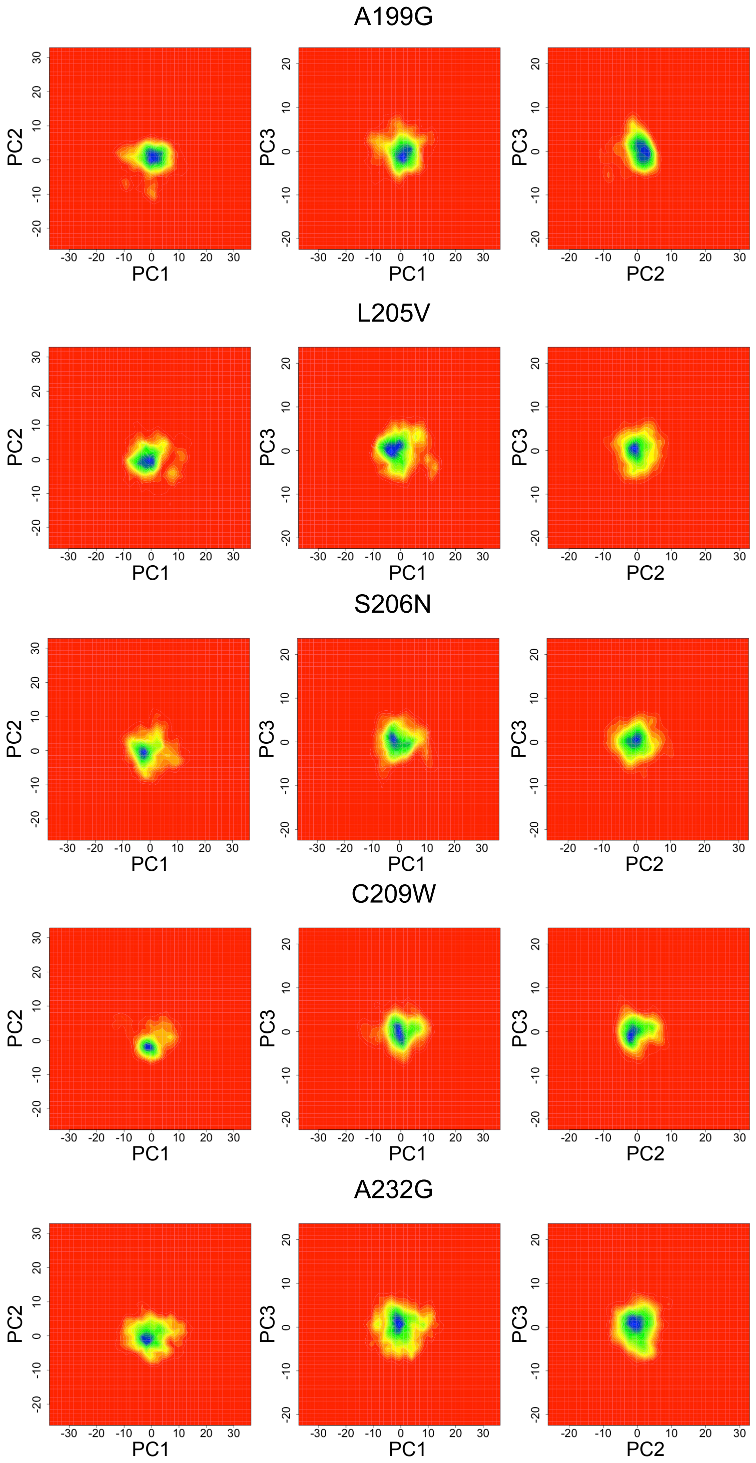
4.2.3 Local Instability of Mutants in Residues of the LA5 Domain Structural *loci*

Another important question is to assess how mutations in different positions along the structure of the LDL-r LA5 domain affect the conformational stability of some important structural *loci*, such as the calcium binding box, which could have important implications for the global stability of the domain[55,61,66,68–71]. An initial comparative assessment of the angular deformation of the octahedral structure of the coordination box is shown in Figure 4.9. In this figure we show the distributions and fluctuation equilibrium values for each of the fifteen angles formed among whichever two residues out of the six forming the coordination box and the calcium in the vertex, see Figure 4.2, panel B. In the case of the wild-type MD simulation (top left chart of Figure 4.9), the outcome observed is the expected for a simulation in equilibrium, in which the structure of the calcium binding box is fairly stable, as can be inferred from the fluctuations of all the angles fairly close to the expected equilibrium values –*i.e.* $180°$ and $90°$. A similar result is observed in the MD simulation of one of the mutants involving a cysteine forming a disulfide bridge (top right). In this case, the distributions for the fluctuation of each angle are also fairly close to the expected equilibrium values, although there is a clear increase in the dispersion of the distributions as indicated by the augment in the number of outliers in each distribution. Thus, apparently, in this case the perturbation caused by the disruption of the disulfide bridge is not causing great distortions in the conformation of the calcium binding box. In the two cases included in the bottom panels in Figure 4.9, the considerable distortions in the octahedral conformation of the calcium binding box are evident in the case of substitutions of coordination residues, as the equilibrium fluctuation values of most of the angles are rather far from the expected angle range.

Probably more interesting than analyzing case by case would be to make quantitative comparisons of the effect of different mutations in specific structural *loci*. In principle,

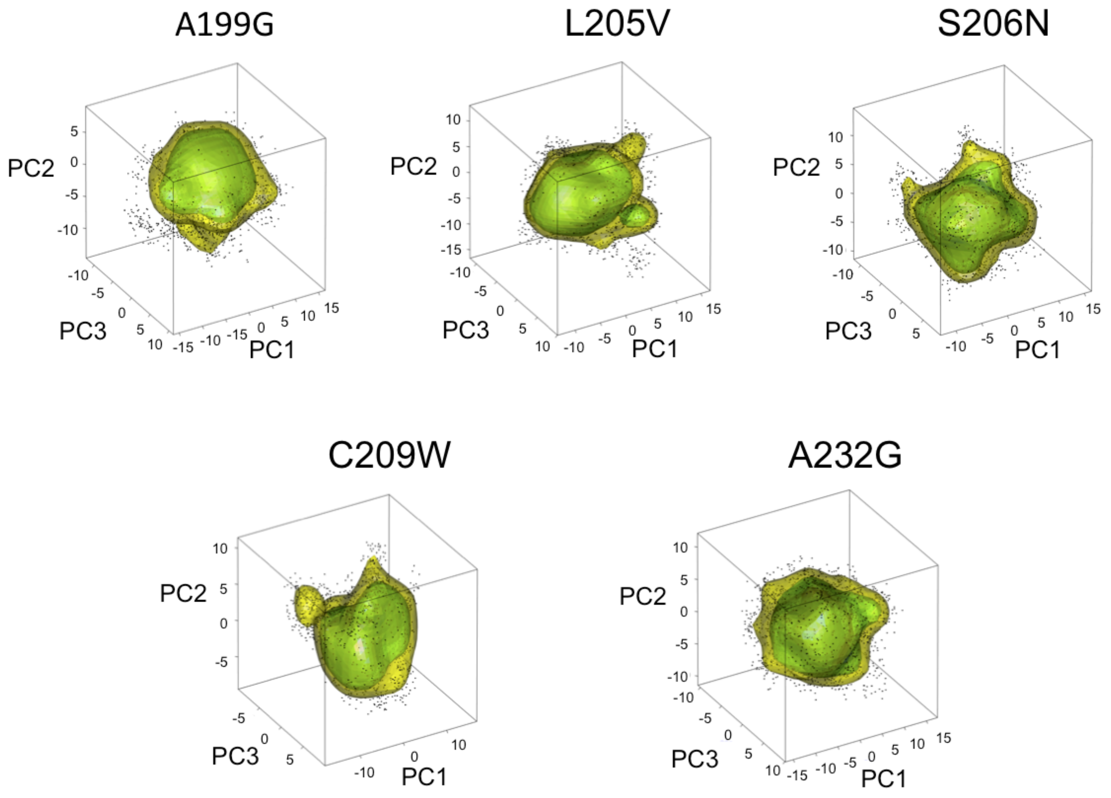
[§] While displaying the videos there are different quality options, thus for a better video experience it is better to select the 720p in the bottom right angle of the video panel, in the **Settings** button. In case you are reading the printed version of this document, please refer to Appendix Table F.3 to obtain the URL link addresses

FIGURE 4.7: Dynamical Evolution of LA5 Mutants in the PCA Space (Stable Mutations)



The MD trajectories are followed along time by projecting the structures at each time step into the space described by the first three PCs. Each subchart is a two-dimension density plot of the projections of the structures into PC1 *vs* PC2, PC1 *vs* PC3 and PC2 *vs* PC3. The color scale goes from red (no occupancy) to blue (high occupancy), passing through intermediate scales of yellow and green. For accessing the more descriptive animations please visit the following links for each example: A199G, L205V, S206N, C209W, A232G

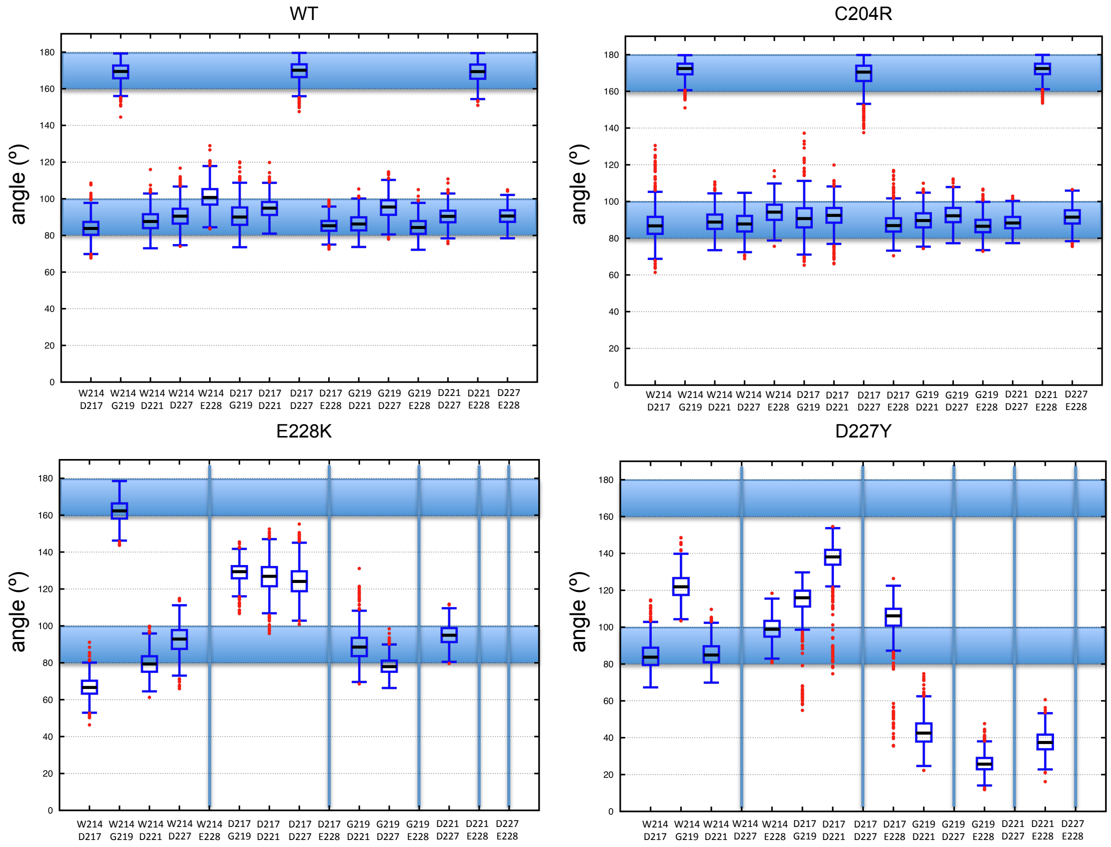
FIGURE 4.8: 3D Density Plots of LA5 Mutants in the PCA Space (Stable Mutations)



We show the 3D density plots of the projections of the structural conformations visited throughout the MD simulations in the transformed space formed by the first three PCs in some stable mutations. We include the same examples depicted on Figure 4.7 to provide a better view of the different states visited in the trajectories

PCA can also be of help to study the changes in the local stability of a specific protein region, but the problem is that in practice the presence of large-scale motions makes it difficult or impossible to resolve small-scale motions because the former have much greater relative amplitudes[143]. Other approaches are better suited and have been applied to examine local instability of some specific residues in enzyme binding sites and allosteric mechanisms[144–146], by estimating the changes in the statistical trends of the distributions of the dihedral angles of those residues during computational simulations. These approaches are based on following the changes in the fluctuations of the ϕ and ψ angles of a residue or residues in a perturbed simulation, and comparing them with those observed in a simulation in equilibrium, using a statistical metric such as the Kullback-Leibler[147] or the Jensen-Shannon[148–152] Divergence. We implemented a similar method to analyze the effects of all the possible mutations generated from SNPs in the LDL-r LA5 domain in the local stability of the six residues forming the calcium binding box. In this bootstrapping assay, for each of the 227 mutants MD simulations, we randomly resampled with replacement 10^5 sets, each containing 10^3 snapshots –*i.e.*

FIGURE 4.9: Angular Fluctuation of the Calcium Binding Box during MD Simulations

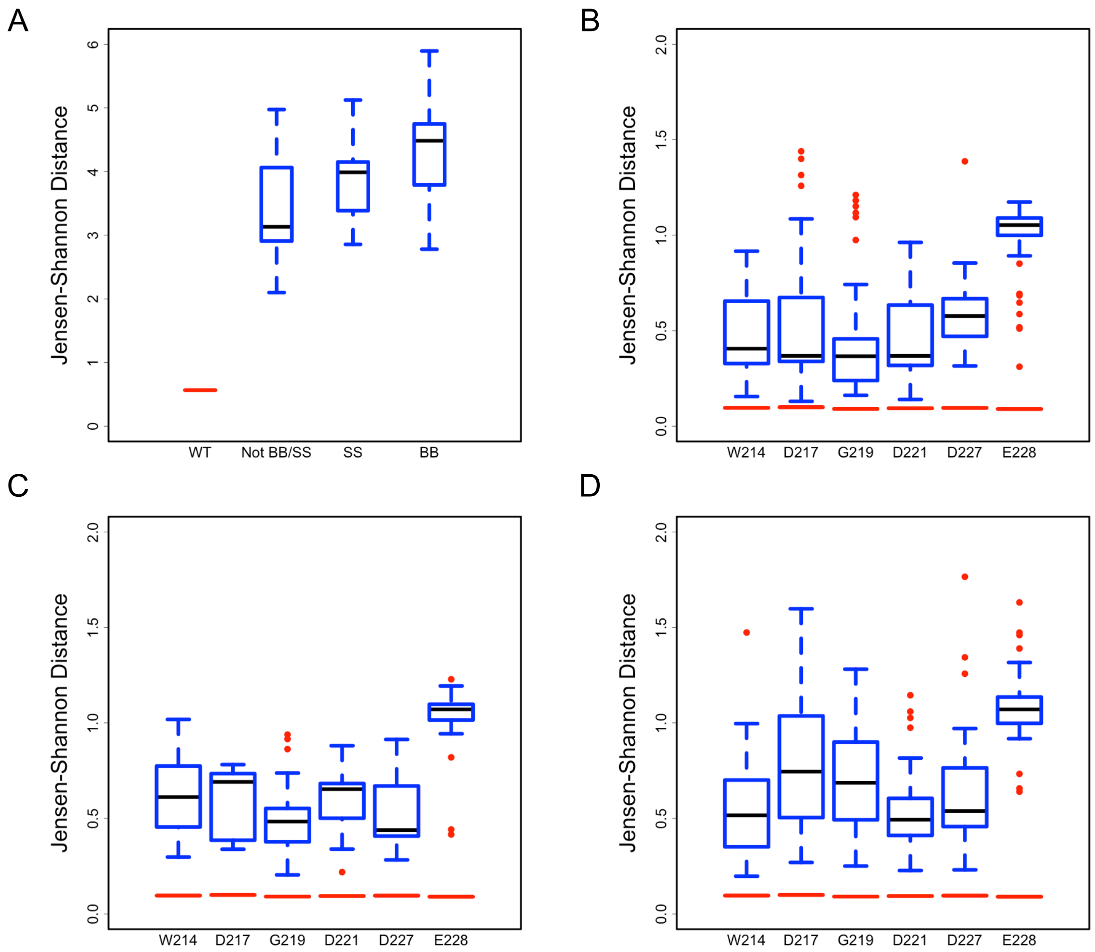


The boxplots of the fluctuations of all the angles formed between the atoms of two out of the six residues forming the calcium binding box in the sides of the angle, with the calcium ion in the vertex. Fifteen different angles could be formed (three of $180°$ and the rest of $90°$), and in each case we include the box-and-whiskers, also including the outliers of the distribution as red dots. The angle range close to which each angle must fluctuate is shadowed in blue. In the x-axis in each tick we include the name and sequence number of the two residues forming the angle. In the two bottom charts in which the amino acid substituted is one of the calcium binding box residues, the missing angles are highlighted with vertical blue lines

5 ns of simulation– from the 20 ns-long MD simulations. Then, each structure in these random sets is analyzed for obtaining the distribution of the ϕ and ψ angles of each of the six residues of the calcium binding box. These temporal distributions are compared using the Jensen-Shannon Distance[153] metric, with the corresponding ones derived from random samples of the simulation of the wild-type LA5 domain, obtained with the same procedure described above, please see the Methodology section for details.

These results are included in Figure 4.10, for the comparative study of the differences in the fluctuation of the distributions of the dihedral angles of the coordinating residues, for all the mutants in our dataset using the simulation of the wild-type

FIGURE 4.10: Assessment of the Structural Distortion of the Calcium Binding Box upon Mutation



Boxplots of the Jensen-Shannon Distance among the distribution of the dihedral angles of the coordinating residues throughout the simulation of mutants in comparison with the corresponding distributions in the simulation of the wild-type LA5 domain. **A**) Comparison of the global deformation of the complete calcium binding box by combining the individual JS_{dist} of the twelve ϕ and ψ angles. In the abscissa, the distributions of the wild-type (WT), mutants not involving residues from the calcium binding box or disulfide bridges (Not BB/SS), mutants in residues from the disulfide bridges (SS) and mutants in residues from the calcium binding box (BB). There exist statistical significant differences among the JS_{dist} distributions of the three mutants subgroups with respect to the wild-type ($\wp - value < 2 \times 10^{-3}$) with a significance $\wp - value < 0.01$. Panels **B–D**) include the distributions of the combination of the JS_{dist} for the two ϕ and ψ angles of each coordinating residue in mutants from the subgroups (Not BB/SS), (SS) and (BB) respectively. In each panel the corresponding values for the WT MD simulation are shown with red horizontal lines and the outliers as red dots

LA5 domain as reference. One of the advantages of using the Jensen-Shannon Divergence[148–152] formulation instead of the classical Kullback-Leibler Divergence[147], is that the former is a smoothed, symmetrized and bounded $(0 \leq JS_{div}(P_1, P_2) \leq 1)$ variation of the latter, and also it is directly related to a metric of the statistical difference among distributions[148–152], see more details in the Methodology section. This metric is the Jensen-Shannon Distance[153] $(JS_{dist}(P_1, P_2))$ which is the square root of the $JS_{div}(P_1, P_2)$, and accordingly is also bounded between $[0, 1]$, where 0 is obtained for identical distributions, while 1 corresponds to complete dissimilarity. As well as the Kullback-Leibler Divergence[147], the JS_{dist} is additive, thus allowing the linear combination of the distance estimates for the probability distributions of whichever two independent variables. Taking advantage of these properties, we were able to obtain some interesting insights regarding the effect of mutations in the conformational stability of the calcium binding box in the LA5 domain. As can be observed in panel A from Figure 4.10, almost all amino acid substitutions in any position in the structure of the domain, cause more or less significant perturbations in the local stability of the calcium binding box, since for the simulations of the 227 mutants, subdivided in three different subgroups, there are significant differences for the estimates of the box in the mutants with respect to the simulation of the wild-type domain. As anticipated intuitively, the less significant perturbations of the coordinating box on average are observed in the subgroup of mutants on which the substituted amino acid is neither one of the coordinating residues nor one of the six disulfide bridge-forming cysteines (Not BB/SS). This line of thought is also justified by the observation that for the subgroup of mutants in the cysteine residues (SS) even more marked perturbations are observed in the calcium binding box, and the most significant differences are of course found on average for the substitutions of any of the coordinating residues (BB).

Notwithstanding the logic trend observed for the averages of the three subgroups of mutants, the inspection of panel A in Figure 4.10 shows a significant overlap for the three JS_{dist} distributions. For example, in the subgroup (Not BB/SS) there are some mutants for which the cumulative JS_{dist} of the binding box are rather low, like some yet undescribed substitutions F**202**{*181*}L, K**223**{*202*}R and also for some mutations found in persons with FH, like F**200**{*179*}L and S**213**{*192*}T. In this same subgroup, rather high structural distortions are observed for a cluster of mutants (G**218**{*197*}S, D, C, A), in a residue located in-between two of the coordinating residues (D**217** and G**219**), showing JS_{dist} estimates as high as those observed for mutants in the binding box residues, corresponding to the maximum values of the (Not BB/SS) distribution. In the (SS) subgroup, the minimum of the distribution corresponds to mutants in residues C**197** and C**204**, which are sequentially and structurally distant to any of the coordinating residues, although in this case the absolute value for the minimum of the distribution is higher than that observed in the (Not BB/SS) subgroup. On the other hand, on the top of the (SS) distribution are located mutants in residues C**216** and C**222**, adjacent to coordinating residues D**217** and D**221** respectively. Many of these SNPs have been reported in FH patients, like (C**216**{*195*}R, Y, F) and (C**222**{*201*}R, G, F). The distribution of the (BB) mutants is more dispersed than the formers, and interestingly the lowest values of binding box distortions are observed for SNPs in residues W**214** and G**219**, which is consistent with the fact that these two amino acids interact with calcium by means of atoms of the backbone, see Figure 4.2, thus the substitution by other amino acids should have less significant effects on the local stability of the coordinating box. In accordance, the JS_{dist} values observed for some mutants in these residues are lower than the mean of the distributions of (Not BB/SS) and (SS). The maximum distortions of the box are found for mutants in residues D**217**, D**227** and E**228**, some of which have been previously described linked to FH, like D**227**{*206*}V and E**228**{*207*}K.

In panels B–C from Figure 4.10 we include a transversal representation of the local instability of each of the six coordinating residues in the mutants subgroups (Not BB/SS), (SS) and (BB) respectively. In this case, the local instability of each residue is approximated by summing up the JS_{dist} between the distributions of its two ϕ and ψ angles in the mutant when compared to the reference native distributions along the MD trajectories. In general, in the three subgroups of mutants, residue E**228** is the one that contributes the most to the instability of the binding box, although in the case of the (BB) mutants, there is an increase in the dispersion and the mean and maximum values of the distributions of almost all the coordinating residues. Also, these results are in agreement with those presented in panel A, which suggests that the amount of distortions caused by a given amino acid substitution in the coordinating box is case specific, since there is an evident overlap of the distributions for each individual residue.

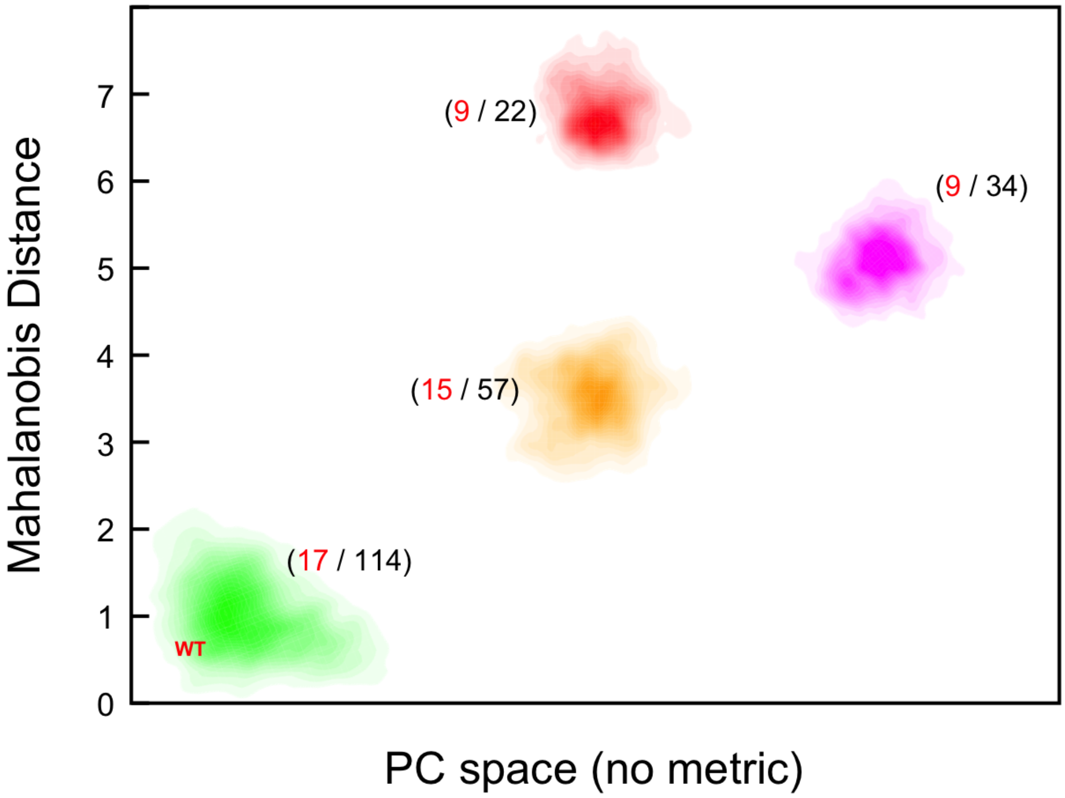
4.2.4 Clustering of LA5 Mutants According to the Extent of Conformational Instability

The image of the problem we are addressing in this work presented on Figures 4.5 and 4.7 for the first three PCs, although quite intuitive for graphically differentiating unstable from stable behaviors during different trajectories, is a reductionism of a more complex problem. It is known that although in PCA the first few eigenvectors usually describe most of the variance in the system, a complete description of the 'essential dynamics' can only be obtained by combining the contributions of the complete set of modes that describes $\approx 90\%$ of motion variance. Thus, in order to do that and to perform a thorough quantitative comparisons of different trajectories, it would be necessary to work in 20^{th}-dimensional spaces, or even beyond, with a significant increase of the computational complexity. Also, to try to go further subjective comparisons of the PCA results from different simulations (Figures 4.5 and 4.7), and to quantitatively compare the effect of amino acid substitutions based on the essential subspaces visited by different mutants during MD simulations, it is necessary not only to work with the complete eigensystem of the trajectories, but also to apply some strategies to successfully combine the conformational ensembles coming from independent MD simulations. The main problem in this regard is that the PCA of two independent trajectories render completely different eigensystems, probably with different dimensions describing dissimilar vector subspaces, in the sense of the orientation of the eigenvectors and the variance described by each. There are different approaches for overcoming these problems[154–157], and in our case we used an alternative approach based on concatenating sections or the complete trajectories of all the mutants into a meta-trajectory, and then performing the PCA to obtain a common PC subspace and eigensystem, which can be used to project and compare all the independent trajectories in the same essential subspace, please see the Methodology section for details.

Another problem for comparing the conformational essential subspaces visited by different mutants during MD simulations is that of contrasting stable simulations like those described in Figure 4.7 against other unstable simulations like in Figure 4.5, not following a Multivariate Normal Distribution in the PC space. Therefore, we used a sampling methodology to compare the essential subspaces independently of the behavior of the systems, as described in detail in the Methodology section. By randomly comparing subsets of each simulation against each other (10^5 random comparisons), it was possible to obtain a statistically significant assessment of the mean distance between the essential subspaces visited by different mutants and the wild-type LA5 domain. As described in the Methodology section, we used the Mahalanobis distance metric[158] for comparing the trajectories, based on the fact that it gives a distance that is normalized by the percentage of variance described by each eigenvector. After performing this exhaustive comparison of the simulations, it was possible to make a clustering accordingly to the Mahalanobis distances, which is presented in Figure 4.11, and at the same time try to quantitatively predict the pathogenicity of each amino acid substitution –*i.e.* establishing a link between conformational instability and the likelihood of the expression of a LDL-r with an impaired capacity of interacting with the LDLs[55,64,101]– and also to group different mutations taking into account their putative relation with FH.

The results included in Figure 4.11 correspond to the analysis of the meta-trajectory of the concatenation of the last 10 ns of each simulation, which is the one that gave better clustering results. The results obtained for the meta-trajectories including all the frames and the last 5 ns of each trajectory, were on one hand rather noisy due to the

FIGURE 4.11: Clustering of LA5 Mutants According to the Extent of Conformational Instability



Clustering of all the mutants in the LDL-r LA5 domain according to the extent of conformational instability introduced by the mutation. For the meta-trajectory of the concatenation of the last 10 ns of each simulation, the average Mahalanobis distance among all pairs of simulations was used to assess the difference in the subspaces explored by each mutant in the PCA N-dimensional space (25-dimensions). Based on these distances, a complete-link based clustering algorithm was used, and the abstract representation of the four more representative clusters is shown. The clusters are color-coded: green (stable mutants), orange (unstable mutants), magenta (very unstable mutants) and red (highly unstable mutants). For each cluster we show in parenthesis the number of mutants found in persons with FH (in red) and the total number of mutants

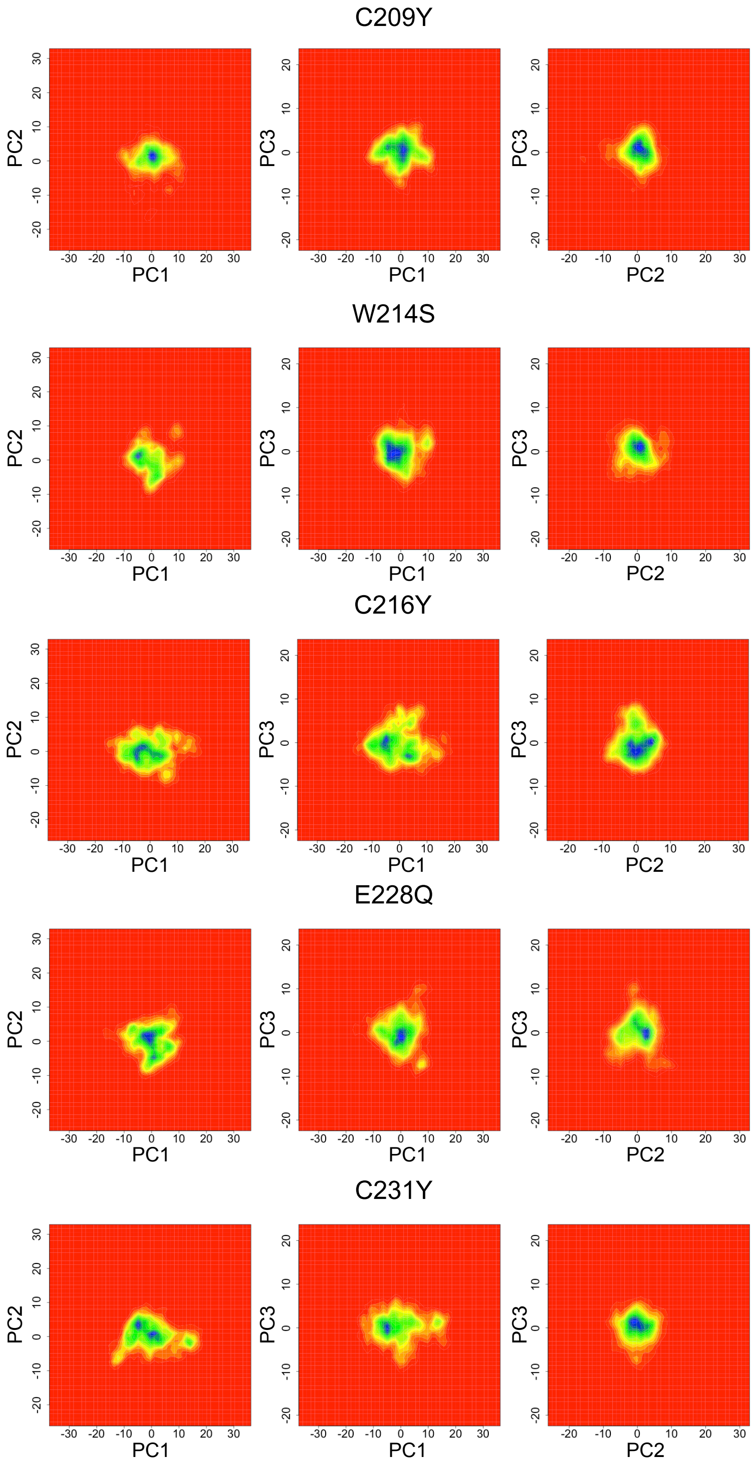
rather high conformational instability in the former case, on which some simulations explored an ample region of the PC space. In the later case, the low number of frames did not permit to make a representative sampling of the trajectories during the resmpling, resulting in some cases on statistically significant errors in the estimation of the mean Mahalanobis distances among simulations. Nevertheless, the last 10 ns meta-trajectory allows both a significant sampling and comparing the trajectories in a region on which most of them are fluctuating in a more limited subspace close to the final conformations the system is converging to. Figure 4.11 is an abstraction to graphically represent the clustering of the trajectories in a 25^{th}-dimensional space, on which the clusters would in reality correspond to hyperspheres. As expected, the projections for the trajectories of stable mutants and the wild-type LDL-r LA5 domain, explored an ellipsoidal subspace close to the $(0_{1^{st}}, 0_{2^{nd}}, 0_{3^{rd}}, \ldots, 0_{25^{th}})$ PC origin, and consequently cluster together in the (stable) cluster, in green in Figure 4.11. This is the largest cluster, comprising 114 mutants, for a detailed list please see the color codes in the Appendix Table F.2. In decreasing order with respect to the number of mutants are the orange cluster (unstable mutants) with 57, magenta cluster (very unstable mutants) with 34 and the red cluster (highly unstable mutants) with 22 mutants respectively.

The number of mutations found in persons with Familial Hypercholesterolemia that have been linked to the disease[33,73] is also unequally distributed among clusters. Despite being the largest cluster, the stable cluster only includes 34% of the known mutations, while the rest 33 out of 50 are distributed in the unstable clusters (Figure 4.11). The detailed case-by-case inspection of the group of known mutations included in the stable cluster gives some interesting hints into the problem introduced above regarding the establishment of a cause-effect among amino acid substitutions and the phenotype from sequencing studies. Because in these studies the genes are scanned in the search for a subset of the known mutations, there could be some cases on which the mutation found might not be the responsible for the disease, but another not identified mutation in the same gene, which is the responsible for affecting the stability of the protein, or by disrupting key binding sites for interacting with partners. The individual inspection of the dynamical evolution of some of the known SNPs classified by us as 'stable' mutations prove that, at least in the time range explored by us in this study, the structure of the LA5 domain is not significantly affected, as can be observed in Figure 4.12 for some examples[¶]. In these specific cases, as well as in others from the other 12 remaining cases, a comparison of the images and the accompanying videos in Figure 4.7 reveal that the conformational behavior is not affected in any case by the mutation –*e.g.* except subtle changes in the length of the ellipsoids in some axes– because the projections are normally distributed along each PCs as expected. Even more importantly, when the analysis is extended for all the PCs that describe as much as 95% of the total variance, the estimations of the Mahalanobis distance among them, on which the clustering in Figure 4.11 is based, unveil that all of them explore the same conformational subspace, along with the wild-type LA5 domain.

Another possibility is that these known mutations that do not affect the conformational stability of the LA5 domain, may impair the binding sites mediating the interaction of the LA5 domain with LDL or other proteins. In Figure 4.13, panel A we show

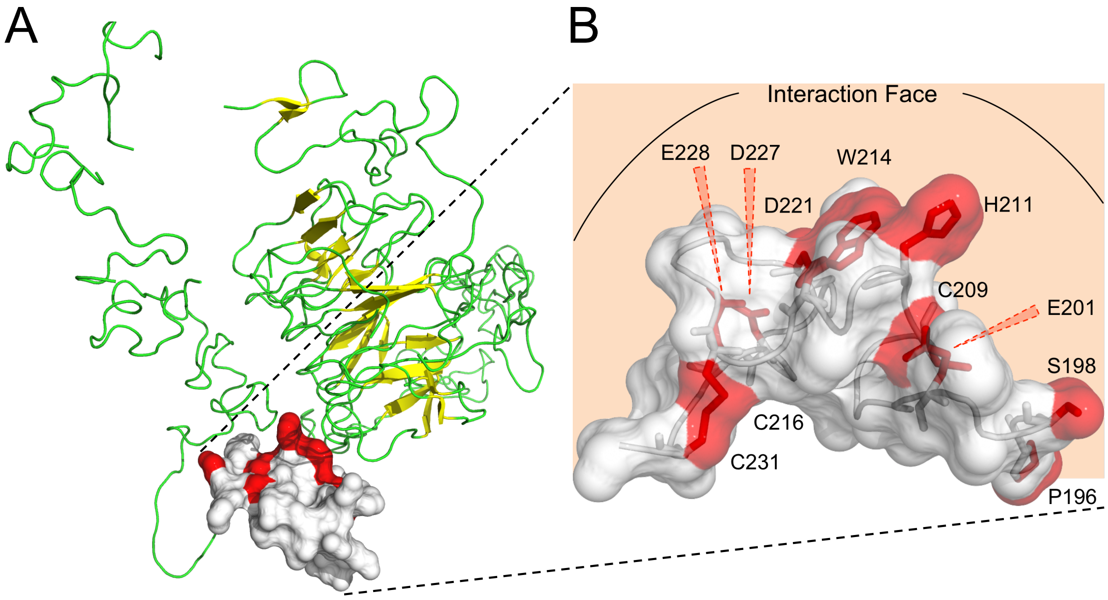
[¶] While displaying the videos there are different quality options, thus for a better video experience it is better to select the 720p in the bottom right angle of the video panel, in the **Settings** button. In case you are reading the printed version of this document, please refer to Appendix Table F.3 to obtain the URL link addresses

FIGURE 4.12: Dynamical Evolution of LA5 Mutants in the PCA Space (Stable FH Mutations)



The MD trajectories are followed along time by projecting the structures at each time step into the space described by the first three PCs. Each subchart is a two-dimension density plot of the projections of the structures into PC1 *vs* PC2, PC1 *vs* PC3 and PC2 *vs* PC3. The color scale goes from red (no occupancy) to blue (high occupancy), passing through intermediate scales of yellow and green. For accessing the more descriptive animations please visit the following links for each example: C209Y, W214S, C216Y, E228Q, C231Y

FIGURE 4.13: The Binding Region of the LDL-r LA5 Domain

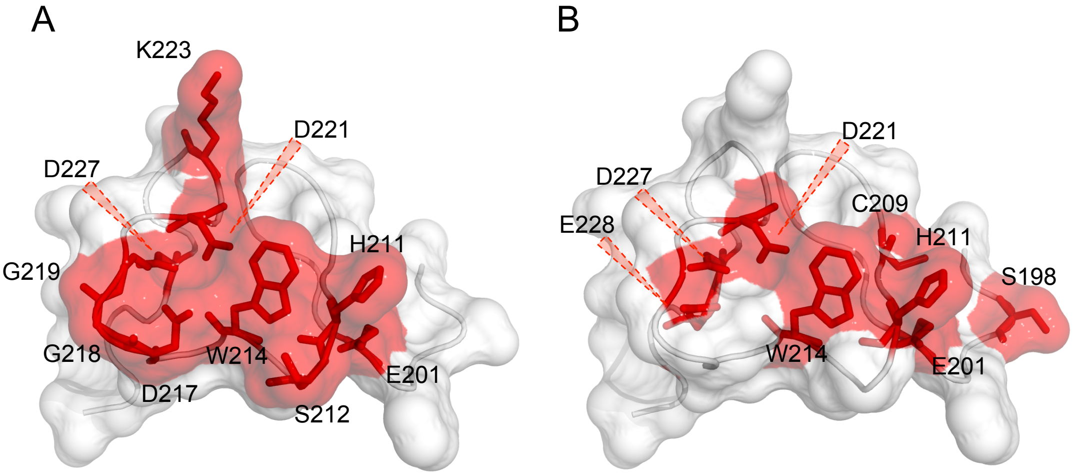


The structure of the LDL-r LA5 domain and the interaction region. **A**) The LA5 domain in the context of the structure of the complete LDL-r extracellular region (PDB id: 1N7D). The LA5 domain is shown in surface representation colored in white, highlighting in red the 11 residues on which occur the 17 mutations which do not affect directly the conformational instability of the domain. **B**) A close look of the LA5 domain and the residues which bear some mutations related to FH that do not destabilize the domain. We highlight the upper convex region of the LA5 domain, which according to recent experimental evidence, is the responsible for the interaction of the domain with other domains from the LDL-r and LDL particles. From the residues highlighted in red we also include their name and position in the sequence. Residues E**201** is oriented towards the top right face, while D**227** and E**228** are oriented towards the top back face

the structural context of the LA5 domain in the extracellular region of the human LDL-r[59]. According to this structural data, the LA5 domain is oriented establishing interactions through residues W**214**, D**217**, G**219**, D**221**, H**211**, S**212** and K**223**, which form a network of hydrophobic and salt-bridges contacts with various residues from the β-propeller domain, in a highly specific structural disposition, also involving a high degree of conservation of the interacting residues in other members of the LDL-r family[59]. Other reports for the direct interaction of LA modules with other proteins, such as the LA3–4 tandem with the Receptor Associated Protein (RAP), also underscored the convex face of these modules as the most important one for interacting with binding partners, which might correspond to a general mode for ligand recognition for LDL receptors[57]. Also, recent structural data from our group indirectly underlined the involvement of the convex face in the interaction with ApoB and ApoE[159]. The 17 known mutations included in the green (stable) cluster, because according to our simulations they cause only marginal distortions in the conformation of the LA5 domain, occur in 11 residues (please see

Appendix Table F.2 for details of the specific amino acid substitutions) 8 of which are located in the 3D structure in the convex face of the domain, with the exception of P**196**, C**216** and C**231** (Figure 4.13, panel B). This might explain the pathogenicity of most of these non-destabilizing known mutations by a disruption of the LA5 binding compatibility with other proteins. Indeed, there is a significant overlap of the interaction patch including residues in the convex face of the LDL-r LA5 domain, that according to experimental data participate in interactions with other domains from the LDL-r[59] and ApoB and ApoE[159], and the patch described by residues which bear 11 of the above mentioned non-destabilizing mutations related to FH that are classified in the stable cluster, please see Figure 4.14.

FIGURE 4.14: The 'Consensus' Binding Region of the LDL-r LA5 Domain



An upper view of of the LA5 module from the interaction face, as defined in Figure 4.13. **A**) The residues in the convex face of the LDL-r LA5 domain participating in interactions with other domains from the LDL-r[59] and ApoB and ApoE[159], are highlighted in red. **B**) The residues which bear some mutations related to FH that do not destabilize the domain and are classified in the stable cluster, are highlighted in red. In these 8 residues are distributed 11 out of the 17 non-destabilizing known mutations

On the other extreme in the scale of conformational instability is the red (highly unstable) cluster, in which are included known FH mutations like C**197**{*176*}G, Y and F, F**200**{*179*}C, C**204**{*183*}Y and E**228**{*207*}K among others, from which C**197**{*176*}G, F and C**204**{*183*}Y are by far the most destabilizing SNPs. Taking into consideration what we have just described before for some mutants in residues from the LA5 domain structural *loci* (see Figure 4.12), on which the global conformation remains more or less unaffected, it is worth saying that mutations from the magenta (very unstable) and red clusters (highly unstable) tend to concentrate in residues forming these structural *loci* such as D**221**, D**227**, E**228** and C**197**, C**204**, see Appendix Table F.2 for a complete description. Notably, as previously described for the instability of the calcium binding box, few of the mutations on W**214** and G**219** belong to the red and magenta categories, which could be explained taking into consideration that these two coordinating residues interact with the calcium ion through backbone atoms. In this same line, for the disulfide-bridge forming cysteines C**216**–C**231**, the number of SNPs in the most unstable clusters is also fairly low in comparison to the other two, which might be related to the fact that it is more or less buried in the core of the LA5 domain. Finally, SNPs in the orange (unstable) cluster, are more or less evenly distributed throughout the structure of the LA5 domain, in some cases flanking residues from the structural *loci*, but in other cases in residues forming them.

4.3 Discussion

4.3.1 Wide-ranging Direct Structural Assessment of the Effect of SNPs in the LDL-r LA5 Domain

The concept of Conformational Diseases[160–164] has gained great popularity, from the realization that some of the events that affect the correct adoption of the native conformation of proteins, widely studied *in vitro* for many years, could play an important role *in vivo* and be the main cause of many serious diseases. It is now well described that changes in the physicochemical conditions in the cell environment or mutations, are related to pathological states arising from alterations in the protein conformational equilibrium. This could result in a reduction in the quantity or the quality of the protein that is available to play its normal role[160,161], and in other cases lead to a transition to aggregation-prone conformations, resulting in the accumulation of protein aggregates[162–164]. In Familial Hypercholesterolemia, an ample number of mutations identified by cascade screening assays of people with the disease or populations of risk[83,85,89,90], have beed reported to be genetic determinants of the disease[74–76]. Although there have been attempts to experimentally asses the real effect of mutations in some of the domains of the LDL-r[56–58,68–70,97–99], much remains to done to explore the consequences of all the biologically accessible mutations, and how they might be related to pathological phenotypes. Indeed, of all the possible mutations in this key protein in the cholesterol metabolism and the main player in this disease, only a small proportion has been catalogued and reported in genetic variations and sequence databases[33,73,77–79]. There are a great variety of computational methodologies available for trying to predict the fate of mutations in proteins[104,107–109,113,114], but they are mainly based on genetic, structural or evolutionary assumptions, and they can not anticipate the real effect of the amino acid substitution at the structural level, and whether it might cause a perturbation in the protein conformational equilibrium. Besides, these methodologies are of course biased by the representation of deleterious and neutral genetic substitutions for different types of proteins in the training sets, and also are strongly dependent on the quality of the multiple alignments used to try to trace evolutionary relationships, and thus may render results with different confidence levels for different groups of proteins. To try to tackle this important problem, and to make a complete exploration of the effect of all biologically accessible mutations caused by SNPs in the structure of the LDL-r LA5 binding domain, and how they might relate to FH, we present here a computational strategy based exclusively on structural information and relying on short MD simulations.

In our study, we have generated at the structural level all the possible mutants arising from SNPs in the cDNA for the LDL-r LA5 binding domain (Figure 4.1), and have then performed short MD simulations in combination with a thorough data mining analysis to try to correlate conformational instability to disease phenotypes in FH. Our proposition is inspired by recent reports demonstrating the feasibility of the use of MD to make detailed studies of the variations in the conformational behavior of small proteins caused by mutations[101,122,123]. Distinctively, instead of concentrating on some specific amino acid substitutions, we explored all the mutational landscape of the key domain for establishing functional interactions with LDLs[55,64], and the one that is encoded in the LDL-r exon which exhibits the higher susceptibility to polymorphisms, and that bears the higher proportion of mutations identified in persons with FH. From the inspection of the 3D structure of the LA5 domain (Figure 4.2) some important structural *loci* can be easily identified, on which in principle the substitution of an amino acid could be accompanied by a significant change in the conformational stability. However, our results point out to a case-dependent scenario where the specific physicochemical perturbations caused by the amino acid substitution, determines whether the structure of the LA5 domain will be significantly affected or exhibit stable dynamical behavior during MD simulations. This finding is illustrated in Figure 4.3 with a selection of trajectories from the 227 possible mutants generated by biologically accessible SNPs. Some of these mutants have been reported as markers of FH in databases[33,73]. The evolution of the $TM - score$[130,131] along the simulations prove that, despite involving in some cases residues from the calcium coordinating box or the disulfide-bridge forming cysteines, some simulations show stable evolution along time, while for others the conformation stability appears to be significantly affected. Also, as described in Appendix Table F.1, there are disparities in the prediction of the deleteriousness of mutations using available approaches[108,119,124,126,127], which highlights the need for computational methods that evaluate directly the fate of mutations in proteins, and allow differentiating substitutions affecting conformation from others that could affect function by other means –*e.g.* by affecting conserved active center or binding site residues.

4.3.2 Local and Global Conformational Instability in the LDL-r Binding Domain: Relating SNPs with FH

An essential step of a wide-ranging study such as this is to take advantage of tools for processing the large amounts of high-dimensional data generated from a multitude of trajectories, and to perform data mining for extracting biologically relevant information. In this sense, we performed PCA in all the trajectories for describing the 'essential dynamics'[132,139–142] and to establish quantitative criteria for comparison of the conformations visited by different types of mutants during MD simulations. In this regard, the preliminary outcomes from this PCA study, as shown in Figure 4.4, allow to rate the structural differences among the average structures of each trajectory, based on the $TM - score$. In this chart clusters of mutants that are structurally different from any other are evident, while other mutants share high structural resemblances. This suggests the possibility of grouping different mutants into clusters according to the extent of the instability introduced by the amino acid substitution, by using a more sphisticated procedure than just comparing the average structures along the trajectories. Further insights of the dynamical evolution was obtained from Principal Component representations depicted on Figures 4.5, 4.7 and the 3D composite equivalent Figures 4.6, 4.8. The evolution of the projections of the conformations visited by some destabilizing mutants (Figure 4.5), most of which are mutations associated to FH (Appendix Table F.1), graphically confirms the scope of the instability caused by each mutation from the distancing from the Multivariate Normal behavior. On the other hand, Figure 4.7 shows the other extreme cases, with stable conformational evolutions around the average. These results reaffirm the discussion regarding case-dependency introduced above, as for examples such as C**209**{*188*}W of Figure 4.7, as well as in others included in Figure 4.12 for some amino acid substitutions in residues from the disulfide-bridges or the calcium coordinating box, we observed little or none modifications of the conformational equilibrium during the MD simulations.

Aiming at understanding the specific effect of mutations in the local stability of the calcium binding box, we also set up a thorough bootstrapping assay for estimating the statistically significant differences among the distributions of the dihedral angles of each calcium coordinating residue. As shown in Figure 4.9 the overall octahedral structure of the binding box is significantly affected in some cases, and it would be expected that substitutions in coordinating residues would determine significant distortions, such as those shown in the two bottom charts of this figure. On the other hand, at least for some mutations in the disulfide-bridge forming cysteines (top right chart in Figure 4.9), the structure of the binding box remains fairly stable during the complete simulation. The quantitative assessment of the stability of the calcium binding box could be obtained by estimating the Jensen-Shannon Distance[153] among the bootstrapped distributions of the ϕ and ψ angles of each coordinating residue, in comparison to the stable equilibrium simulation of the wild-type LA5 domain. The results depicted in Figure 4.10, panel A prove that though almost all mutations in the LA5 domain cause more or less some instability in the binding box, the proportion of the instability depend on the structural environment around the amino acid to be mutated. While it is clear that, on average, mutations in residues not directly forming part of the structural *loci*, cause less instability than disulfide-bridge cysteine mutations, and less than coordinating residues, respectively, the significant superposition of the distributions of JS_{dist} for these three subgroups demonstrates that for some particular cases the instability introduced can be considerably high. For example, mutations of G**218**, structurally close to two binding box residues, cause significant distortions in the binding box, while mutations of C**197**, C**204** cause mild distortions as they are structurally distant, but mutations of C**216**, C**222** destabilize the box thanks to their structural proximity. The hotspots regarding binding box destabilization among coordinating residues concentrate on residues D**217**, D**227** and E**228** while for W**214** and G**219** the distortions are less significant, which can be rationalized from the way in which each residue interacts with the calcium, the first three by means of the side-chain, and the last two with backbone atoms. The variability of the local instability and the abundant presence of outliers in the distributions of the JS_{dist} of each coordinating residues (Figure 4.10, panels B–D) justify this case-dependency for destabilization of the binding box by mutations.

4.3.3 From Molecular Dynamics to a Strategy for Computational Diagnosis in Conformational Diseases

The final goal of our work is to devise a rational and quantitative way of identifying SNP mutations that could destabilize the structure of the LDL-r LA5 binding domain, and also to group different types of mutations according to the possible destabilizing effects. As pointed out above, to do this thoroughly, it is indispensable to work with the complete eigensystem describing the 'essential dynamics' of the system, and to put all the trajectories on the same comparative context. The resampling presented here for comparing the essential subspace more probably visited by each mutant during the MD simulations gave us the possibility of doing that, and the results in Figure 4.11 give a good summary of how different mutants group together according to their possible pathogenicity. This figure, along with Appendix Table F.2, offers a color-coded classification of the destabilizing effect of each mutant, as inferred from a clustering based on the Mahalanobis distance among trajectories, estimated from an exhaustive all-to-all comparison of the essential subspace explored in each simulation. In principle we believe that the discrimination achieved with this strategy is fairly promising, as we were able to successfully classify 66% of the known mutations into one of the three clusters of (unstable), (very unstable) or (highly unstable) mutants. Also, regarding the discrepancies with the rest 17 known mutations that are classified by us as 'stables', the results in Figure 4.12 prove that at least in the time-range of this study, these mutations do not significantly alter the conformational behavior of the LA5 domain.

This might be an indication that they could be related to the disease not by affecting the conformation of the LA5 binding domain, but by impairing the interaction with other domains from the LDL-r binding region or other proteins from the cholesterol metabolism pathway, or by causing a decrease in the expression efficiency with the concomitant reduction of the quantity of membrane receptor molecules. The pathologic nature of some of these 17 mutations could be explained following this rationale, as in most cases they cause a substitution in residues presented to the interacting face of the LA5 domain, please see Figures 4.13 and 4.14. From the structural point of view, there could be two different scenarios, one in which the substituted amino acid is essential for establishing direct contacts with specific residues from the binding partners, or a second one in which the mutated residue, even not being directly involved in the formation of non-covalent bonds, could change the physicochemical compatibility of the interaction patch. Examples of the former case would be mutation W**214**{*193*}S, which would abolish the multiple hydrophobic interactions among this residue and amino acids E**602**, K**603**; and D**221**{*200*}N, G that would also eliminate the interaction with K**603** from the β-propeller domain. Also, the substitution S**198**{*177*}L which would break an hydrogen bond with LA4 domain residue R**185**, can be included in this case. It could also be the case of mutations H**211**{*190*}D, Y, L in a residue which forms a cluster of histidines with β-propeller domain residues H**583** and H**607**, which are responsible for determining the acid sensitivity essential for endosomal LDL release[58,59,165]. Experimental data for tyrosine substitution of these histidines proved a significant reduction of the expression of LDL-r molecules in fibroblasts[97], and the reduction of the efficiency of LDL release in the endosome[58,165]. Actually, as can be inferred from recent structural experiments from our group, residues W**214**, D**221**, D**227** and E**201** positioned in the convex face of the LA5 domain, are the responsible for the interaction with ApoB and ApoE[159], which is essential for correct internalization of LDLs and cholesterol metabolism.

There is another possible scenario, in which can be included the other 3 known mutations out of the remaining 9 clustered in the 'stable' cluster –*i.e.* excluding those described above implicated in direct interactions with other domains or proteins– when the mutation could affect a residue and change the physicochemical compatibility of the interaction patch. In this group can be included mutations C**209**{*188*}Y, D**227**{*206*}E and

E**228**{*207*}Q, all involving residues located in the convex interaction face of the LA5 domain, see Figures 4.13 and 4.14, panel B. For these substitutions, notwithstanding there is no direct evidence of their implication in the formation of interactions upon binding to other domains, they could be related to FH by modifying the physicochemical and steric conformation of the interaction surface, significantly impairing the binding efficiency. Thus, our results demonstrate that most of the known mutations that do not cause destabilizations of the conformational equilibrium of the LA5 domain, occur in residues from the interaction region. Considering this, our computational estimations of the conformational instability caused by mutations, combined with the experimental knowledge of the LA5 domain interacting residues, makes possible to anticipate the disease phenotype for 44 out of the 50 SNPs known to be related to Familial Hypercholesterolemia. Following this idea, it would be possible to make a tentative computational diagnosis for this disease, taking into consideration both the degree of conformational instability and the possible impairment of the interacting region, as proposed in Appendix Table F.2. In this table, all the mutations not reported in genetic variation databases are classified solely taking into consideration its destabilizing effect. However, for mutations in residues from the interaction region, we tentatively evaluated the possible impairment that the amino acid substitution could cause in the binding efficiency. Anticipating the effect of mutation in protein binding regions in the binding efficiency is difficult, specially if, as in our case, there is no detailed knowledge of the binding partner. The information extracted from the structure of the LDL-r complete extracellular region[59] is limited because its low resolution does not allow an accurate appraisal of some non-covalent interactions that could exist between the LA5 and β-propeller domains. Also, in the case of the interaction with lipoprotein peptides[159], the structural information is indirect, because there is no information of the exact conformation of the peptides in the complexes. However, we indirectly approximated the effect of mutations in binding, by using qualitative structural criteria –*e.g.* steric and physicochemical differences between the wild-type and substituted residue– and combined this information with conformational instability, to propose the disease phenotype in mutants in interacting residues, please see Appendix Table F.2. However, in these cases, due to the uncertainties in the evaluation of the effect of mutation in binding as discussed above, our predictions are less conclusive.

The remaining 6 cases correspond to substitutions in residues P**196** and the disulfide-bridge forming cysteines C**216**–C**231**. These cases are special because these residues are not in the interaction face of the LA5 domain, but on the other hand they are located in the N- and C-terminal ends. These amino acid substitutions could cause a significant distortion of the possible orientations of the LA5 domain with respect to domains LA4 and LA6, also altering the interaction efficiency of the LDL-r binding domain. For example, P**196** is located in the linker connecting domains LA4–5, in which it may play a role in reducing the flexibility of the linker, to guarantee specific conformations of these domains to interact with LDL[59]. Mutations in residues C**216**–C**231**, on the other hand, would cause a destabilization of the native conformation of the linker connecting domains LA5–6, which could cause an increase of the length of the linker region, contributing to the adoption of conformations in which these two domains are not correctly oriented for binding, as previously studied experimentally[166]. This might explain the relation with FH for those known mutations that have been classified by us as stables, as in those cases the specific physicochemical or steric changes caused by the substitutions could affect the binding interface compatibility among the LA5 domain and lipoproteins or other biological binding partners. Also, it might be the improbable but non negligible case that some of these stable known mutations were false positives, and are not related to FH in any way, but have been identified by chance during sequencing of the gene in persons with FH, on which the real genetic variation responsible for the phenotype is in another region of the same LDL-r gene, or other genes from the cholesterol uptake pathway[25–31].

According to our calculations, the structure of the LDL-r LA5 domain has some hotspots on which there is an accumulation of SNPs resulting in substitutions that might affect its conformational stability. Those are D**221**, D**227**, E**228** and C**197**, C**204**, C**222**, which in some cases coincide with residues whose substitution destabilize the calcium binding box, as described in the previous section. The same trend is observed for mutations on coordinating residues W**214** and G**219**, which besides affecting the less the stability of the binding box, also appear to accept few mutations that significantly destabilize the complete structure of the LA5 domain, whilst as exposed above, there are some exceptions that oppose this trend depending on the case-specific physicochemical and structural environment associated to the amino acid substitution. Recent experimental reports from our group of low-resolution studies of the oxidative unfolding of the wild-type and mutated LA5 domain[69] have proven *in vitro* that mutants E**208**{*187*}K and D**221**{*200*}G can fold into a native-like conformation, though less efficiently than the wild-type, because of the accumulation of intermediates with a scrambled disulfide pattern. For two other mutants: D**224**{*203*}A and D**227**{*206*}E the oxidative folding leads through a path of scrambled isomers and it is impossible to reach the native structure. The last of the two mutations described before has also been assayed experimentally *in vitro*, along with amino acid substitution D**224**{*203*}G in another report[167], also concluding in the existence of folding defects in the final LA5 species. These mutations have been classified by us as unstable, stable, unstable, stable and unstable, and although these findings have not been tested experimentally with high resolution techniques, they suggest that some mutations that do not destabilize the LA5 structure when folded, could affect the oxidative folding, an event that can not be accounted for computationally without the use of more sophisticated MD simulation techniques and longer time-scales.

There is a lot that needs to be done to understand the complete picture of the interaction among the LDL-r binding domain and LDLs, and how the efficiency of the interaction is affected by the independent contribution of different LA domains, and the interactions established among them during LDL-recognition. With our work we believe we have provided a clear view of the complete mutational landscape of the LA5 domain, with a quantitative classification of the conformational instability caused by each different biologically accessible amino acid substitutions, which could be useful for planning experimental tests to measure the real extent of the structural instability of the domain upon mutation. By extension, as small disulfide-rich domains are fairly widespread in the protein universe, our outcomes could also serve as a reference for studying how mutations could affect the structure and the function of other proteins bearing this kind of interacting domains associated to other pathologies, or even in other small proteins. Though undoubtedly our approach is more computationally expensive and requires more data processing and analysis than others available for predicting the deleteriousness of mutations[108,109,115,119,124–128], with the advent of great advances in the field of MD simulations for reaching longer simulation times[168–173], which together with the emergence of online services for performing client-based and high-throughput MD simulations[174–176], performing studies such as ours will become feasible in the near future.

4.4 Conclusions

In this work we have presented an alternative strategy for studying the complete SNP mutational space of a protein using Molecular Dynamics, and try to correlate the conformational instability introduced by amino acid substitutions with a disease phenotype. Differently to other bioinformatics approaches for predicting mutation deleteriousness based on protein family or evolutionary information, ours rely solely on the knowledge of the structure of a protein, as a starting point for performing short Molecular Dynamics simulations. We have been able to quantitatively classify different types of mutations according to the extent of local and global conformational instability caused by the amino acid change, and also have found that our predictions are in fairly good agreement with pathologic mutations reported in databases. Also, with our method, it is possible to directly appraise the real effect of mutations in the structure of the protein, and to differentiate mutations associated with the disease that directly affect the conformation, from others that do not, which can be useful to give some hints of mutations that might be affecting the function of the protein at other levels, instead of impairing its folding or conformational stability. Our results display a layout of all the biologically accessible mutations in the LDL receptor LA5 domain responsible for interacting with LDL, an impairment of which is the cause of one of the most common and serious diseases in human populations, Familial Hypercholesterolemia. As performing this wide-ranging studies experimentally is rather difficult, our work might be of help for proposing new SNP candidates for being studied *in vitro* to finally assess the real cause-effect of SNPs in this disease. Also, as the LA5 domain is fairly abundant in the protein universe, our propositions could also be of value for studying the effect of mutations in the conformation of other proteins bearing this kind of domains and related to other diseases, for which there is less experimental, mutational or evolutionary information. Besides, our approach could be used to study other small proteins or independent folding domains associated to other molecular functions, like SH3 or PDZ.

4.5 Methodology

4.5.1 LA5 Domain Coding Sequence, Structure and Complete SNP Mutational Map Generation

We started from the protein sequence for the complete human LDL-r accessible in Uniprot[77] (ID: LDLR_HUMAN, AC: P01130), from where we extracted the DNA coding sequence for the LA5 domain by accessing to the entry for this gene in the Ensembl database[177] (ID: ENSG00000130164). The protein sequence for the LA5 domain corresponds to residues 195–233 in the sequence of the complete receptor, while in the structure of the domain used as the starting point for the structural analysis (PDB id: 1AJJ), are included residues 196–232. Thus, we just extracted the coding sequence for amino acids 196–232, leaving out the codons for the N- and C-terminal serine and valine. The cDNA sequence was then processed with an *ad hoc* script for generating all the biologically accessible mutants arising from the substitution of a single nitrogen base (SNPs), please see Figure 4.1. All the non-synonymous SNPs were identified –*i.e.* 256 non-synonymous SNPs coding for 227 different single amino acid substituted protein variants– and then we generated all the corresponding mutations in the structure of LA5 domain using the program SCWRL[129], for finding the best rotamers for the side chain of the mutated amino acid. Then, we organized the mutants with a specific code for each one (Appendix Table F.2), to be further processed before running the MD simulations.

4.5.2 Setting up the Systems for Molecular Dynamics Simulation Production

Each of the 227 mutants, plus the wild-type LA5 domain, was solvated in a cubic water box with approximately 5500 TIP3 water molecules, and neutralized with $Na^{+}Cl^{-}$ counter ions using the solvate package in VMD[178]. We set up a thorough procedure for preparing the systems previous to run the production MD simulations, including multiple cycles of step-descending minimization/equilibration steps in a preparation phase of about $5\ ns$ of simulation, which encompasses: **a**) short CPT dynamics of water molecules with the protein atoms fixed to eliminate the possible potential strains in the water box, **b**) slow release of the protein atoms by imposing decreasing elastic restraints and **c**) very slow heating of the systems to the final simulation temperature $(310\ K)$ using a gradient temperature ramp. We followed the evolution of the systems during the preparation phase to check for the fluctuations of the different energies –*e.g.* the potential, kinetic and total energies– and the duration of the preparation $(5\ ns)$ was set in consequence to guarantee the stabilization of all the system variables. Also, during the slow heating temperature ramp step, we also checked in the final step that the temperature of the systems was stabilized at the desired temperature. Then, the production MD simulations were run for each mutant, in a production phase of $20\ ns$ using NAMD[179] and the CHARMM[180] force field. The simulations were run using Langevin Dynamics, with periodic boundary conditions and Particle Mesh Ewald (PME) for modeling long-range electrostatic interactions with a cutoff distance of 14 Å. The Nosé-Hoover thermostat was used for pressure coupling of the system and the friction coefficients of atoms to be used in the Langevin formulations were set to 0.5 and $60\ ps^{-1}$ for protein atoms and water molecules and ions respectively. All the simulations were run mainly in the *Marenostrum* Supercomputer, and also in the *CaesarAugusta* and *Terminus* clusters. The trajectories were analyzed with VMD[178] and a set of *ad hoc* TCL and Perl scripts.

4.5.3 Principal Component Analysis of MD Trajectory Data

Principal Components Analysis (PCA) is a procedure from the field of multidimensional statistics based on performing a linear transformation of data, very useful for capturing the correlations among variables and significantly reducing the number of degrees of freedom, and at the same time hierarchically decomposing the variance in data. This technique has been extensively used for analyzing MD trajectory data aiming at describing the 'essential dynamics'[132,139–142], and to underscore the motions that determine the characteristics of the conformational ensemble of the system. The procedure for performing PCA on MD data starts by removing the translational and rotational components of movement throughout the trajectory, thus commonly it is necessary to align all the snapshots using one of the available methodologies for finding the best solution to transforming structural data[181]. Then, the trajectory is centered to the reference structure S_{ref} –*e.g.* the initial or an average structure can be used– by subtracting the reference structure to the aligned snapshots and it is represented as a matrix of the type ($T_c = [3N \times F]$), on which the rows are the coordinates of the N residues of the system, and the columns the number of frames or snapshots F of the trajectory. Subsequently, the covariance or correlation matrix is calculated from the product of the trajectory matrix by its transpose ($\Sigma = \frac{1}{3N} T_c \cdot T_c^T$), which has the dimensions $[3N \times 3N]$. The eigenvalue decomposition of the covariance matrix renders a set of eigenvalues and orthogonal eigenvectors –*i.e.* $3N$ eigenvectors result from the matrix decomposition if $F > 3N$, but there are six zero eigenvectors for the translational and rotational movements which are excluded from the subsequent analyses– organized in the form ($\Lambda = V^T \cdot \Sigma \cdot V$), where Λ is the diagonal matrix of the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_{3N-6})$ and V is the matrix of the $3N-6$ eigenvectors paired to the eigenvalues. The eigenvalues are sorted in descending order with respect to the amount of variance of the original data described by the pairing eigenvectors.

After obtaining the eigensystem from the covariance matrix as described above, it is possible to take advantage of the significant reduction of the multidimensionality of data –*e.g.* usually just a few eigenvectors are enough for describing most of the variance (see Appendix Figure G.1), and in practice there is no need to work with all the eigenvectors, but to retrieve the number of eigenvectors sufficient to describe, for instance, 90 or 95% of the total variance in the system. One of the possibilities is to project the coordinates in the cartesian space coming from the simulation into the eigenspace, which is known as the Principal Components representation. Following the symbology described above, the structure at time S_{t_i} represented as a vector of the type ($S_{t_i} = [1 \times 3N]$) can be projected into the Principal Components space according to ($P_{t_i} = S_{t_i} \cdot V$), resulting in a vector of scalar values of the projections $(p_1, p_2, \ldots, p_{(3N-6)})$ of the structure at timestep t_i into the eigenvectors $(v_1, v_2, \ldots, v_{(3N-6)})$. Hence, the projections of all the snapshots in a trajectory are obtained by multiplying the transpose of the trajectory matrix and the matrix of eigenvectors ($P = T_c^T \cdot V$).

$$S_{t_i} = S_{ref} + p_1 \cdot v_1 + p_2 \cdot v_2 + \ldots + p_{(3N-6)} \cdot v_{(3N-6)} \tag{4.1}$$

As shown in Equation 4.1[132], the projections in the PC space are the weights for scaling the pairing eigenvectors for regenerating the coordinates of structure S_{t_i}, by adding the scaled eigenvectors to the reference structure S_{ref}. Seen from another angle, the higher the values of the elements in the vector resulting from the series of products

$(p_1 \cdot v_1 + p_2 \cdot v_2 + \ldots + p_{(3N-6)} \cdot v_{(3N-6)})$ the higher the structural differences between S_{t_i} and S_{ref}. Formally, for a simulation in equilibrium, on which the conformational ensemble must oscillate harmonically around the reference structure, the geometrical representation expected for the distribution of the projections must be a set of ellipsoids in the $3N - 6$ dimensions. This can be deduced from the analysis of the *Probability Density Function* of Multivariate Normal Distributions:

$$f(x,y) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp(-\frac{1}{2}(x-\mu_{(x)})\Sigma(y-\mu_{(y)})^T) \tag{4.2}$$

where the *Moment Generation Function* is the exponential element, and defines the geometrical ellipsoidal representations for independent variables, each distributing following a normal distribution. In this equation, the $(x\{y\} - \mu_{(x\{y\})})$ parameters correspond to the dispersion of the distribution in the $x\{y\}$-plane. In the case of the eigensystems generated from PCA, the eigenvectors are orthogonal, thus the projections in these eigenvectors should distribute as Multivariate Normal Distributions as described above in a simulation in equilibrium. Specifically, in the case of the eigensystem of a simulation, the *Moment Generation Function* in Equation 4.2 is generated from $(T_c^T \cdot \Lambda \cdot V)$, which results in a matrix of the projections in each PC scaled by the eigenvalues $(\lambda_1 p_1, \lambda_2 p_2, \ldots, \lambda_{(3N-6)} p_{(3N-6)})$ in each frame f_i –*i.e.* with dimensions $[F \times (3N-6)]$. In this representation the λ_i is the factor of proportionality with the square of the length –*i.e.* dispersion– of the distribution in this component –*i.e.* the structural differences among the snapshots and the reference structure– and the projections are the variables determining the Gaussian behavior along PC$_i$. Accordingly, a deviation from this geometrical behavior can be an indicator of perturbations in the system, and the amount of this deviation as estimated from the projections can be taken as a quantitative measure of the perturbations.

We also carried out a procedure for quantitatively compare the PCA subspace explored by different mutants and the wild-type LA5 domain. In order to do so, we concatenated different subsections of all the trajectories –*e.g.* the complete, last 10 and 5 ns– into meta-trajectories using the VMD[178] *CATDCD* utility. After doing so, we recalculated the complete eigensystems for each meta-trajectory and obtained the projections of the frames of each independent simulation in the principal components using the meta-trajectory eigensystem. This is necessary because when comparing essential subspaces from different simulations, the subspace metrics are dependent on the dimensions of each specific subspace and the dimensions of the full vector subspace[143] –*i.e.* the PCA analysis from different simulations renders a different set of eigenvectors (in number and orientation) and a different set of eigenvalues (different percentage of variance described by each mode). By using this approach it is possible to put all the different simulations in a common PC space, and obtain a single set of eigenvectors for the complete system on which any independent trajectory can be easily projected into. After performing this procedure we intended to assess quantitatively the effect of mutations on the structure of the LA5 domain by calculating the distance among the subspace explored by each mutant and the wild-type domain. In order to do so, we utilize a metric routinely used in the field of multivariate statistics which is the Mahalanobis distance[158] $(MD_{pp'})$, which in contrast to the classic Euclidean distance, accounts for the correlations on data and is independent of data transformations. In the specific case of PCA, the Mahalanobis distance between a pair of points p and p' in the PC space is defined in Equation 4.3:

$$MD_{pp'} = \sqrt{\sum_{i=0}^{N} \frac{(proj_{p_i} - proj_{p'_i})^2}{\lambda_i}} \tag{4.3}$$

on which $proj_{p\{p'\}_i}$ are the corresponding projections in the N-dimensional PCA space, and λ_i is the corresponding eigenvalue for the PC$_i$. By using $MD_{pp'}$ it is possible to make a more realistic assessment of the distance among points in the PCA space, by normalizing the contributions of all the PCs according to the percentage of variance the pairing eigenvector describes, out of the total variance in the system.

For obtaining the mean distance among trajectories there is an additional problem in our case, which is related to the fact that for some mutants there is a significant deviation from the ellipsoidal behavior expected for simulations in equilibrium as described above. In those cases, obtaining the mean distance between a pair of simulations is rather complicated because non compliance to the ellipsoidal behavior makes it impossible to use the mean of the projections in the N-dimensional distribution. Thus, in this case, we set a resampling strategy to overcome these drawbacks and make it possible to estimate the real average distance between whichever two trajectories, notwithstanding whether they follow a Multivariate Normal Distribution. Specifically, for each pair of simulations, we resampled with replacement a subset of snapshots –*e.g.* 5 or 2.5 ns depending on the meta-trajectory– from each trajectory, and calculated $MD_{pp'}$ for all possible pairs of points in the N-dimensional PCA space –*e.g.* 20–25 eigenvectors. We repeated this step 10^5 times for each pair of simulations and from that exhaustive subspace comparison, we obtained in all cases normal distributions for the Mahalanobis distances among points in the trajectories, with rather low standard deviations. From this comparison we obtained the mean MD_{T_1,T_2} among whichever two trajectories. After calculating the distance matrix among trajectories according to the procedure described above, we performed a clustering –*i.e.* using a complete-link clustering procedure– of the trajectories according to the PCA subspace explored in each case. All the manipulation of MD data for PCA analysis was performed with a set of *ad hoc* TCL and Perl scripts, alongside with the package PCAZIP[‖]. We compressed all the trajectories using PCAZIP, taking into consideration only the backbone atoms of the LDL-r LA5 domain and retrieving in each case the number of eigenvalues and eigenvectors sufficient to describe 95% of the total variance in the system. Using one of the tools from the PCAZIP package we then extracted all the metrics and data used in the statistical analyzes in our study –*e.g.* eigenvectors, projections, etc. The processing of PCA data, and all the resampling, clustering and statistical analyses were done in the R statistical package[182] with a group of *ad hoc* R scripts.

4.5.4 Estimating Local Instability by Comparing Dihedral Angles using the Jensen-Shannon Distance

The trends of the distributions of ϕ and ψ angles of an amino acid during a MD simulation can be used as an indirect measure of residue local instability. When measured in a simulation perturbed in some way –*e.g.* by introducing a mutation[101] or run in potential accelerated conditions[144]– local instability estimates can be approximated by comparing the fluctuations of the dihedral angles of a residue or residues in the perturbed simulation, with a reference simulation in equilibrium. In our study, we carried out an evaluation of the changes in the local stability of the calcium binding box caused by mutations in the LA5 domain. For each mutant, we followed the temporal evolution throughout the MD simulations of the ϕ and ψ angles of the six residues of the calcium binding box (see Figure 4.2), which were calculated using the program DSSP[183,184] for each frame in the trajectory. Then, the distribution of each dihedral angle was compared with the corresponding one in the simulation of the wild-type LA5 domain, using the Jensen-Shannon Distance[153]. In order to guarantee the statistical robustness of our estimations, as well as to make a thorough comparison of the distributions and to avoid the bias of the results caused by outliers, we set up a bootstrapping assay for comparing the distributions of angle fluctuations. This is of great importance because, although in the simulations in equilibrium the dihedral angles distribute normally for the coordinating residues (not shown), in some mutants the perturbations caused could result in non-normal distributions, with a significant number of outliers. In that sense, it is important to choose a metric independent of the characteristics of the distributions to be compared, such as the Jensen-Shannon Distance. In the jackknife bootstrapping assay, from the MD simulation of each mutant we randomly extract with replacement 10^3 snapshots, which correspond to $5\ ns$ of simulation, and compare the random distributions of the two dihedral angles of the six coordinating residues, with the corresponding random distributions obtained from the simulation of the wild-type LA5 domain, following

[‖] PCAZIP repository: http://holmes.cancres.nottingham.ac.uk/pcazip/

a similar procedure as described above. We repeated these steps 10^5 times, for each mutant, and then calculated the average Jensen-Shannon Distance from the bootstrapped sets for the ϕ and ψ angles of each binding box residue.

The Jensen-Shannon Distance[153] $(JS_{dist}(P_1, P_2))$ is a metric of the statistical difference among probability distributions (P_1 and P_2), and is the square root of the Jensen-Shannon Divergence[148–152] $(JS_{div}(P_1, P_2))$, which is a special case of the λ-divergence $(\lambda_{div}(P_1, P_2))$, defined in Equation 4.4:

$$\lambda_{div}(P_1, P_2) = \lambda KL_{div}(P_1, \lambda P_1 + (1-\lambda)P_2) + (1-\lambda)KL_{div}(P_2, \lambda P_1 + (1-\lambda)P_2) \quad (4.4)$$

on which $(KL_{div}(P_1, P_2) = \sum_{i=1}^{N} \log(P_{1_i}/P_{2_i})P_{1_i})$ is the Kullback-Leibler Divergence[147]. The λ-divergence is a generalization of the original symmetrized Kullback-Leibler Divergence, and in the case of $(\lambda = 1/2)$, it transforms into the Jensen-Shannon Divergence, as defined in Equation 4.5:

$$JS_{div}(P_1, P_2) = \frac{1}{2}KL_{div}(P_1, M) + \frac{1}{2}KL_{div}(P_2, M) \quad (4.5)$$

on which M is the average of the probability distributions P_1 and P_2. The JS_{div} has many important characteristics, such as it is symmetrical and smoother in comparison to KL_{div}, and also it is bounded $(0 \leq JS_{div}(P_1, P_2) \leq 1)$ provided that $(\log_2)$ is used for calculating the KL_{div}. Therefore, the JS_{dist} also takes values between $[0, 1]$ and formally corresponds to the expected information gain when deciding (by means of a sample of length 1) between two distributions, given a uniform prior over the distributions[153]. Besides being a special case of the symmetrized Kullback-Leibler Divergence, the JS_{div} and in consequence the JS_{dist} are additive, which make them very useful for obtaining cumulative divergence estimates by means of the linear combination of divergence or distance estimates for the probability distributions of independent variables. We also assessed the statistical differences of the binding box instability among different subgroups of mutants –*e.g.* mutants not involving residues from the calcium binding box or disulfide bridges (Not BB/SS), mutants in residues from the disulfide bridges (SS) and mutants in residues from the calcium binding box (BB)– with the JS_{dist} value obtained for the inner comparison of the wild-type with itself, used as control in the bootstrapping assay (Figure 4.10). We used the non-parametric Mann-Withney-Wilcoxon rank-sum test with a significance $\wp - value < 0.01$ and we found significant differences in all cases, with an upper limit for the comparisons of $(\wp - value < 2 \times 10^{-3})$ obtained for the subgroup (Not BB/SS). All the bootstrapping and statistical analyses were implemented in R[182] with a group of *ad hoc* R scripts.

4.6 Bibliography

[1] Viviane Z Rocha et al. Extensive xanthomas and severe subclinical atherosclerosis in homozygous familial hypercholesterolemia. *J Am Coll Cardiol*, **61**: 2193, 2013. (see p. 152)

[2] Daniëlla M Oosterveer et al. The risk of tendon xanthomas in familial hypercholesterolaemia is influenced by variation in genes of the reverse cholesterol transport pathway and the low-density lipoprotein oxidation pathway. *Eur Heart J*, **31**: 1007–12, 2010. (see p. 152)

[3] Daniëlla M Oosterveer et al. Differences in characteristics and risk of cardiovascular disease in familial hypercholesterolemia patients with and without tendon xanthomas: a systematic review and meta-analysis. *Atherosclerosis*, **207**: 311–7, 2009. (see p. 152)

[4] Lucia Pietroleonardo and Thomas Ruzicka. Skin manifestations in familial heterozygous hypercholesterolemia. *Acta Dermatovenerol Alp Panonica Adriat*, **18**: 183–7, 2009. (see p. 152)

[5] Duck-Chul Lee et al. Changes in fitness and fatness on the development of cardiovascular disease risk factors hypertension, metabolic syndrome, and hypercholesterolemia. *J Am Coll Cardiol*, **59**: 665–72, 2012. (see p. 152)

[6] Sekar Kathiresan and Deepak Srivastava. Genetics of human cardiovascular disease. *Cell*, **148**: 1242–57, 2012. (see p. 152)

[7] Daniel Zambón et al. Higher incidence of mild cognitive impairment in familial hypercholesterolemia. *Am J Med*, **123**: 267–74, 2010. (see p. 152)

[8] Ornella Guardamagna et al. The type of LDLR gene mutation predicts cardiovascular risk in children with familial hypercholesterolemia. *J Pediatr*, **155**: 199–204.e2, 2009. (see p. 152)

[9] Carolyn M Hutter, Melissa A Austin, and Steve E Humphries. Familial hypercholesterolemia, peripheral arterial disease, and stroke: a HuGE minireview. *Am J Epidemiol*, **160**: 430–5, 2004. (see p. 152)

[10] Peter van Vliet. Cholesterol and late-life cognitive decline. *J Alzheimers Dis*, **30 Suppl 2**: S147–62, 2012. (see p. 152)

[11] Matthias Orth and Stefano Bellosta. Cholesterol: its regulation and role in central nervous system disorders. *Cholesterol*, **2012**: 292598, 2012. (see p. 152)

[12] Carlos Ramírez et al. ApoB100/LDLR-/- hypercholesterolaemic mice as a model for mild cognitive impairment and neuronal damage. *PLoS ONE*, **6**: e22712, 2011. (see p. 152)

[13] J Goldstein and M Brown. The LDL receptor. *Arterioscler Thromb Vasc Biol*, **29**: 431–438, 2009. (see p. 152)

[14] Anne K Soutar and Rossi P Naoumova. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nature clinical practice Cardiovascular medicine*, **4**: 214–25, 2007. (see pp. 152, 155)

[15] Melissa A Austin et al. Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. *Am J Epidemiol*, **160**: 407–20, 2004. (see p. 152)

[16] A F Vuorio et al. Familial hypercholesterolaemia in Finland: common, rare and mild mutations of the LDL receptor and their clinical consequences. Finnish FH-group. *Ann Med*, **33**: 410–21, 2001. (see p. 152)

[17] V Gudnason et al. Common founder mutation in the LDL receptor gene causing familial hypercholesterolaemia in the Icelandic population. *Hum Mutat*, **10**: 36–44, 1997. (see p. 152)

[18] D C Rubinsztein, D R van der Westhuyzen, and G A Coetzee. Monogenic primary hypercholesterolaemia in South Africa. *S Afr Med J*, **84**: 339–44, 1994. (see p. 152)

[19] M N Slimane et al. Phenotypic expression of familial hypercholesterolaemia in central and southern Tunisia. *Atherosclerosis*, **104**: 153–8, 1993. (see p. 152)

[20] S Moorjani et al. Homozygous familial hypercholesterolemia among French Canadians in Québec Province. *Arteriosclerosis*, **9**: 211–6, 1989. (see p. 152)

[21] H C Seftel et al. Prevalence of familial hypercholesterolemia in Johannesburg Jews. *Am J Med Genet*, **34**: 545–7, 1989. (see p. 152)

[22] H C Seftel et al. A host of hypercholesterolaemic homozygotes in South Africa. *Br Med J*, **281**: 633–6, 1980. (see p. 152)

[23] A David Marais. Familial hypercholesterolaemia. *Clin Biochem Rev*, **25**: 49–68, 2004. (see p. 152)

[24] Ana-Barbara Garcia-Garcia et al. Reduced penetrance of autosomal dominant hypercholesterolemia in a high percentage of families: importance of genetic testing in the entire family. *Atherosclerosis*, **218**: 423–30, 2011. (see p. 152)

[25] C R Pullinger et al. Familial ligand-defective apolipoprotein B. Identification of a new mutation that decreases LDL receptor binding affinity. *J Clin Invest*, **95**: 1225–34, 1995. (see pp. 152, 157, 186)

[26] N B Myant. Familial defective apolipoprotein B-100: a review, including some comparisons with familial hypercholesterolaemia. *Atherosclerosis*, **104**: 1–18, 1993. (see pp. 152, 157, 186)

[27] T L Innerarity et al. Familial defective apolipoprotein B-100: low density lipoproteins with abnormal receptor binding. *Proc Natl Acad Sci USA*, **84**: 6919–23, 1987. (see pp. 152, 157, 186)

[28] Delphine Allard et al. Novel mutations of the PCSK9 gene cause variable phenotype of autosomal dominant hypercholesterolemia. *Hum Mutat*, **26**: 497, 2005. (see pp. 152, 157, 186)

[29] Jonathan Cohen et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet*, **37**: 161–5, 2005. (see pp. 152, 157, 186)

[30] T P Leren. Mutations in the PCSK9 gene in Norwegian subjects with autosomal dominant hypercholesterolemia. *Clin Genet*, **65**: 419–22, 2004. (see pp. 152, 157, 186)

[31] Marianne Abifadel et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet*, **34**: 154–6, 2003. (see pp. 152, 157, 186)

[32] Ludovic Villéger et al. The UMD-LDLR database: additions to the software and 490 new entries to the database. *Hum Mutat*, **20**: 81–7, 2002. (see pp. 152, 157)

[33] K E Heath et al. Low-density lipoprotein receptor gene (LDLR) world-wide website in familial hypercholesterolaemia: update, new features and mutation analysis. *Atherosclerosis*, **154**: 243–6, 2001. (see pp. 152, 155, 157, 176, 180, 181)

[34] Akl C Fahed and Georges M Nemer. Familial hypercholesterolemia: the lipids or the genes? *Nutr Metab (Lond)*, **8**: 23, 2011. (see pp. 152, 153, 157)

[35] M Vrablík, R Ceska, and A Horínek. Major apolipoprotein B-100 mutations in lipoprotein metabolism and atherosclerosis. *Physiol Res*, **50**: 337–43, 2001. (see p. 152)

[36] Brown M. S. Goldstein J. L. "Familial Hypercholesterolemia" in: *The Metabolic and Molecular Bases of Inherited Disease* ed. by Sly W. S. Valle D. Scriver C. R. Beaudet A. C. 7th Edition. New York, USA: McGraw Hill Book Co., 1995. 1981–2030 (see p. 152)

[37] Børge G Nordestgaard et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: Consensus Statement of the European Atherosclerosis Society. *Eur Heart J*, 2013. (see p. 153)

[38] H A Neil et al. Extent of underdiagnosis of familial hypercholesterolaemia in routine practice: prospective registry study. *BMJ*, **321**: 148, 2000. (see p. 153)

[39] J Yochem et al. A gp330/megalin-related protein is required in the major epidermis of Caenorhabditis elegans for completion of molting. *Development*, **126**: 597–606, 1999. (see p. 153)

[40] J Yochem and I Greenwald. A gene for a low density lipoprotein receptor-related protein in the nematode Caenorhabditis elegans. *Proc Natl Acad Sci USA*, **90**: 4572–6, 1993. (see p. 153)

[41] M D Adams et al. The genome sequence of Drosophila melanogaster. *Science*, **287**: 2185–95, 2000. (see p. 153)

[42] C P Schonbaum, S Lee, and A P Mahowald. The Drosophila yolkless gene encodes a vitellogenin receptor belonging to the low density lipoprotein receptor superfamily. *Proc Natl Acad Sci USA*, **92**: 1485–9, 1995. (see p. 153)

[43] Gwang-Woong Go and Arya Mani. Low-density lipoprotein receptor (LDLR) family orchestrates cholesterol homeostasis. *Yale J Biol Med*, **85**: 19–28, 2012. (see p. 153)

[44] Anders Nykjaer and Thomas E Willnow. The low-density lipoprotein receptor gene family: a cellular Swiss army knife? *Trends Cell Biol*, **12**: 273–80, 2002. (see pp. 153, 154)

[45] Joachim Herz and Hans H Bock. Lipoprotein receptors in the nervous system. *Annu Rev Biochem*, **71**: 405–34, 2002. (see pp. 153, 154)

[46] B W Howell and J Herz. The LDL receptor gene family: signaling functions during development. *Curr Opin Neurobiol*, **11**: 74–81, 2001. (see pp. 153, 154)

[47] C K Garcia et al. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science*, **292**: 1394–8, 2001. (see p. 153)

[48] H Barnes et al. Tyrosine-phosphorylated low density lipoprotein receptor-related protein 1 (Lrp1) associates with the adaptor protein SHC in SRC-transformed cells. *J Biol Chem*, **276**: 19119–25, 2001. (see p. 153)

[49] Y Li et al. The YXXL motif, but not the two NPXY motifs, serves as the dominant endocytosis signal for low density lipoprotein receptor-related protein. *J Biol Chem*, **275**: 17187–94, 2000. (see p. 153)

[50] M Gotthardt et al. Interactions of the low density lipoprotein receptor gene family with cytosolic adaptor and scaffold proteins suggest diverse biological functions in cellular communication and signal transduction. *J Biol Chem*, **275**: 25616–24, 2000. (see p. 153)

[51] M Trommsdorff et al. Interaction of cytosolic adaptor proteins with neuronal apolipoprotein E receptors and the amyloid precursor protein. *J Biol Chem*, **273**: 33556–60, 1998. (see p. 153)

[52] L Goretzki and B M Mueller. Low-density-lipoprotein-receptor-related protein (LRP) interacts with a GTP-binding protein. *Biochem J*, **336 (Pt 2)**: 381–6, 1998. (see p. 153)

[53] D J Owen and P R Evans. A structural explanation for the recognition of tyrosine-based endocytotic signals. *Science*, **282**: 1327–32, 1998. (see p. 153)

[54] J Herz and D K Strickland. LRP: a multifunctional scavenger and signaling receptor. *J Clin Invest*, **108**: 779–84, 2001. (see pp. 153, 154)

[55] Stephen C Blacklow. Versatility in ligand recognition by LDL receptor family proteins: advances and frontiers. *Current opinion in structural biology*, **17**: 419–26, 2007. (see pp. 154–156, 160, 167, 174, 181)

[56] Miklos Guttman and Elizabeth A Komives. The structure, dynamics, and binding of the LA45 module pair of the low-density lipoprotein receptor suggest an important role for LA4 in ligand release. *Biochemistry*, **50**: 11001–8, 2011. (see pp. 154, 156, 180)

[57] Carl Fisher, Natalia Beglova, and Stephen C Blacklow. Structure of an LDLR-RAP complex reveals a general mode for ligand recognition by lipoprotein receptors. *Mol Cell*, **22**: 277–83, 2006. (see pp. 154, 156, 178, 180)

[58] Natalia Beglova et al. Cooperation between fixed and low pH-inducible interfaces controls lipoprotein release by the LDL receptor. *Mol Cell*, **16**: 281–92, 2004. (see pp. 154, 156, 180, 184)

[59] Gabby Rudenko et al. Structure of the LDL receptor extracellular domain at endosomal pH. *Science*, **298**: 2353–8, 2002. (see pp. 154, 178, 179, 184–186)

[60] H Jeon et al. Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair. *Nat Struct Biol*, **8**: 499–504, 2001. (see p. 154)

[61] D Fass et al. Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. *Nature*, **388**: 691–3, 1997. (see pp. 154, 160, 167)

[62] Gang Ren et al. Model of human low-density lipoprotein and bound receptor based on cryoEM. *Proc Natl Acad Sci USA*, **107**: 1059–64, 2010. (see p. 154)

[63] D W Russell, M S Brown, and J L Goldstein. Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. *J Biol Chem*, **264**: 21682–8, 1989. (see p. 154)

[64] Carl Fisher, Dunia Abdul-Aziz, and Stephen C Blacklow. A two-module region of the low-density lipoprotein receptor sufficient for formation of complexes with apolipoprotein E ligands. *Biochemistry*, **43**: 1037–44, 2004. (see pp. 154, 155, 174, 181)

[65] J Boren et al. Identification of the low density lipoprotein receptor-binding site in apolipoprotein B100 and the modulation of its binding activity by the carboxyl terminus in familial defective apo-B100. *J Clin Invest*, **101**: 1084–93, 1998. (see pp. 154, 155)

[66] Sara Cheek, S Sri Krishna, and Nick V Grishin. Structural classification of small, disulfide-rich protein domains. *Journal of Molecular Biology*, **359**: 215–37, 2006. (see pp. 154, 160, 167)

[67] C E Dann et al. Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. *Nature*, **412**: 86–90, 2001. (see p. 154)

[68] Dunia Abdul-Aziz et al. Folding and binding integrity of variants of a prototype ligand-binding module from the LDL receptor possessing multiple alanine substitutions. *Biochemistry*, **44**: 5075–85, 2005. (see pp. 154, 156, 160, 167, 180)

[69] Xabier Arias-Moreno et al. Scrambled isomers as key intermediates in the oxidative folding of ligand binding module 5 of the low density lipoprotein receptor. *J Biol Chem*, **283**: 13627–37, 2008. (see pp. 154, 156, 160, 167, 180, 186)

[70] Xabier Arias-Moreno et al. Mechanism of low density lipoprotein (LDL) release in the endosome: implications of the stability and Ca2+ affinity of the fifth binding module of the LDL receptor. *J Biol Chem*, **283**: 22670–9, 2008. (see pp. 154, 156, 160, 167, 180)

[71] Natalia Beglova and Stephen C Blacklow. The LDL receptor: how acid pulls the trigger. *Trends Biochem Sci*, **30**: 309–17, 2005. (see pp. 154, 156, 160, 167)

[72] T Yamamoto et al. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell*, **39**: 27–38, 1984. (see p. 154)

[73] Peter D Stenson et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*, **21**: 577–81, 2003. (see pp. 155, 176, 180, 181)

[74] M A Lehrman et al. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science*, **227**: 140–6, 1985. (see pp. 155, 180)

[75] M S Brown and J L Goldstein. Familial hypercholesterolemia: defective binding of lipoproteins to cultured fibroblasts associated with impaired regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity. *Proc Natl Acad Sci USA*, **71**: 788–92, 1974. (see pp. 155, 180)

[76] J L Goldstein and M S Brown. Familial hypercholesterolemia: identification of a defect in the regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity associated with overproduction of cholesterol. *Proc Natl Acad Sci USA*, **70**: 2804–8, 1973. (see pp. 155, 180)

[77] UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, **41**: D43–7, 2013. (see pp. 155, 180, 188)

[78] Thomas A Peterson et al. DMDM: domain mapping of disease mutations. *Bioinformatics*, **26**: 2458–9, 2010. (see pp. 155, 180)

[79] S T Sherry et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**: 308–11, 2001. (see pp. 155, 180)

[80] A Taylor et al. Mutation screening in patients for familial hypercholesterolaemia (ADH). *Clin Genet*, **77**: 97–9, 2010. (see p. 155)

[81] A Taylor et al. Multiplex ligation-dependent probe amplification analysis to screen for deletions and duplications of the LDLR gene in patients with familial hypercholesterolaemia. *Clin Genet*, **76**: 69–75, 2009. (see p. 155)

[82] Steve E Humphries et al. What is the clinical utility of DNA testing in patients with familial hypercholesterolaemia? *Curr Opin Lipidol*, **19**: 362–8, 2008. (see pp. 155, 157)

[83] M Varret et al. Genetic heterogeneity of autosomal dominant hypercholesterolemia. *Clin Genet*, **73**: 1–13, 2008. (see pp. 155, 157, 180)

[84] A Taylor et al. Multiplex ARMS analysis to detect 13 common mutations in familial hypercholesterolaemia. *Clin Genet*, **71**: 561–8, 2007. (see p. 155)

[85] Emily S van Aalst-Cohen et al. Diagnosing familial hypercholesterolaemia: the relevance of genetic testing. *Eur Heart J*, **27**: 2240–6, 2006. (see pp. 155, 180)

[86] Olaf A Bodamer et al. Use of denaturing HPLC to provide efficient detection of mutations causing familial hypercholesterolemia. *Clin Chem*, **48**: 1913–8, 2002. (see p. 155)

[87] S W Fouchier et al. The molecular basis of familial hypercholesterolemia in The Netherlands. *Hum Genet*, **109**: 602–15, 2001. (see p. 155)

[88] K E Heath et al. A molecular genetic service for diagnosing individuals with familial hypercholesterolaemia (FH) in the United Kingdom. *Eur J Hum Genet*, **9**: 244–52, 2001. (see p. 155)

[89] A Taylor et al. Mutation detection rate and spectrum in familial hypercholesterolaemia patients in the UK pilot cascade project. *Clin Genet*, **77**: 572–80, 2010. (see pp. 155, 180)

[90] A M Medeiros et al. Update of the Portuguese Familial Hypercholesterolaemia Study. *Atherosclerosis*, **212**: 553–8, 2010. (see pp. 155, 180)

[91] Adrian David Marais, Jean Catherine Firth, and Dirk Jacobus Blom. Familial hypercholesterolemia in South Africa. *Semin Vasc Med*, **4**: 93–5, 2004. (see p. 155)

[92] Trond P Leren et al. Application of molecular genetics for diagnosing familial hypercholesterolemia in Norway: results from a family-based screening program. *Semin Vasc Med*, **4**: 75–85, 2004. (see p. 155)

[93] Ian Hamilton-Craig. Case-finding for familial hypercholesterolemia in the Asia-Pacific region. *Semin Vasc Med*, **4**: 87–92, 2004. (see p. 155)

[94] Joep C Defesche et al. Advanced method for the identification of patients with inherited hypercholesterolemia. *Semin Vasc Med*, **4**: 59–65, 2004. (see p. 155)

[95] M A Umans-Eckenhausen et al. Review of first 5 years of screening for familial hypercholesterolaemia in the Netherlands. *Lancet*, **357**: 165–8, 2001. (see p. 155)

[96] Renée M Ned and Eric J G Sijbrands. Cascade Screening for Familial Hypercholesterolemia (FH). *PLoS Curr*, **3**: RRN1238, 2011. (see pp. 156, 157)

[97] Zhenze Zhao and Peter Michaely. Role of an intramolecular contact on lipoprotein uptake by the LDL receptor. *Biochim Biophys Acta*, **1811**: 397–408, 2011. (see pp. 156, 180, 184)

[98] Xabier Arias-Moreno et al. Thermodynamics of protein-cation interaction: Ca(+2) and Mg(+2) binding to the fifth binding module of the LDL receptor. *Proteins*, **78**: 950–61, 2010. (see pp. 156, 180)

[99] Sha Huang et al. Mechanism of LDL binding and release probed by structure-based mutagenesis of the LDL receptor. *J Lipid Res*, **51**: 297–308, 2010. (see pp. 156, 180)

[100] Nutjaree Jeenduang et al. Molecular modeling of D151Y and M391T mutations in the LDL receptor. *Biochem Biophys Res Commun*, **377**: 355–60, 2008. (see p. 156)

[101] S Cuesta-López, F Falo, and J Sancho. Computational diagnosis of protein conformational diseases: short molecular dynamics simulations reveal a fast unfolding of r-LDL mutants that cause familial hypercholesterolemia. *Proteins*, **66**: 87–95, 2007. (see pp. 156, 158, 161, 174, 181, 193)

[102] Misha Soskine and Dan S Tawfik. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*, **11**: 572–82, 2010. (see p. 157)

[103] Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*, **19**: 596–604, 2009. (see p. 157)

[104] Nobuhiko Tokuriki et al. How protein stability and new functions trade off. *PLoS Comput Biol*, **4**: e1000002, 2008. (see pp. 157, 158, 180)

[105] Z Wang and J Moult. SNPs, protein structure, and disease. *Hum Mutat*, **17**: 263–70, 2001. (see p. 157)

[106] Lucía Conde et al. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Research*, **32**: W242–8, 2004. (see p. 158)

[107] Yongwook Choi et al. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, **7**: e46688, 2012. (see pp. 158, 180)

[108] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, **4**: 1073–81, 2009. (see pp. 158, 160, 180, 181, 187)

[109] Carles Ferrer-Costa, Modesto Orozco, and Xavier De La Cruz. Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. *Proteins*, **61**: 878–87, 2005. (see pp. 158, 160, 180, 187)

[110] C Ferrer-Costa, M Orozco, and X De La Cruz. Sequence-based prediction of pathological mutations. *Proteins*, **57**: 811–819, 2004. (see p. 158)

[111] P C Ng and S Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, **11**: 863–74, 2001. (see p. 158)

[112] S Sunyaev et al. Prediction of deleterious human alleles. *Hum Mol Genet*, **10**: 591–7, 2001. (see p. 158)

[113] Lei Bao and Yan Cui. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**: 2185–90, 2005. (see pp. 158, 180)

[114] Nathan O Stitziel et al. Structural location of disease-associated single-nucleotide polymorphisms. *Journal of Molecular Biology*, **327**: 1021–30, 2003. (see pp. 158, 180)

[115] Carles Ferrer-Costa, Modesto Orozco, and Xavier De La Cruz. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of Molecular Biology*, **315**: 771–86, 2002. (see pp. 158, 160, 187)

[116] Christopher T Saunders and David Baker. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology*, **322**: 891–901, 2002. (see p. 158)

[117] D Chasman and R M Adams. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, **307**: 683–706, 2001. (see p. 158)

[118] S Sunyaev, V Ramensky, and P Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, **16**: 198–200, 2000. (see p. 158)

[119] Abel González-Pérez and Nuria López-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, **88**: 440–9, 2011. (see pp. 158, 160, 181, 187)

[120] Yana Bromberg and Burkhard Rost. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, **35**: 3823–35, 2007. (see p. 158)

[121] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, **30**: 3894–900, 2002. (see p. 158)

[122] Man Hoang Viet et al. Effect of the Tottori Familial Disease Mutation (D7N) on the Monomers and Dimers of A40 and A42. *ACS Chem Neurosci*, 2013. (see pp. 158, 181)

[123] Yu-Shan Lin and Vijay S Pande. Effects of familial mutations on the monomer structure of A. *Biophys J*, **103**: L47–9, 2012. (see pp. 158, 181)

[124] Carles Ferrer-Costa et al. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, **21**: 3176–8, 2005. (see pp. 160, 181, 187)

[125] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**: 3812–4, 2003. (see pp. 160, 187)

[126] Ivan A Adzhubei et al. A method and server for predicting damaging missense mutations. *Nat Methods*, **7**: 248–9, 2010. (see pp. 160, 181, 187)

[127] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, **39**: e118, 2011. (see pp. 160, 181, 187)

[128] Boris Reva, Yevgeniy Antipin, and Chris Sander. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol*, **8**: R232, 2007. (see pp. 160, 187)

[129] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**: 778–95, 2009. (see pp. 160, 188)

[130] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**: 889–95, 2010. (see pp. 161, 181)

[131] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**: 702–10, 2004. (see pp. 161, 181)

[132] Charles A Laughton, Modesto Orozco, and Wim Vranken. COCO: a simple tool to enrich the representation of conformational variability in NMR structures. *Proteins*, **75**: 206–16, 2009. (see pp. 161, 163, 164, 182, 189, 190)

[133] Gia G Maisuradze, Adam Liwo, and Harold A Scheraga. Principal component analysis for protein folding dynamics. *J Mol Biol*, **385**: 312–29, 2009. (see p. 161)

[134] A Palazoglu et al. Probing Protein Folding Dynamics Using Multivariate Statistical Techniques. *Advanced Control of Chemical Processes*, **7**: 171–176, 2009. (see p. 161)

[135] Lee-Wei Yang et al. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics*, **25**: 606–14, 2009. (see pp. 161, 163)

[136] S Stein et al. Principal Components Analysis: A Review of its Application on Molecular Dynamics Data. *Annual Reports in Computational Chemistry*, **2**: 233–261, 2006. (see p. 161)

[137] Aruna Rajan, Peter L Freddolino, and Klaus Schulten. Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PLoS ONE*, **5**: e9890, 2010. (see p. 163)

[138] T Meyer, C Ferrer-Costa, and A Pérez… Essential dynamics: a tool for efficient trajectory compression and management. *J. Chem. Theory …*, 2006. (see p. 163)

[139] H J Berendsen and S Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, **10**: 165–9, 2000. (see pp. 163, 182, 189)

[140] D Van Aalten, B De Groot, and J Findlay… A comparison of techniques for calculating protein essential dynamics. *Journal of …*, 1997. (see pp. 163, 182, 189)

[141] A Amadei et al. An efficient method for sampling the essential subspace of proteins. *J Biomol Struct Dyn*, **13**: 615–25, 1996. (see pp. 163, 182, 189)

[142] A Amadei, A B Linssen, and H J Berendsen. Essential dynamics of proteins. *Proteins*, **17**: 412–25, 1993. (see pp. 163, 182, 189)

[143] Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol*, **1084**: 193–226, 2014. (see pp. 169, 191)

[144] Patrick G Blachly et al. Utilizing a Dynamical Description of IspH to Aid in the Development of Novel Antimicrobial Drugs. *PLoS Comput Biol*, **9**: e1003395, 2013. (see pp. 169, 193)

[145] Christopher L McClendon et al. Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *J Chem Theory Comput*, **8**: 2115–2126, 2012. (see p. 169)

[146] Christopher L McClendon et al. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J Chem Theory Comput*, **5**: 2486–2502, 2009. (see p. 169)

[147] S Kullback and R A Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**: 79–86, 1951. (see pp. 169, 172, 194)

[148] J Briët and P Harremoës. Properties of classical and quantum Jensen-Shannon divergence. *Physical review A*, 2009. (see pp. 169, 172, 194)

[149] A Majtey, P Lamberti, and D Prato. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Physical review A*, **72**: 052310, 2005. (see pp. 169, 172, 194)

[150] F Topsøe. Jensen-Shannon Divergence and norm-based measures of Discrimination and Variation. *Technical Report*, 2003. (see pp. 169, 172, 194)

[151] J Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, **37**: 145–151, 1991. (see pp. 169, 172, 194)

[152] Rao C. R. "Differential Metrics in Probability Spaces" in: *Differential geometry in statistical inference* ed. by Shanti S. Gupta. vol. 10 Hayward, California, USA: Institute of Mathematical Statistics, 1987. 217–238 (see pp. 169, 172, 194)

[153] D Endres and J Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, **49**: 1858 –1860, 2003. (see pp. 170, 172, 183, 193, 194)

[154] P Absil, A Edelman, and P Koev. On the largest principal angle between random subspaces. *Linear Algebra and its Applications*, 2006. (see p. 174)

[155] H Gunawan, O Neswan, and W Setya-Budhi. A formula for angles between subspaces of inner product spaces. *Contributions to Algebra …*, 2005. (see p. 174)

[156] A Amadei, M A Ceruso, and A Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins*, **36**: 419–24, 1999. (see p. 174)

[157] J Miao and A Ben-Israel. On principal angles between subspaces in Ri sup¿ ni/sup¿. *Linear Algebra and its Applications*, 1992. (see p. 174)

[158] Prasanta Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, **2**: 49–55, 1936. (see pp. 174, 192)

[159] Juan Martínez-Oliván et al. LDL receptor/lipoprotein recognition: endosomal weakening of apo B and apo E binding to the convex face of the LR5 repeat. *FEBS J*, 2014. (see pp. 178, 179, 184, 185)

[160] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, **75**: 333–66, 2006. (see p. 180)

[161] Niels Gregersen, Lars Bolund, and Peter Bross. Protein misfolding, aggregation, and degradation in disease. *Mol Biotechnol*, **31**: 141–50, 2005. (see p. 180)

[162] Damian C Crowther. Familial conformational diseases and dementias. *Hum Mutat*, **20**: 1–14, 2002. (see p. 180)

[163] R R Kopito and D Ron. Conformational disease. *Nat Cell Biol*, **2**: E207–9, 2000. (see p. 180)

[164] R W Carrell and D A Lomas. Conformational disease. *Lancet*, **350**: 134–8, 1997. (see p. 180)

[165] Zhenze Zhao and Peter Michaely. The epidermal growth factor homology domain of the LDL receptor drives lipoprotein release through an allosteric mechanism involving H190, H562, and H586. *J Biol Chem*, **283**: 26528–37, 2008. (see p. 184)

[166] Natalia Beglova, Christopher L North, and Stephen C Blacklow. Backbone Dynamics of a Module Pair from the Ligand-Binding Domain of the LDL Receptor. *Biochemistry*, **40**: PMID: 11258891, 2808–2815, 2001. (see p. 186)

[167] C L North and S C Blacklow. Evidence that familial hypercholesterolemia mutations of the LDL receptor cause limited local misfolding in an LDL-A module pair. *Biochemistry*, **39**: 13127–35, 2000. (see p. 186)

[168] John E Stone et al. GPU-accelerated molecular modeling coming of age. *J Mol Graph Model*, **29**: 116–25, 2010. (see p. 187)

[169] Mark S Friedrichs et al. Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem*, **30**: 864–72, 2009. (see p. 187)

[170] James C Phillips, John E Stone, and Klaus Schulten. "Adapting a message-driven parallel application to GPU-accelerated clusters" in: *High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008. International Conference for*. IEEE 2008. 1–9 (see p. 187)

[171] K Bowers, R Dror, and D Shaw. Zonal methods for the parallel execution of range-limited N-body simulations. *Journal of Computational Physics*, **221**: 303–329, 2007. (see p. 187)

[172] Blake G Fitch et al. "Blue Matter: Strong scaling of molecular dynamics on Blue Gene/L" in: *Computational Science–ICCS 2006*. Springer, 2006. 846–854 (see p. 187)

[173] David E Shaw. A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. *J Comput Chem*, **26**: 1318–28, 2005. (see p. 187)

[174] Adam Hospital and Josep Ll Gelpi. High-throughput molecular dynamics simulations: toward a dynamic view of macromolecular structure. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **3**: 364–377, 2013. (see p. 187)

[175] Adam Hospital et al. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics*, **28**: 1278–9, 2012. (see p. 187)

[176] Tim Meyer et al. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, **18**: 1399–409, 2010. (see p. 187)

[177] Paul Flicek et al. Ensembl 2013. *Nucleic Acids Research*, **41**: D48–55, 2013. (see p. 188)

[178] W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *J Mol Graph*, **14**: 33–8, 27–8, 1996. (see pp. 189, 191)

[179] James C Phillips et al. Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**: 1781–802, 2005. (see p. 189)

[180] B R Brooks et al. CHARMM: the biomolecular simulation program. *J Comput Chem*, **30**: 1545–614, 2009. (see p. 189)

[181] W Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **34**: 827–828, 1978. (see p. 190)

[182] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0 R Foundation for Statistical Computing, Vienna, Austria, 2012. (see pp. 193, 194)

[183] Robbie P Joosten et al. A series of PDB related databases for everyday needs. *Nucleic Acids Research*, **39**: D411–9, 2011. (see p. 193)

[184] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**: 2577–637, 1983. (see p. 193)

CHAPTER 5

Conclusiones y Perspectivas

Contents

5.1 Conclusiones

Los resultados de las investigaciones que se han realizado para alcanzar los objetivos de esta Tesis Doctoral, permiten extraer las siguientes conclusiones:

A nivel de Secuencia de Proteínas: Predicción de secuencias priónicas basada en consideraciones composicionales en todos los proteomas completos anotados en bancos de datos

1. Es posible modelar las características composicionales de secuencias priónicas conocidas para identificar una gran cantidad de proteínas con posible actividad priónica en genomas completos de organismos de todas las clasificaciones taxonómicas

2. Las proteínas con posible actividad priónica se distribuyen desigualmente en diferentes clasificaciones funcionales, están vinculadas a procesos biológicos diversos, regulan diferentes funciones moleculares y se localizan en diferentes compartimentos celulares en diferentes clasificaciones taxonómicas y grupos de organismos

3. Las proteínas con posible actividad priónica regulan importantes procesos y funciones moleculares mediadas por la formación de complejas redes de interacción entre biomoléculas, como la regulación de la expresión génica, interacción entre proteínas y ADN o ARN, o la formación de biofilmes extracelulares en microorganismos

A nivel de Estructura de Proteínas: Identificación de regiones de inestabilidad conformacional en proteínas utilizando información geométrica y físico-química de interfaces enterradas

4. Es posible identificar regiones localmente inestables de las proteínas a partir del análisis de ficheros de coordenadas atómicas
5. Las regiones localmente inestables están asociadas a interfases con alta polaridad y baja densidad de empaquetamiento
6. La conservación de estas propiedades estructurales es compatible con variaciones a nivel de secuencia, lo cual puede estar relacionado con mecanismos moleculares para el desarrollo de nuevas funciones conservando una dinámica local característica

A nivel de Dinámica de Proteínas: Predicción de fenotipos patológicos asociados a enfermedades conformacionales causados por Mutaciones de Nucleótido Simple utilizando Dinámica Molecular: caso de estudio “Hypercolesterolamia Familiar”

7. Las simulaciones de Dinámica Molecular permiten analizar la inestabilización estructural causada por mutaciones en proteínas, permitiendo incluso estudiar todas las posibles mutaciones generadas por SNPs en proteínas pequeñas o módulos de plegamiento autónomo
8. De todas las posibles mutaciones del dominio LA5 del receptor de LDL, con cambio de un solo amino ácido, aproximadamente el $50\,\%$ son mutaciones que desestabilizan la estructura del dominio
9. Casi la totalidad de las mutaciones descritas como causantes de la Hypercolesterolamia Familiar corresponden a mutantes que desestabilizan la estructura del dominio o a mutaciones en residuos del sitio de interacción con otras proteínas

5.2 Perspectivas

En los proyectos que forman parte del cuerpo de esta Tesis Doctoral, nuestros resultados específicos podrían contribuir de manera significativa en la continuación de estos proyectos en nuestro propio grupo, o incluso en el desarrollo de proyectos similares en otros grupos de investigación. Las posibles perspectivas futuras de nuestro trabajo podrían incluir:

A nivel de Secuencia de Proteínas: Predicción de secuencias priónicas basado en consideraciones composicionales en todos los proteomas completos anotados en bancos de datos

- El desarrollo de nuestra base de datos con predicciones de proteínas con posible actividad priónica en los genomas completos de todos los organismos anotados en bancos de datos de secuencias biológicas, así como la libre distribución de nuestro algoritmo, podría contribuir a la realización de estudios a nivel genómico del papel de los priones en el funcionamiento celular y en el desarrollo de enfermedades causadas por estas proteínas en uno o varios organismos, o en differentes clasificaciones taxonómicas
- La utilización de nuestra bases de datos de posibles priones o de nuestro método de predicción podría ser de gran interés para experimentalistas interesados en identificar y comprobar experimentalmente la prionogenicidad de una o varias de estas proteínas en uno o varios genomas. De hecho, en estos momentos tenemos una colaboración en la cual estamos intentando comprobar *in vitro* e *in vivo* el carácter prionogénico de varias posibles proteínas priónicas no descritas hasta la fecha en bacterias y humano, y también tenemos información de proyectos en otros grupos de investigación intentando comprobar experimentalmente nuestras predicciones en varios genomas de plantas

A nivel de Estructura de Proteínas: Identificación de regiones de inestabilidad conformacional en proteínas utilizando información geométrica y físico-química de interfaces enterradas

- Una de las principales ventajas de nuestro método, a diferencia de otros disponibles, es que permite la rápida identificacion de regiones de inestabilidad local, utilizando pocos recursos computacionales. Por esta razón, podría ser fácilmente utilizado para analizar todas las proteínas de una familia de las cuales se disponga de información estructural, para estudiar cómo está codificada la información que determina la dinámica intrínseca de las proteínas
- En este sentido, tenemos un proyecto en curso para el desarrollo de una base de datos con predicciones de regiones de inestabilidad local para todas las proteínas de un solo dominio de estructura conocida. Esta base de datos permitiría hacer estudios de la relación entre inestabilidad conformacional y función, y de cómo las proteínas pertenecientes a diferentes familias funcionales o clasificaciones de plegamiento han diversificado sus funciones a lo largo de la evolución, conservando o variando simultáneamente los determinantes estructurales de la inestabilidad conformacional

A nivel de Dinámica de Proteínas: Predicción de fenotipos patológicos asociados a enfermedades conformacionales causados por Mutaciones de Nucleótido Simple utilizando Dinámica Molecular: caso de estudio “Hypercolesterolemia Familiar”

- Los resultados generados en este estudio podrían ser de gran ayuda para establecer criterios concretos que relacionen los distintos tipos de mutaciones y el contexto estructural en donde ocurren, con un fenotipo patológico en enfermedades conformacionales. Dada la abundancia de estos dominios de interacción (LA) en muchas otras proteínas y receptores de membrana, involucrados en muchos importantes procesos de regulación, transducción de señales y metabolismo de compuestos, nuestros resultados podrían servir no solo de referencia de comparación para otros dominios LA de secuencia y función diferente, sino como inspiración para desarrollar proyectos similares para el estudio de otras proteínas específicas desde el punto de vista computacional

- Dada la gran dificultad de realizar estudios mutacionales masivos para estudiar *in vitro* el efecto de todas las posibles mutaciones en este u otros dominios de interacción, nuestro método podría servir como una herramienta de ‘diagnóstico anticipado’ para identificar las posibles mutaciones que con mayor probabilidad puedan ser las causantes de la enfermedad, de entre todas las posibles. De esta manera se reduciría significativamente el número de ejemplos a ser estudiados en detalle en el laboratorio para intentar relacionar mutación y fenotipo

- El gran desarrollo de diferentes métodos computacionales para alcanzar mayores tiempos de simulación, así como la creciente disponibilidad de plataformas para realizar simulaciones de Dinámica Molecular online sin la necesidad de formación especializada, podría abrir las puertas para realizar rutinariamente estudios similares al nuestro en el futuro, para evaluar el efecto de las mutaciones en la estabilidad conformacional de una proteína, utilizando solamente información estructural, de manera no sesgada y sin asunciones evolutivas o funcionales previas

APPENDIX A

Representation of Prion Predictions in Gene Ontology Classifications

We tested the significance of the number of predictions found in all taxa according to the belonging of proteins bearing putative PrDs to different classifications in Gene Ontology –*i.e.* Molecular Function, Biological Processes and Cellular Component. We compared the abundance of predictions in a given class with the expected frequency obtained by randomly selecting a set of the same size in the proteomes over 1×10^6 randomizations. In each taxon we represent the $z - score$ for a number of representative GO terms. The GO terms description might be trimmed in some cases to fit in the chart. Given the large quantity of data included in these charts, they have to be considerably reduced to fit in the page margins of this document. In each case, we provide a link to a high quality figure for a closer look into our results (Figure A.1 (High-Res Image); Figure A.2 (High-Res Image); Figure A.3 (High-Res Image)).

FIGURE A.1: Significance Over- or Under-representation of PrD Predictions According to Gene Ontology Molecular Function Classifications

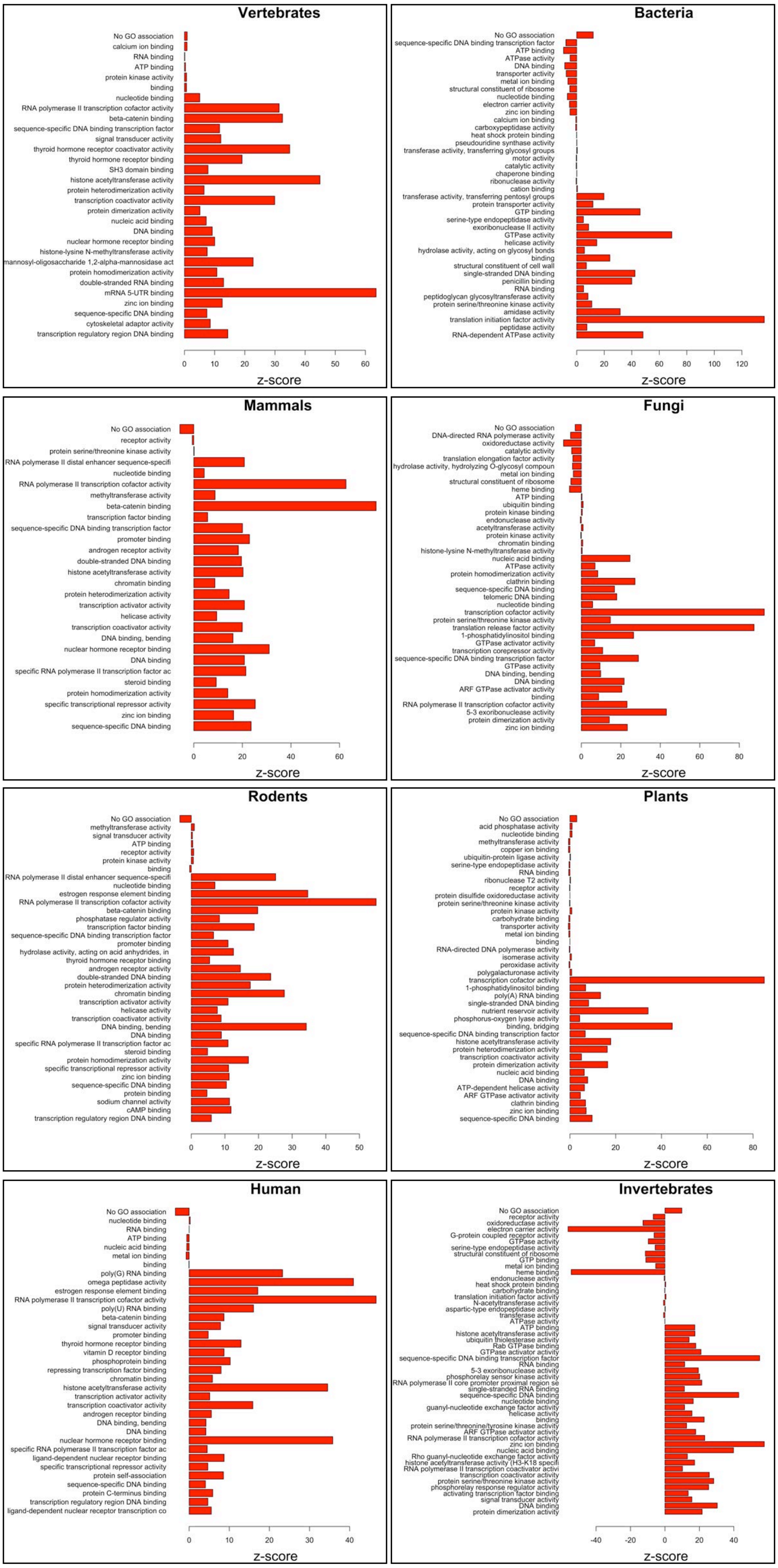


FIGURE A.2: Significance Over- or Under-representation of PrD Predictions According to Gene Ontology Biological Processes Classifications

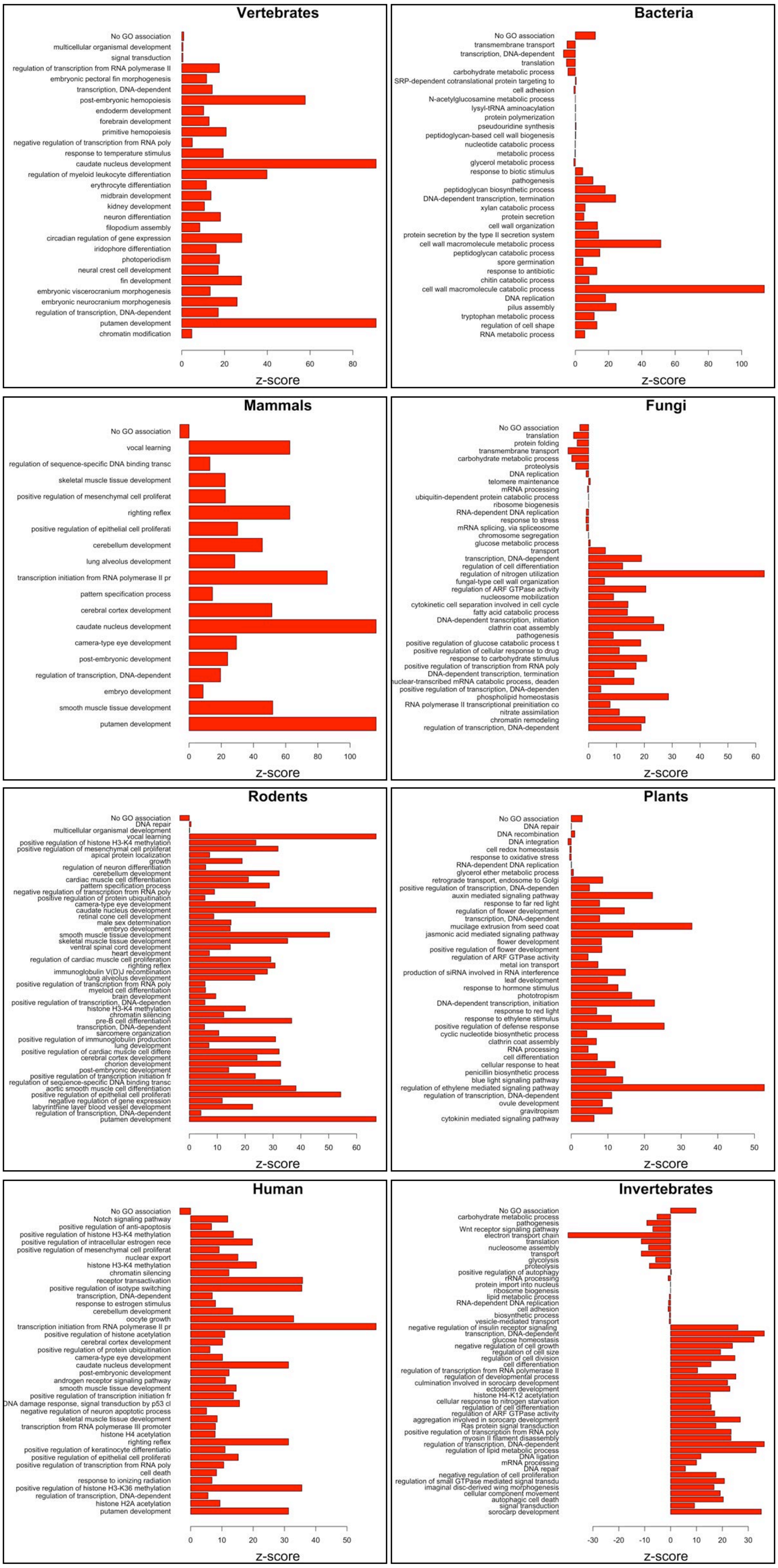
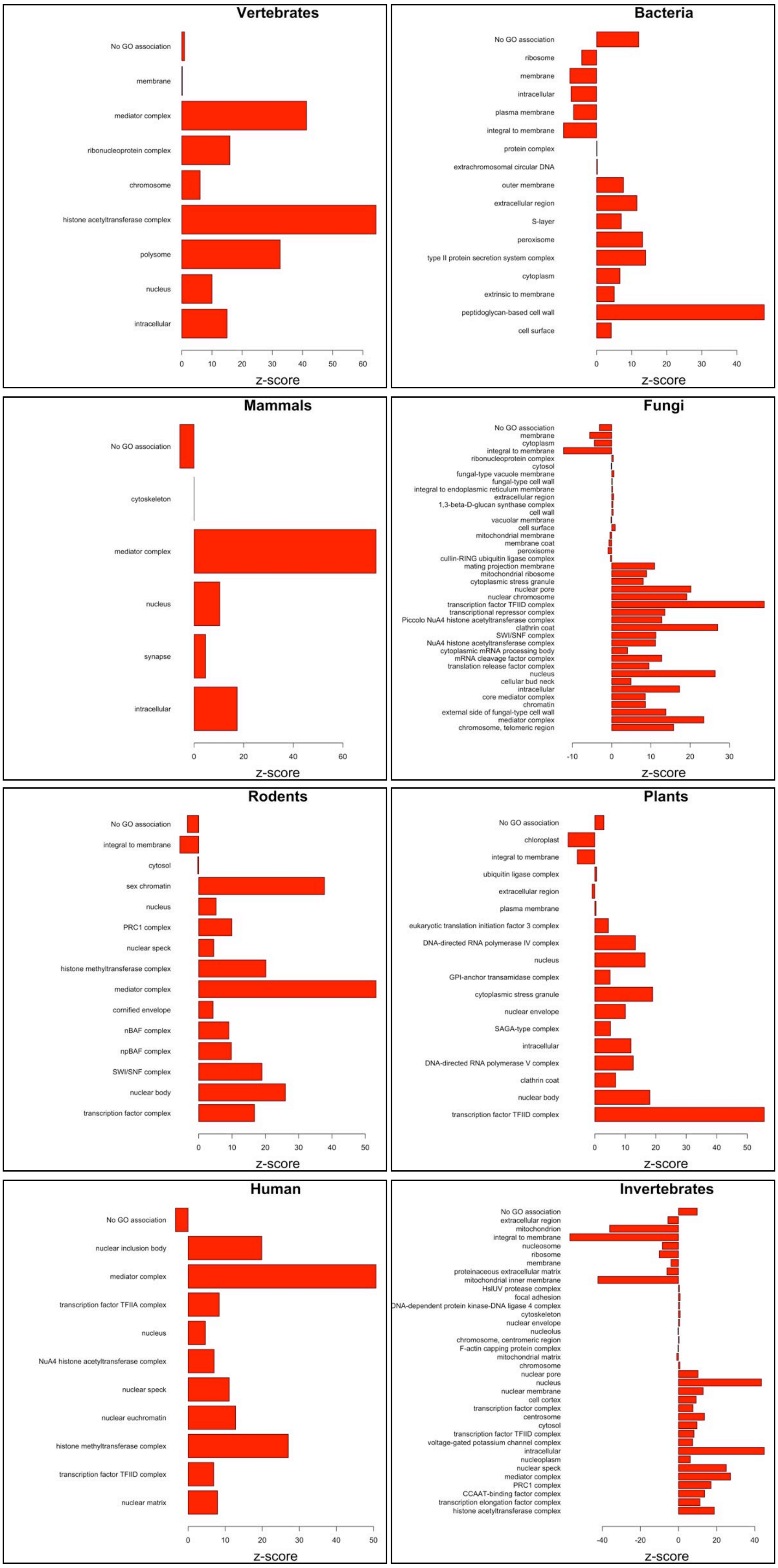


FIGURE A.3: Significance Over- or Under-representation of PrD Predictions According to Gene Ontology Cellular Component Classifications



Appendix B

Sequence Datasets

Table B.1: These 29 proteins were predicted as prions using a HMM model and were then studied experimentally to test their aggregation propensity and prionogenicity in a previous work. These sequences experimentally validated as real prions were used as the positive training set for obtaining the amino acid propensities in prion domains in our study

Gene	Sequence
CYC8_YEAST	QPNDQGNPLNTRISAQSANATASMVQQQHPAQQTPINSSATMYSNGASPQLQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAHAQAQAQAQAQAQAQAQAQAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQLQPLPRQQLQQKGVSVQMLNPQQGQPYITQPTVIQAHQLQPFSTQAMEHPQSSQLPPQQQQLQSVQHPQQLQGQPQAQAPQPLIQHNVEQN
YBM6_YEAST	SANDYYGGTAGEKSQYSRPSNPPPSSAHQNKTQERGYPPQQQQQYYQQQQQHPGYYNQQGYNQQGYNQQGYNQQGYNQQGYNQQGYNQQGHQQPVYVQQQPPQRGN
CBK1_YEAST	YNSSTNHHEGAPTSGHGYYMSQQQDQQHQQQQQYANEMNPYQQIPRPPAAGFSSNYMKEQGSHQSLQEHLQRETGNLGSGFTDVPALNYPATPPPHNNYAASNQMINTPPPSMGGLYRHNNNSQSMVQNGNGSGNAQLPQLSPGQYSIESEYNQNLNGSSSSSPFHQPQTLRSNGSYSSGLRSVKSFQRLQQEQENVQVQQQLSQAQQQNSRQQQQQLQYQQQQQQQQQQQHMQIQQQQQQQQQQQQSQSPVQSGFNNG
Q6Q7I0_YEASX	SDSNQGNNQQNYQQYSQNGNQQQGNNRYQGYQAYNAQAQPAGGYYQNYQGYSGYQQGGYQQYNPDAGYQQQYNPQGGYQQYNPQGGYQQQFNPQGGRGNYKNFNYNNNLQGYQAGFQPQSQGMSLNDFQKQQKQ

Continued on next page. . .

Table B.1: (continued)

Gene	Sequence
RNQ1_YEAST	SGSGGGSQSMGASGLAALASQFFKSGNNSQGQGQGQGQGQGQGQGQGQGSFTALASLASSFMNSNNNNQQGQNQSSGGSSFGALASMASSFMHSNNNQNSNNSQQGYNQSYQNGNQNSQGYNNQQYQGGNGGYQQQQGQSGGAFSSLASMAQSYLGGGQTQSNQQQYNQQGQNNQQQYQQQGQNYQHQQQGQQQQQGHSSSFSALASMASSYLGNNSNSNSSYGGQQQANEYGRPQQNGQQQSNEYGRPQYGGNQNSNGQHESFNFSGNFSQQNNNGNQNRY
GPR1_YEAST	NNNNNDNDNDNNNSNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNSNNIKNNVDNNNTNPADNIPTLSNEAFTPSQQFSQERVNNNADRCENSSFTNVQQHFQAQTYKQ
NEW1_YEAST	GSNNASKKSSYQQQRNWKQGGNYQQGGYQSYNSNYNNYNNYNNYNNYNNYNNYNKYNGQGYQKSTYKQSAVTPNQSG
PUF2_YEAST	NSYFNNQQVVYSGNQNQNQNGNSNGLDELNSQFDSFRIANGTNLSLPIVNLPNVSNNNNNNYNNSGYSSQMNPLSRSVSHNNNNNTNNYNNNDNDNNNNNNNNNNNNNNNNNNNNNNNSNNSNNNNNNNDTSLYRYRSYGY
NRP1_YEAST	SGNNNIAPNYRYNNNINNNNNNNINNMTNNRYNINNNINGNGNGNGNNSNNNNNNHNNNNHNNNNHHNGSINSNSNTNNNNNNNNNGNNSNNCNSNIGMGGCGSN
SWI1_YEAST	DFFNLNNNNNNNNNTTTTTTTTNNNNNTNNNNNTNNNNNNPANNTNNNNNSTGHSSNTNNNTNNNNNTNTGASGVDDFQNFFDPKPFDQNLDSNNNNNSNSNNNDNNNSNTVASSTNFTSPTAVVNNAAPANVTGGKAANFIQNQSPQFNSPYDSNNSNTNLNSLSPQAILAKNSIIDSSNLPLQAQQQLYGGNNNNNSTGIANDNVITPHFITNVQSISQNSSSSTPNTNSNSTPNANQQFLPFNNSASNNGNLTSNQLISNYAASNSMDRSSSASNEFVPNTSDNNNNSNNHNMRNNSNNKTSNNNNVTAVPAATPANTNNSTSNANTVFSERAAMFAALQQKQQQRFQALQQQQQQQQNQQQQNQQPQQQQQQQQNPKFLQSQRQQQQ
SAP30_YEAST	QGGGYASNNNGSCNNNNGSNNNNNNNNNNNNNNSNNSNNNNGPTSSGRTNGKQRLTAAQQQY
GTS1_YEAST	QQQYAMAMQQQQQQQQQLAVAQAQAQAQAQAQAQVQAQAQAQAQAQAQAQQIQMQQLQMQQQQQAPLSFQQMSQGGNLPQGYFYTQ

Continued on next page. . .

Table B.1: (continued)

Gene	Sequence
YP022_YEAST	QQAQQPQQVQQSQQPQQIQQLQQLQFPQQLRAPLQQPMLQQQMHPQQASPTFPSYDPRIRNNGQNGNQFFNLIFDNRTGVNGFEVDAANNNGNGNDQNMNINPAVQQQRYQDRNFASSSYQQPLQPLTQDQQQEQYFQQQKLAQQQQQQQQQQQQQQQLPPQN
MED3_YEAST	QAQAQAQAQAQVYAQQSTVQTPITASMAAALPNPTPSMINSVSPTNVMGTPLTNMMSPMGNAYSMGAQNQGGQVSMSQFNGSGNGSNPNTNTNSNNTPLQSQLNLNNLTPANILNMSMNNDFQQQQQQQQQQQQPQPQYNMNMGMNNMNNGG
RLM1_YEAST	GPNSAKPGNTNNPGTFPPVQTAVNNGNSSNISSTNNTNNNNNNNNNNSSNNNSNNGNDNNSNNSNNSYYSNN
LSM4_YEAST	QQINSNNNSNSNGPGHKRYYNNRDSNNNRGNYNRRNNNNGNSNRRPYSQNRQYNNSNSSNINNSINSINSNNQNMNNGLGGSVQHHFNSSSPQKVEF
YBI1_YEAST	QSSNSFQSHNAPSHQSNYHPHYNHMKYNNTGSYYYYNNNNNSSVNPHNQAGLQSINRSIPSAPYGAYNQNRANDVPYMNTQKKHHRFSANNNLNQQKYKQYPQYTSNPMVTAHLKQTYPQLYYNSNVNAHNNNNNSNNNNNNNNNSNNNNNLYNQTQFSTRYFNSNSSPSLTSSTSNSSSPYNQS
PUB1_YEAST	NNNNNNYQQRRNYGNNNRGGFRQYNSNNNNNMNMGMNMNMNMNMNNSRGMPPSSMGMPIGAMPLPSQGQPQQSQTIGLPPQVNPQ
HRP1_YEAST	QQKSSNNGGNNGGNNMNRRGGNFGNQGDFNQMYQNPMMGGYNPMMNPQAMTDYYQKMQEYYQQMQKQTGMDYTQMYQQQMQQMAMMMPGFAMPPNAMTLNQPQQDSNATQGSPAPSDSDNNKSNDVQTIGNTSNTDSGSPPLNLPNGPKGPSQYNDDHNSGYGYNRDRGDRDRNDRDRDYNHRSGGNHRRNGRGGRGGYNRRNNGYHPYNR
MRN1_YEAST	MVVSYNNNNNNNNNNNNNNISNNNNNNNMFPPFPSSDDFAMYQQSSSSGPYQETYASGPQNFGDAVYPMNGN

Continued on next page...

TABLE B.1: (continued)

Gene	Sequence
MOT3_YEAST	NADHHLQQQQQRQQHQQQHQQQHQHQHQQQHTILQNVSNTNNIGSDSLASQPFNTTTVSSNKDDVMVNSGARELPMPLHQQQYIYPYYQYTSNNSNNNNVTAGNNMSASPIVHNNSNNSNNSNISASDYTVANNSTSNNNNNNNNNNNNNNNIHPNQFTAAANMNSNAAAAAYYSFPTANMPIPQQDQQYMFNPASYISHYYSAVNSNNNGNNAANNGSNNSSHSAPAPAPGPPHHHHHHSNTHNNLNNGGAVNTNNAPQHHPTIITDQFQFQLQQNPSPNLNLNINPAQ
KSP1_YEAST	GFSNNNNKQYRQNRNYNNNNNNSNNNHGSNYNNFNNGNSYIKGWNKNFNKYRRPSSSSYTGKSPLSRYNMSYNHNNNSSINGY
NUP59_YEAST	FGIRSGNNNGGFTNLTSQAPQTTQMFQSQSQLQPQPQPQPQQQQQHLQFNGSSDASSLRFGNSLSNTVNANNYSSNIGNNSINNNNIKNGTNNISQHGQGNNPSWVNN
PDR1_YEAST	YAQPTNGQNNTQVQSNKPINAQQQIPTSVQVPFMNTNEINNNNNNNNNKNNINNINNNNSNN
URE2_YEAST	MNNNGNQVSNLSNALRQVNIGNRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNSGRNGSQNNDNENNIKNTLEQHRQQQQ
NGR1_YEAST	QQQQQQQLQQQHQQLDQEDNNGPLLIKTANNLIQNNSNMLPLNALHNAPPMHLNEGGISNMRVNDSLPSNTYNTDPTNTTVFVGGLVPKTTEFQLRSLFKPFGPILNVRIPNGKNCGFVKFEKRIDAEASIQGLQGFIVGGSPIRLSWGRPSSSNAKTNSTIMGASQYMSSNGLRAPSAASSVDNSKQILEQYAEDKRRLFLHQQQQQQQQQQDGNFSMEQMAHNNYYNYNNYDYHRNKNGSHSDLVNLQRSNVPYMQEDGALYPHQYSSPSYSLHPTGNQFSNATNNLPQFGNAMSISMQLPNGNSNKTASSMNTNPNTNMIMNSNMNMNMNVNPVPYGMGNGANMY
RBS1_YEAST	QVNKPQQQFYDSRRGRGGRRRGTNNYKDAYRGQSRRNKENGGYQSGYSSPYLVYPPPQMGGNSLPTYPLMYNPAGPAPGPAPSPMVMGNNTVFMNPYMYNMNPQGSCSFGTPIPMYPPYQYQYQYQYNTQYHSGPYSNTPSYNSNNYTRSSANKYHHFQGKNSYSG

Continued on next page. . .

TABLE B.1: (continued)

Gene	Sequence
NSP1_YEAST	NFNTPQQNKTPFSFGTANNNSNTTNQNSSTGAGAFGTGQS TFGFNNSAPNNTNNANSSITPAFGSNNTGNTAFGNSNPTSN VFGSNNSTTNTFGSNSAGTSLFGSSSAQQTKSNGTAGGNTF GSSSLFNNSTNSNTTKPAFGGLNFGGGNNTTPSSTGNANTS NNLFGATANAN
GLN3_YEAST	QYNHGSLGNSVSKSSLFPYNSSTSNSNINQPSINNNSNTNAQ SHHSFNIYKLQNNNSSSSAMNITNNNNSNNSNIQ

TABLE B.2: These proteins were predicted using a HMM model and were then studied experimentally to test their aggregation propensity and prionogenicity in a previous work. These 18 proteins resulted as negatives in all four experimental tests and in accordance were used as the negative dataset for estimating the predictive performance of our methodology

Gene	Sequence
ENT2_YEAST	NSQGTGYKQVTNEPKNNPFLSNQYTGLPSTNIVPTQTGYGF GNQPQSPPTNSPQQNPTGISYSQPQQQQQPQQQPQYMQN FQQQPQYAQNFQQQPQYTQNYQQQPQYIQPHQQQQQQ QQQQQQQQGYTPDQG
MCM1_YEAST	GNDMQRQQPQQQQPQQQQQVLNAHANSLGHLNQDQVPA GALKQEVKSQLLGGANPNQNSMIQQQQHHTQNSQPQQQQ QQQPQQQMSQQQMSQHPRPQQGIPHPQQSQPQQQQQQQ QQLQQQQQQQQQQPLTGIHQPHQQAFANAASPYLNAEQN AAYQQYFQEPQQGQY
NAB2_YEAST	NAQSLGQSDIAQQQQQQQQQQPDIAQQQPQQQPQQQP QQQPQQQPQQQPQQQPQQQPQQQPQLQPLQPQLGTQNA MQTDAPATPSPISAFSGVVNAAAPPQFAPVDNSQRFTQRGG GAVGKNRRGGRGGNRGGRNNNS
TAF12_YEAST	QESTQQQRVQQQRVQQQQQQQQQQQQQQQQQQQQQQQ QRQGQNQRKISSSNSTEIPSVTGPDALKSQQQQQN
KC11_YEAST	NKQLQMQQLQMQQLQQQQQQQQYAQKTEADMRNSQYK PKLDPTSYEAYQHQTQQKYLQEQQKRQQQQKLQEQQLQEQ QLQQQQQQQQQLRATGQPPSQPQAQTQSQQFGARYQPQQ Q

Continued on next page...

TABLE B.2: (continued)

Gene	Sequence
MED2_YEAST	NNINNNINSTKNGKDNNNESNKNNNGDEKNKNNNEDNENNNNSSEKNNNNNNNNNNNNDDNGNNNNNNSGDNNNTTNNDSNNKNNS
AKL1_YEAST	QQQGQRYQQAQNQTGTQGNTFPDESQYQSRVEQQQQQDQPKGPANYSQRNFYTGRDRSNKPMQLGGTIAGDSGNRRVNFQNISQNYATNSQSGYLPSQNSPAIPMVRPVISMNQQQAQQIQAQQLQAQQMQAKQQMQAKQQMQVQQQLQVQQQMQIQNANNNG
PUF4_YEAST	QNHMPLMNSANNKHHGRNNNSMSSHNDNDNIGNSNYNNKDTGRSNVGKMKNMKNSYHGYYNNNNNNNNNNNNNNNSNATNSNS
PCF11_YEAST	QVQMQLRQVFSQDQQVLQERMRYHELQQQQQQQQQQQQQQQQQQQQYHETKDMVGSYTQNSNSAIPLFGNNSDTTNQQNS
SKG6_YEAST	QPLNYQDQYQQQEQSPVYNGHTQYPGNGYSGNPQQQGYTAQFVQNPQWYGVPTPQQQQHNHPQ
EPL1_YEAST	IQHLQQQQQQQQQQQQQAQQQKQKSQNNNSNSSNSLKKLNDSLINSEAKQNSSITQKNSS
SNF2_YEAST	QFAAKQRQELQMQRQQQGISGSQQNIVPNSSDQAELPNNASSHISASASPHLAPNMQLNGNETFSTSAHQSPIMQTQMPLNSNGGNNMLPQRQSSVGSLNATNFSPTPANNGENAAEKPDNSNHNNLNLNNSELQPQNRSLQEHNIQDSNVMPGSQINSPMPQQAQMQQAQFQAQQAQQAQQAQQAQQAQARLQQG
SCD6_YEAST	GLGRGRGNYRGNRGNRGRGGQRGNYQNRNNYQNDSGAYQNQNDSYSRPANQFSQPPSNVEF
YAK1_YEAST	MNSSNNNDSSSSNSNMNNSLSPTLVTHSDASMGSGRASPDNSHMGRGIWNPSYVNQGSQRSPQQQHQNHHQQQQQQQQQQQNSQ
YL177_YEAST	NNSSQKYYPQKQQQQQQQQQQQQQQSIFDPGRRSSYISDALIHGNAATQQPQYSQPVYINNNPSLQVPYTAPSEYTQQQQYSSPFNARRNTQ
CAF40_YEAST	MFSAQKPIYGNGAGVNMGGGGPSTNNPGSMSMPGVPTSMGPGMNQQIPSGGPMLMGNTPNNNNSNENGENNGNNGNNGGNDANATRNNPNMVNNRG

Continued on next page. . .

TABLE B.2: (continued)

Gene	Sequence
NRD1_YEAST	QQYVQPMMQQPYGYAPNQPLPSQGPAAAAPPVPQQQFDPT AQLNSLMNMLNQQQQQQQS
PDC2_YEAST	NNQNHLSMSQASHNPDYNSNHSNNAIENTNNRGSNNNNN NNGSSNNINDNDSSVKYLQQNTVDNSTKTGNPGQPN

Appendix C

Scripts

Code C.1: This *ad hoc* script comes with a man page (run [./prion_parse_proteome.pl man] in a UNIX/Linux console) which explains the functionality and parameters needed for running in a Linux environment and the required libraries dependencies. It is designed to read genomes in a Swissprot format and to run in a multicore environment to speed up the prediction in large protein sequence sets, as those distributed in Uniprot. We only show some sections of the 500 lines of the original script. For a complete version please download it at the following address: prion_parse_proteome.pl

```
my @seq_ids = keys %sequences;
my $total_sequences = scalar @seq_ids;
my $fragment_size = int ($total_sequences / $CORES);
my @threads;
for my $i (0 .. ($CORES - 1))
{
        my @partial_array = splice (@seq_ids, 0, $fragment_size);
        my $t = threads->new(\&parse_proteome, \@partial_array);
        push(@threads,$t);
}
my $t = threads->new(\&parse_proteome, \@seq_ids);
push(@threads,$t);

my %cores;
my %organism_total_proteins;
foreach my $thread (@threads)
{
        my @pack = @{$thread->join};
        my %partial_results = %{$pack[0]};
        my %partial_organism_total_proteins = %{$pack[1]};

        while (my ($organism, $predictions_ref) = each (%partial_results))
        {
                $organism_total_proteins{$organism} +=
    $partial_organism_total_proteins{$organism};
                my @predictions;
                while (my ($seq_id, $info_ref) = each (%{$predictions_ref}))
                {
                        my $score = $info_ref->{'Score'};
                        my $core = $info_ref->{'Seq'};
                        my $window_pos = $info_ref->{'Window'};
                        my $protein_id = $1 if ($seq_id =~ m/^>(\w+);/);

                        push @predictions, "\t$protein_id\tWindow Position=$window_pos; Score=$score | Prion Domain: $core\n";
                }
                push @{$cores{$organism}}, @predictions;
        }
}

while (my ($organism, $predictions_ref) = each (%cores))
{
        my @predictions = @{$predictions_ref};
        my $proteins = $organism_total_proteins{$organism};

        my $total_predictions = scalar (@predictions);
        print PRED_FILE_FILTERED ">$organism: Total=$total_predictions\n@predictions\n";
}

close (PRED_FILE_FILTERED);

#A subroutine to move a window along a sequence and report a score relative to the prionogenicity of the stretch
sub parse_proteome
{
        my $ids = shift;
        my @identifiers = @{$ids};

        my %cores;
        my %organism_total_proteins;

        foreach my $seq_id (@identifiers)
        {
                my $seq = $sequences{$seq_id};
                next unless (length ($seq) >= $WINDOW);

                my $organism = &get_organism($seq_id);
                $organism_total_proteins{"$organism"}++;

                my %results;
                for my $i (0 .. (length ($seq) - $WINDOW))
                {
                        my $domain = substr ($seq, $i, $WINDOW);
                        my @domain = split (//, $domain);

                        my $prolines = $domain;
                        $prolines =~ s/[^P]/-/g;
                        @prolines_number = $prolines =~ m/[P]/g;

                        my $domain_score = 0;
                        foreach my $aa (@domain)
                        {
                                $domain_score += $aa_scores{'PrD'}{$aa};
                        }
```

Code C.2: This *ad hoc* script (`surface_analyzer.pl`) is designed and tested to run in a UNIX/Linux environment and requires the previous installation of other third-party software, such as NACCESS and CALC_VOL for estimating the accessible surface areas and atom packing, as described in the Methodology section. It is designed to read PDB files along with a group of command line parameters defining the size of the probe used to decompose the structure of the protein being analyzed. We only show some sections of the 800 lines of the original script

```perl
sub compare_fragments
{
        my ($chain_asa_ref, $fragment_asa_ref, $buried_file, $pymol_file) = @_;

        my %chain_asa = %{$chain_asa_ref};
        my %fragment_asa = %{$fragment_asa_ref};

        my %polar_buried_residues;
        my %volume_info;
        my $fragment_polar_asa = 0;
        my $fragment_apolar_asa = 0;
        my $fragment_vol = 0;
        my $theoretic_fragment_vol = 0;
        my @intrafase_obj;
        my @polar_intrafase_obj;
        my @apolar_intrafase_obj;
        while (my ($resnum, $resname_info_ref) = each (%fragment_asa))
        {
                while (my ($resname, $atom_info_ref) = each (%{$resname_info_ref}))
                {
                        my @output;
                        my @atomnames;
                        my @polar_atomnames;
                        my @apolar_atomnames;
                        my $polar_asa_cum = 0;
                        my $apolar_asa_cum = 0;
                        my $vol_cum = 0;
                        my $theoretic_vol_cum = 0;
                        while (my ($atomname, $asa) = each (%{$atom_info_ref}))
                        {
                                if ($fragment_asa{$resnum}{$resname}{$atomname} > $chain_asa{$resnum}{$resname}{$atomname})
                                {
                                        if ($polarities{$resname}{$atomname} == 1)
                                        {
                                                $polar_asa_cum += $asa;
                                                push @polar_atomnames, $atomname;
                                                push @{$polar_buried_residues{$resname}{$resnum}}, $atomname;
                                        }
                                        elsif ($polarities{$resname}{$atomname} == 0)
                                        {
                                                $apolar_asa_cum += $asa;
```

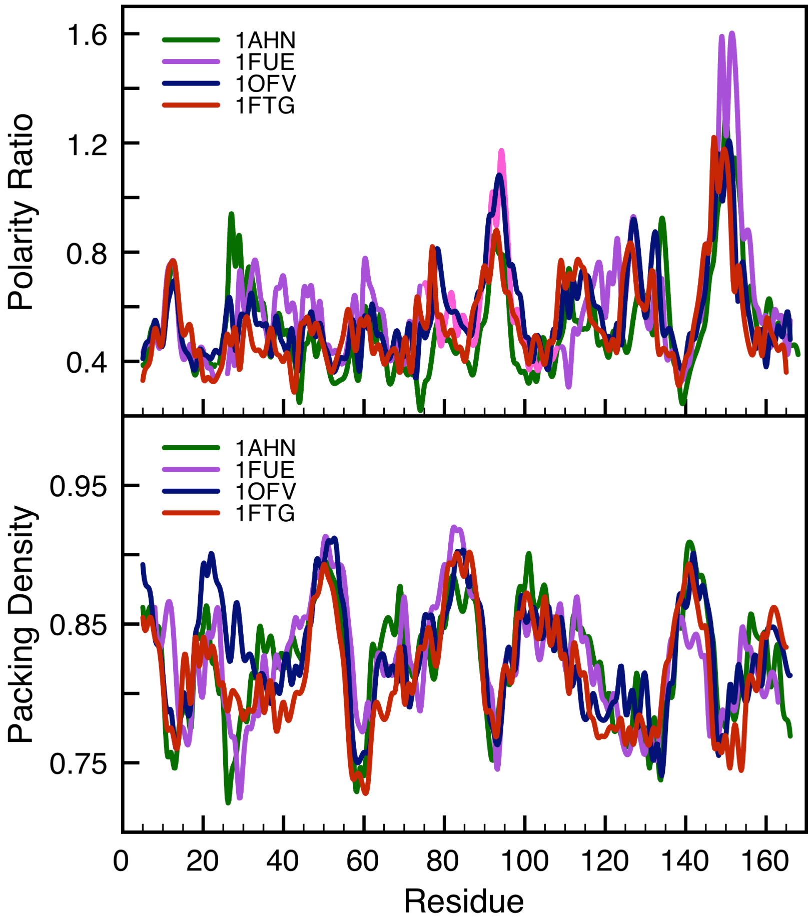
```
                                            push @apolar_atomnames, $atomname
;
                                    }
                                    my $volume = $pdb_vol_info{$resnum}{
$resname}{$atomname};
                                    my $theoretic_vol = $standard_vol{
$resname}{$atomname};
                                    $vol_cum += $volume unless ($volume == -1
.00);
                                    $theoretic_vol_cum += $theoretic_vol
unless ($volume == -1.00);
                                    push @output, "\t\t$atomname = ASA:
$fragment_asa{$resnum}{$resname}{$atomname}(ASA0:$chain_asa{$resnum}{$resname
}{$atomname}); Vol:$volume(Vol0: $theoretic_vol)\n";
                                    push @atomnames, $atomname;
                            }
                    }
                    $fragment_polar_asa += $polar_asa_cum;
                    $fragment_apolar_asa += $apolar_asa_cum;
                    $fragment_vol += $vol_cum;
                    $theoretic_fragment_vol += $theoretic_vol_cum;
                    push @output, "\t\tPolar ASA = $polar_asa_cum | APolar
ASA = $apolar_asa_cum | Volume = $vol_cum | Standard Volume =
$theoretic_vol_cum\n";
                    if (scalar (@output) > 1)
                    {
                            my $corrected_resid = $resnum + $residue_offset;
                            print $buried_file "\t$resname-$corrected_resid:\
n";
                            foreach my $line (@output)
                            {
                                    print $buried_file $line;
                            }
                    }
                    if (scalar (@output) > 1)
                    {
                            my $atomnames = lc join ("+", @atomnames);
                            my $polar_atomnames = lc join ("+",
@polar_atomnames);
                            my $apolar_atomnames = lc join ("+",
@apolar_atomnames);
                            my $corrected_resid = $resnum + $residue_offset;
                            #push @residues, "res_$corrected_resid";
                            push @intrafase_obj, "(resi $corrected_resid and
(name $atomnames))";
                            push @polar_intrafase_obj, "(resi
$corrected_resid and (name $polar_atomnames))" unless (scalar(
@polar_atomnames) == 0);
                            push @apolar_intrafase_obj, "(resi
$corrected_resid and (name $apolar_atomnames))" unless (scalar(
@apolar_atomnames) == 0);
                    }
            }
    }
    my $intrafase_obj = join (" or ", @intrafase_obj);
```

```perl
        my $polar_intrafase_obj = join (" or ", @polar_intrafase_obj);
        my $apolar_intrafase_obj = join (" or ", @apolar_intrafase_obj);
        print $pymol_file "create intrafase, ($intrafase_obj)\n";
        print $pymol_file "create intrafase_polar, ($polar_intrafase_obj)\n";
        print $pymol_file "create intrafase_apolar, ($apolar_intrafase_obj)\n";
        print $buried_file ">>>Fragment resume: Polar ASA = $fragment_polar_asa | APolar ASA = $fragment_apolar_asa | Volume = $fragment_vol | Standard Volume = $theoretic_fragment_vol\n";
        #print $pymol_file "@residues\n";
        print $pymol_file "\n";
        $volume_info{'Vol'} = $fragment_vol;
        $volume_info{'T_Vol'} = $theoretic_fragment_vol;
        return (\%volume_info, \%polar_buried_residues);
}
```

APPENDIX D

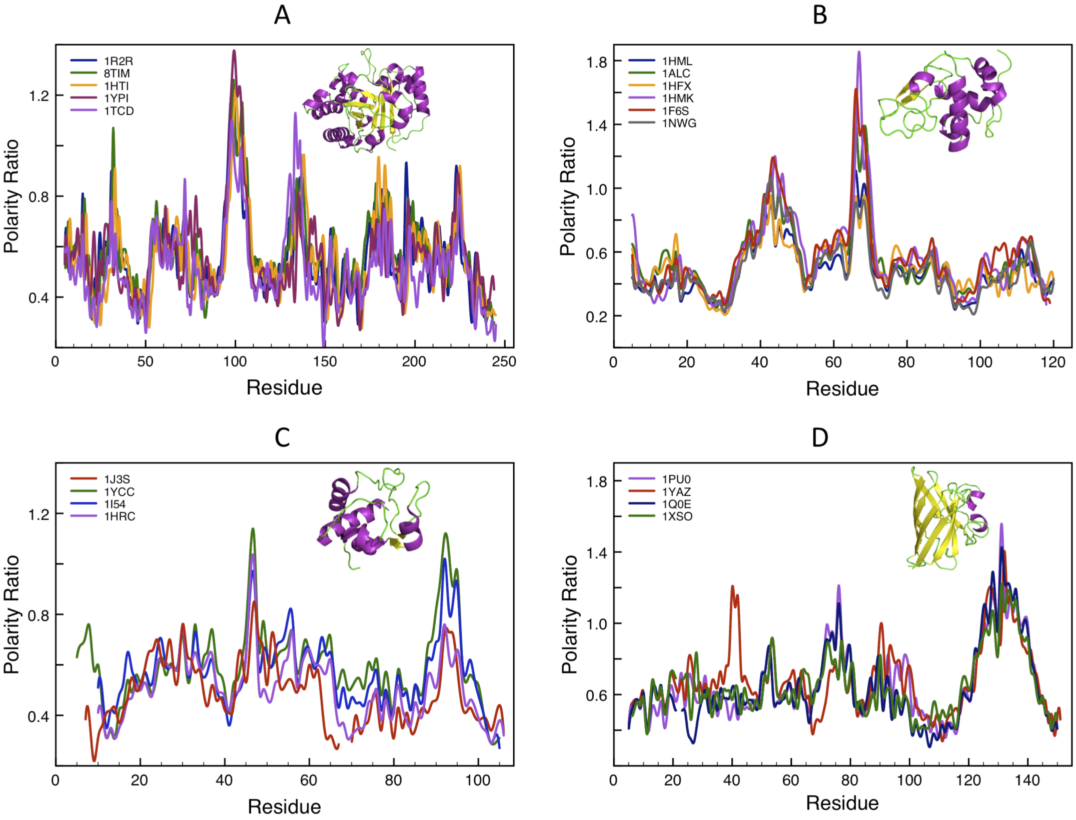
Polarity Ratio and Packing Density Profiles in Some Protein Families

FIGURE D.1: Conservation of the Polarity and Packing Density Profiles in the Flavodoxin Family



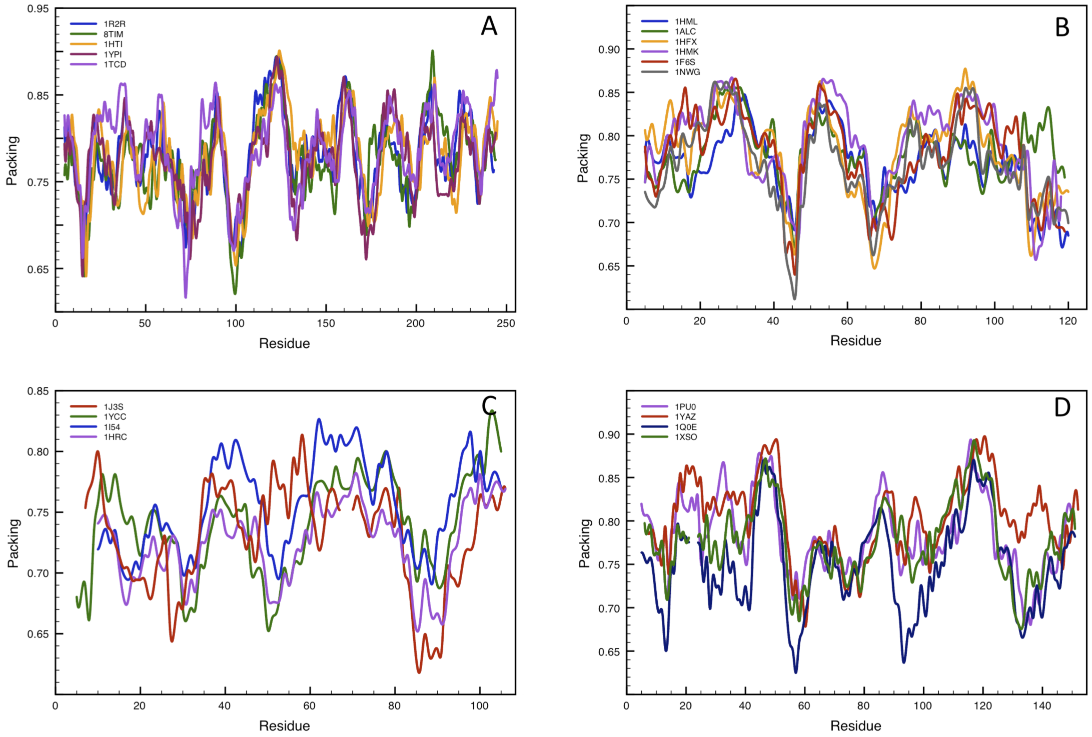
The polarity and packing density profiles of some members of the Flavodoxin family with known structure are shown. The different members of this family were structurally aligned and the polarity and packing profiles obtained were superposed taking the structural alignment as template. The proteins analyzed were (PDB id: 1AHN), (PDB id: 1FUE), (PDB id: 1OFV) and (PDB id: 1FTG)

FIGURE D.2: Conservation of Polarity Profiles in Representative Folds in SCOP (Structural Classification of Proteins) Database



A) SCOP class α/β (fold TIM α/β barrel), **B**) SCOP class $\alpha+\beta$ (fold Lysozyme-like), **C**) SCOP class all α (fold Cytochrome *c*) and **D**) SCOP class all β (fold Immunoglobulin-like β-sandwich). The PDB ids of the proteins analyzed are indicated and a cartoon representation of each protein fold is shown. The Apoflavodoxin profiles shown in Figure D.1, represent an additional example of polarity ratio conservation in class α/β (fold Flavodoxin-like)

FIGURE D.3: Conservation of Packing Density Profiles in Representative Folds in SCOP (Structural Classification of Proteins) Database



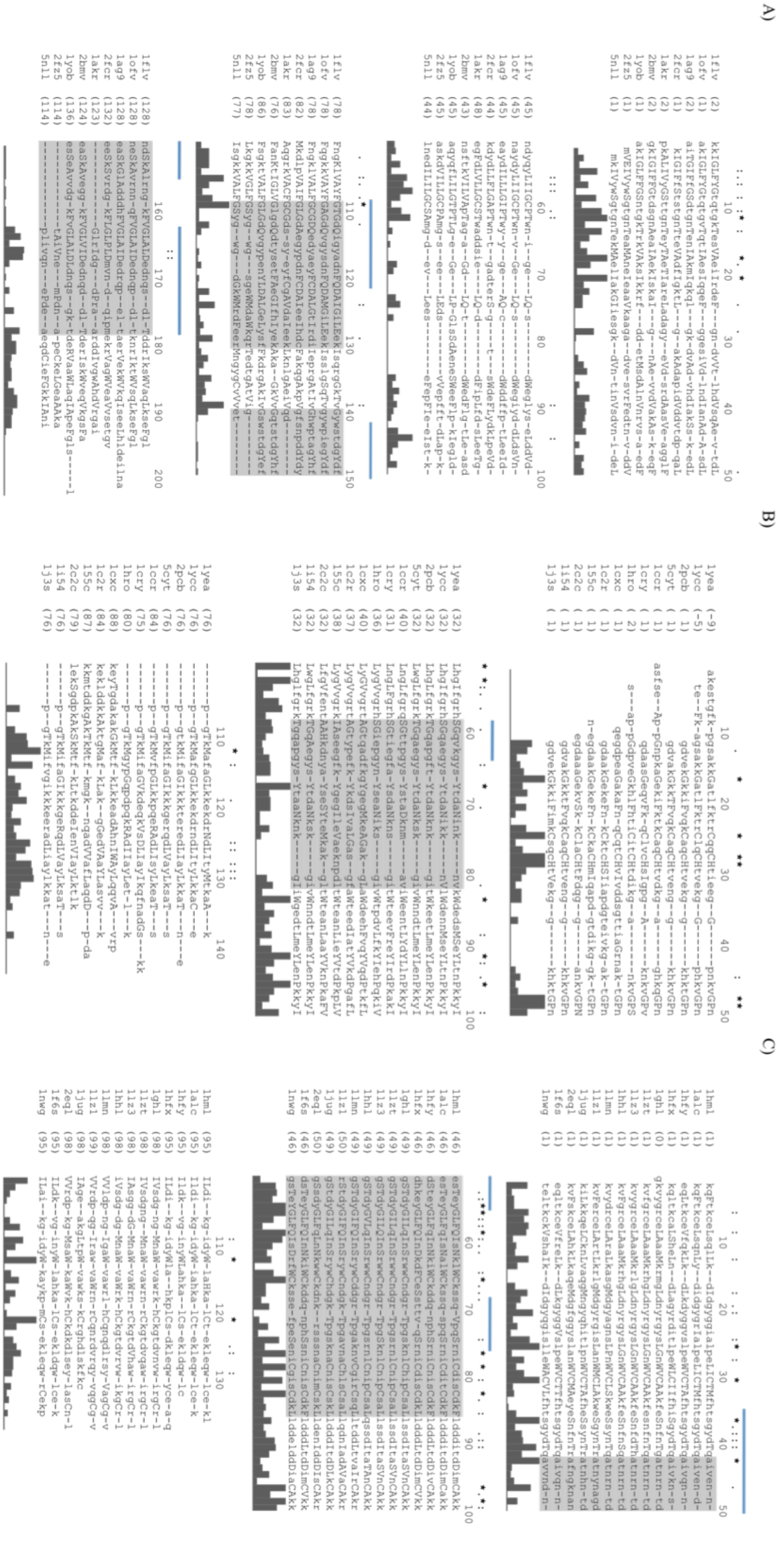
A) SCOP class α/β (fold TIM α/β barrel), **B**) SCOP class $\alpha+\beta$ (fold Lysozyme-like), **C**) SCOP class all α (fold Cytochrome *c*) and **D**) SCOP class all β (fold Immunoglobulin-like β-sandwich). The PDB IDs of the proteins analyzed are indicated and a cartoon representation of each protein fold is shown. The Apoflavodoxin profiles shown in Figure D.1, represent an additional example of polarity ratio conservation in class α/β (fold Flavodoxin-like)

APPENDIX E

Sequence conservation of LIPs in three protein families

Each protein family was aligned as described in the Methodology section and the corresponding structural alignments were processed using JOY to obtain a detailed representation of the alignment including structural information. In the alignments, the solvent accessible residues are in lower case and buried residues in uppercase. We also show in the top line of each alignment fifty-column block the qualitative representation of conservation, and in the bottom line we include the histogram with the quantitative estimates of conservation scores in a given position as calculated by CLUSTAL. With this data, the average conservation scores for alignment columns (corresponding to a given aligned residue) have been calculated for the complete alignments, for the experimentally determined unstable regions, and for the LIPs. We report the PDB codes of the proteins aligned, colored in light grey the experimentally unstable regions, and remarked the LIPs with a light blue line. **A**) Structural alignment of Flavodoxin family (the identity percentage $PID = 34\%$). The unstable regions 87–108 and 118–152 are highlighted. The average column conservation score is 24%. However, the average column conservation score for buried and exposed residues in LIPs are 29 and 9%, respectively, and for buried and exposed residues in the experimentally unstable regions are 30 and 10%, respectively. **B**) Structural alignment of the Cytochrome *c* family ($PID = 47\%$) showing the 40–57 flexible region. The global average column conservation score is 33%, while the average column conservation scores for buried and exposed residues in LIPs are 49 and 30%, respectively, and for buried and exposed residues in the experimentally unstable regions are 42 and 25%, respectively. **C**) Structural alignment of α-Lactalbumin family ($PID = 50.4\%$) with the unstable region 40–80 in grey. The global average column conservation score is 43%, while the average column conservation scores for buried and exposed residues in LIPs are 73 and 41%, respectively, and for buried and exposed residues in the experimentally unstable regions are 70 and 44%, respectively.

FIGURE E.1: Sequence Conservation of LIPs in Three Protein Families



APPENDIX F

LDL-r LA5 domain Missense Mutations

TABLE F.1: We compile the list of non-synonymous SNPs in the coding sequence for the LDL-r LA5 domain. In the first column we include the initial and mutated codons and in the second column the position of the SNP in the LDL-r gene. In column 3 there is a description of the mutation in the protein sequence in the numeration starting from the first translated amino acid (bold) and from the ATG in the gene (italics between braces), respectively. In column 4 we include the results for the prediction of the fate of each mutation –*e.g.* Deleterious (D), Neutral (N)– using the program PMUT, and in column 5 the predictions using the program CONDEL. In the last column we include the two-letter code of the countries where each mutation has been identified. All these data has been extracted from database HGMD Professional, version 2013.3 and the LDL Receptor Database

wt → mt	SNP	wt → mt	PMUT	CONDEL	Country
CCC→ACC	c.586C→A	P**196**{*175*}T	N	N	NL
TGC→CGC	c.589T→C	C**197**{*176*}R	D	D	GB, IT
TGC→GGC	c.589T→G	C**197**{*176*}G	N	D	FR
TGC→TTC	c.590G→T	C**197**{*176*}F	D	D	US
TGC→TGG	c.591C→G	C**197**{*176*}W	D	D	PL
TGC→TAC	c.590G→A	C**197**{*176*}Y	D	D	SV
TCG→TTG	c.593C→T	S**198**{*177*}L	N	N	
TTC→TGC	c.599T→G	F**200**{*179*}C	N	N	
TTC→TTA	c.600C→A	F**200**{*179*}L	N	N	TW, DE
GAG→AAG	c.601G→A	E**201**{*180*}K	N	N	VE, RU
TGC→TTC	c.611G→T	C**204**{*183*}F	D	D	
TGC→TCC	c.611G→C	C**204**{*183*}S	N	D	JP
TGC→TAC	c.611G→A	C**204**{*183*}Y	D	D	FR, IT
AGT→AGG	c.618T→G	S**206**{*185*}R	D	D	RU
GAG→AAG	c.622G→A	E**208**{*187*}K	N	N	NL, IL
TGC→TAC	c.626G→A	C**209**{*188*}Y	D	D	CZ, RU
ATC→AAC	c.629T→A	I**210**{*189*}N	N	D	
CAC→GAC	c.631C→G	H**211**{*190*}D	N	D	PT

Continued on next page...

TABLE F.1: (continued)

wt → mt	**SNP**	**wt → mt**	**PMUT**	**CONDEL**	**Country**
CAC→CTC	c.632A→T	H**211**{*190*}L	D	N	AT
CAC→TAC	c.631C→T	H**211**{*190*}Y	N	D	US
AGC→ACC	c.638G→C	S**213**{*192*}T	N	N	GB
TGG→TCG	c.641G→C	W**214**{*193*}S	D	D	FR
CGC→AGC	c.643C→A	R**215**{*194*}S	N	N	VE
TGT→CGT	c.646T→C	C**216**{*195*}R	D	D	ES
TGT→TTT	c.647G→T	C**216**{*195*}F	D	D	
TGT→TAT	c.647G→A	C**216**{*195*}Y	D	D	GR, DE
GGC→GAC	c.656G→A	G**219**{*198*}D	N	N	GB
GAC→AAC	c.661G→A	D**221**{*200*}N	N	D	
GAC→GGC	c.662A→G	D**221**{*200*}G	N	D	NL, IT, AT, PT, ES
GAC→TAC	c.661G→T	D**221**{*200*}Y	N	D	PT, ES, IT, DE, FI
GAC→GTC	c.662A→T	D**221**{*200*}V	N	D	ES
TGC→CGC	c.664T→C	C**222**{*201*}R	D	D	DK
TGC→GGC	c.664T→G	C**222**{*201*}G	D	D	GB
TGC→TTC	c.665G→T	C**222**{*201*}F	D	D	NL
TGC→TAC	c.665G→A	C**222**{*201*}Y	D	D	FI, IT
GAC→GCC	c.671A→C	D**224**{*203*}A	N	D	ZA
GAC→AAC	c.670G→A	D**224**{*203*}N	N	N	PT
GAC→GGC	c.671A→G	D**224**{*203*}G	N	N	IT
GAC→GTC	c.671A→T	D**224**{*203*}V	N	D	DE
TCT→TGT	c.677C→G	S**226**{*205*}C	N	D	ES
TCT→CCT	c.676T→C	S**226**{*205*}P	N	D	US
GAC→GAG	c.681C→G	D**227**{*206*}E	N	D	GB, NL, IT, ZA
GAC→GTC	c.680A→T	D**227**{*206*}V	N	D	FR
GAG→GCG	c.683A→C	E**228**{*207*}A	D	D	CN
GAG→CAG	c.682G→C	E**228**{*207*}Q	N	N	GB
GAG→AAG	c.682G→A	E**228**{*207*}K	N	D	PT, NL, ES, IT
TGC→CGC	c.691T→C	C**231**{*210*}R	D	D	CN
TGC→GGC	c.691T→G	C**231**{*210*}G	N	D	NO
TGC→TGG	c.693C→G	C**231**{*210*}W	D	D	FR
TGC→TAC	c.692G→A	C**231**{*210*}Y	D	D	KR

TABLE F.2: A list of all the possible non-synonymous SNPs that can be generated in the coding sequence for the LDL-r LA5 domain. In the first column are the initial and mutated codons, in the second column the mutation in the protein sequence in the numeration starting from the first translated amino acid (bold) and from the ATG in the gene (italics between braces), respectively. In column 3 are the indexes used to identify the mutations. In columns 4 and 5 are the predictions of the fate of each mutation –*e.g.* Deleterious (D), Neutral (N)– using the programs PMUT and CONDEL, respectively. Due to the degenerate nature of the genetic code in some cases different SNPs could result in the same amino acid change. In these cases the corresponding rows are surrounded by horizontal lines. All mutations identified in persons with FH (Appendix Table F.1) are highlighted with an asterisk (*), and those in residues located in the interaction sites for the β-propeller and ApoB and ApoE, are highlighted with (⊕). The colored bullets in the sixth column correspond to a code for the clustering of mutants according to their conformational instability, as described in Figure 4.11, on which green corresponds to (stable mutants), orange to (unstable mutants), magenta to (very unstable mutants) and red to (highly unstable mutants). In the seventh column there is a classification according to our study for each mutation, considering both the contribution of the effect on the conformational stability of the domain (sixth column), and the occurrence of the mutated residue in the domain interaction sites with other domains from the LDL-r, or other proteins. In the latter case, the effect of the mutation in binding was rationalized following qualitative criteria of the physicochemical and steric change in a given residue upon mutation. For those cases the phenotype prediction are in italics

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
*CCC→ACC	P**196**{*175*}T	M001	N	N	🟢	N
CCC→GCC	P**196**{*175*}A	M002	N	N	🟢	N
CCC→TCC	P**196**{*175*}S	M003	N	N	🟠	D
CCC→CAC	P**196**{*175*}H	M004	N	N	🟠	D
CCC→CGC	P**196**{*175*}R	M005	N	N	🟢	N
CCC→CTC	P**196**{*175*}L	M006	N	N	🟢	N
TGC→AGC TGC→TCC	C**197**{*176*}S	M007	N	N	🔴	D
*TGC→CGC	C**197**{*176*}R	M008	D	D	🟣	D
*TGC→GGC	C**197**{*176*}G	M009	N	D	🔴	D
*TGC→TAC	C**197**{*176*}Y	M010	D	D	🔴	D
*TGC→TTC	C**197**{*176*}F	M011	D	D	🔴	D
*TGC→TGG	C**197**{*176*}W	M012	D	D	🟣	D
TCG→ACG	⊕S**198**{*177*}T	M013	N	N	🟠	*D*
TCG→CCG	⊕S**198**{*177*}P	M014	N	N	🟠	*D*
TCG→GCG	⊕S**198**{*177*}A	M015	N	N	🟢	*D*
TCG→TGG	⊕S**198**{*177*}W	M016	D	D	🟣	*D*
*TCG→TTG	⊕S**198**{*177*}L	M017	N	N	🟢	*D*
GCC→ACC	A**199**{*178*}T	M018	N	N	🟢	N
GCC→CCC	A**199**{*178*}P	M019	N	N	🟢	N

Continued on next page...

TABLE F.2: (continued)

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
GCC→TCC	A**199**{*178*}S	M020	N	N	🟢	N
GCC→GAC	A**199**{*178*}D	M021	N	N	🟢	N
GCC→GGC	A**199**{*178*}G	M022	N	N	🟢	N
GCC→GTC	A**199**{*178*}V	M023	N	N	🟠	D
TTC→ATC	F**200**{*179*}I	M024	N	N	🟢	N
TTC→CTC *TTC→TTA TTC→TTG	F**200**{*179*}L	M025	N	N	🟠	D
TTC→GTC	F**200**{*179*}V	M026	N	N	🟣	D
TTC→TAC	F**200**{*179*}Y	M027	N	N	🟢	N
TTC→TCC	F**200**{*179*}S	M028	N	N	🟢	N
*TTC→TGC	F**200**{*179*}C	M029	N	N	🔴	D
*GAG→AAG	⊕E**201**{*180*}K	M030	N	N	🟢	*D*
GAG→CAG	⊕E**201**{*180*}Q	M031	N	D	🟢	*D*
GAG→GCG	⊕E**201**{*180*}A	M032	N	D	🔴	*D*
GAG→GGG	⊕E**201**{*180*}G	M033	N	D	🔴	*D*
GAG→GTG	⊕E**201**{*180*}V	M034	N	D	🟣	*D*
GAG→GAT GAG→GAC	⊕E**201**{*180*}D	M035	N	D	🟣	*D*
TTC→ATC	F**202**{*181*}I	M036	N	D	🟠	D
TTC→TTG TTC→TTA TTC→CTC	F**202**{*181*}L	M037	N	N	🟣	D
TTC→GTC	F**202**{*181*}V	M038	N	D	🟢	N
TTC→TAC	F**202**{*181*}Y	M039	N	D	🟢	N
TTC→TCC	F**202**{*181*}S	M040	N	D	🟠	D
TTC→TGC	F**202**{*181*}C	M041	N	D	🟣	D
CAC→AAC	H**203**{*182*}N	M042	N	N	🟢	N
CAC→GAC	H**203**{*182*}D	M043	N	N	🟠	D
CAC→TAC	H**203**{*182*}Y	M044	N	N	🟠	D
CAC→CCC	H**203**{*182*}P	M045	D	N	🟢	N
CAC→CGC	H**203**{*182*}R	M046	N	N	🟠	D
CAC→CTC	H**203**{*182*}L	M047	N	N	🟢	N
CAC→CAA CAC→CAG	H**203**{*182*}Q	M048	N	N	🟠	D

Continued on next page. . .

TABLE F.2: (continued)

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
*TGC→TCC TGC→AGC	C**204**{*183*}S	M049	N	D	🟣	D
TGC→CGC	C**204**{*183*}R	M050	D	D	🟢	N
TGC→GGC	C**204**{*183*}G	M051	N	D	🔴	D
*TGC→TAC	C**204**{*183*}Y	M052	D	D	🔴	D
*TGC→TTC	C**204**{*183*}F	M053	D	D	🟣	D
TGC→TGG	C**204**{*183*}W	M054	D	D	🟢	N
CTA→ATA	L**205**{*184*}I	M055	N	N	🟢	N
CTA→GTA	L**205**{*184*}V	M056	N	N	🟢	N
CTA→CAA	L**205**{*184*}Q	M057	N	N	🟢	N
CTA→CCA	L**205**{*184*}P	M058	N	N	🟢	N
CTA→CGA	L**205**{*184*}R	M059	N	N	🟢	N
*AGT→AGG AGT→AGA AGT→CGT	S**206**{*185*}R	M060	D	D	🟣	D
AGT→GGT	S**206**{*185*}G	M061	N	N	🟢	N
AGT→TGT	S**206**{*185*}C	M062	N	D	🟢	N
AGT→AAT	S**206**{*185*}N	M063	N	N	🟢	N
AGT→ACT	S**206**{*185*}T	M064	N	N	🟠	D
AGT→ATT	S**206**{*185*}I	M065	N	D	🟠	D
GGC→AGC	G**207**{*186*}S	M066	N	N	🟢	N
GGC→CGC	G**207**{*186*}R	M067	D	N	🔴	D
GGC→TGC	G**207**{*186*}C	M068	N	D	🟣	D
GGC→GAC	G**207**{*186*}D	M069	N	N	🟢	N
GGC→GCC	G**207**{*186*}A	M070	N	N	🟠	D
GGC→GTC	G**207**{*186*}V	M071	N	D	🟠	D
*GAG→AAG	E**208**{*187*}K	M072	N	N	🟠	D
GAG→CAG	E**208**{*187*}Q	M073	N	N	🟢	N
GAG→GCG	E**208**{*187*}A	M074	N	N	🟣	D
GAG→GGG	E**208**{*187*}G	M075	N	D	🟢	N
GAG→GTG	E**208**{*187*}V	M076	N	D	🟣	D
GAG→GAT GAG→GAC	E**208**{*187*}D	M077	N	N	🟣	D
TGC→AGC TGC→TCC	⊕C**209**{*188*}S	M078	N	D	🟢	*D*

Continued on next page. . .

TABLE F.2: (continued)

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
TGC→CGC	⊕**C209**{*188*}R	M079	D	D	🟣	*D*
TGC→GGC	⊕**C209**{*188*}G	M080	N	D	🟣	*D*
*TGC→TAC	⊕**C209**{*188*}Y	M081	D	D	🟢	*D*
TGC→TTC	⊕**C209**{*188*}F	M082	D	D	🟣	*D*
TGC→TGG	⊕**C209**{*188*}W	M083	D	D	🟢	*D*
ATC→CTC	**I210**{*189*}L	M084	N	N	🟠	D
ATC→GTC	**I210**{*189*}V	M085	N	N	🟣	D
ATC→TTC	**I210**{*189*}F	M086	N	D	🟢	N
*ATC→AAC	**I210**{*189*}N	M087	N	D	🟠	D
ATC→ACC	**I210**{*189*}T	M088	N	D	🟢	N
ATC→AGC	**I210**{*189*}S	M089	N	D	🟢	N
ATC→ATG	**I210**{*189*}M	M090	N	D	🟠	D
CAC→AAC	⊕**H211**{*190*}N	M091	N	D	🟢	*D*
*CAC→GAC	⊕**H211**{*190*}D	M092	N	D	🟢	*D*
*CAC→TAC	⊕**H211**{*190*}Y	M093	N	D	🟢	*D*
CAC→CCC	⊕**H211**{*190*}P	M094	D	N	🟢	*D*
CAC→CGC	⊕**H211**{*190*}R	M095	N	N	🟣	*D*
*CAC→CTC	⊕**H211**{*190*}L	M096	D	N	🟢	*D*
CAC→CAG CAC→CAA	⊕**H211**{*190*}Q	M097	N	D	🟠	*D*
TCC→ACC	⊕**S212**{*191*}T	M098	N	N	🟠	*D*
TCC→CCC	⊕**S212**{*191*}P	M099	N	N	🟢	*D*
TCC→GCC	⊕**S212**{*191*}A	M100	N	N	🟢	*D*
TCC→TAC	⊕**S212**{*191*}Y	M101	D	N	🟠	*D*
TCC→TGC	⊕**S212**{*191*}C	M102	N	N	🟢	*D*
TCC→TTC	⊕**S212**{*191*}F	M103	D	N	🟢	*D*
AGC→CGC AGC→AGA AGC→AGG	**S213**{*192*}R	M104	D	N	🟢	N
AGC→GGC	**S213**{*192*}G	M105	N	N	🟠	D
AGC→TGC	**S213**{*192*}C	M106	N	N	🟢	N
AGC→AAC	**S213**{*192*}N	M107	N	N	🟢	N
*AGC→ACC	**S213**{*192*}T	M108	N	N	🔴	D
AGC→ATC	**S213**{*192*}I	M109	N	N	🟢	N

Continued on next page. . .

TABLE F.2: (continued)

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
TGG→AGG TGG→CGG	⊕W**214**{*193*}R	M110	D	D	🟠	*D*
TGG→GGG	⊕W**214**{*193*}G	M111	D	D	🔴	*D*
*TGG→TCG	⊕W**214**{*193*}S	M112	D	D	🟢	*D*
TGG→TTG	⊕W**214**{*193*}L	M113	D	D	🟢	*D*
TGG→TGT TGG→TGC	⊕W**214**{*193*}C	M114	D	D	🟠	*D*
*CGC→AGC	R**215**{*194*}S	M115	N	N	🟠	D
CGC→GGC	R**215**{*194*}G	M116	N	N	🟢	N
CGC→TGC	R**215**{*194*}C	M117	N	D	🟠	D
CGC→CAC	R**215**{*194*}H	M118	N	N	🟢	N
CGC→CCC	R**215**{*194*}P	M119	D	N	🟢	N
CGC→CTC	R**215**{*194*}L	M120	N	N	🟢	N
TGT→AGT TGT→TCT	C**216**{*195*}S	M121	N	D	🟢	N
*TGT→CGT	C**216**{*195*}R	M122	D	D	🟠	D
TGT→GGT	C**216**{*195*}G	M123	D	D	🟣	D
*TGT→TAT	C**216**{*195*}Y	M124	D	D	🟢	N
*TGT→TTT	C**216**{*195*}F	M125	D	D	🟢	N
TGT→TGG	C**216**{*195*}W	M126	D	D	🟢	N
GAT→AAT	⊕D**217**{*196*}N	M127	N	D	🟢	*D*
GAT→CAT	⊕D**217**{*196*}H	M128	N	D	🟢	*D*
GAT→TAT	⊕D**217**{*196*}Y	M129	N	D	🟠	*D*
GAT→GCT	⊕D**217**{*196*}A	M130	N	D	🟠	*D*
GAT→GGT	⊕D**217**{*196*}G	M131	D	D	🟣	*D*
GAT→GTT	⊕D**217**{*196*}V	M132	D	D	🟢	*D*
GAT→GAG GAT→GAA	⊕D**217**{*196*}E	M133	N	D	🟠	*D*
GGT→AGT	⊕G**218**{*197*}S	M134	N	D	🟠	*D*
GGT→CGT	⊕G**218**{*197*}R	M135	D	D	🟢	*D*
GGT→TGT	⊕G**218**{*197*}C	M136	N	D	🟢	*D*
GGT→GAT	⊕G**218**{*197*}D	M137	N	D	🟠	*D*
GGT→GCT	⊕G**218**{*197*}A	M138	N	N	🟢	*D*
GGT→GTT	⊕G**218**{*197*}V	M139	D	D	🟢	*D*
GGC→AGC	⊕G**219**{*198*}S	M140	N	N	🟢	*D*

Continued on next page...

TABLE F.2: (continued)

wt → mt	wt → mt	Index	PMUT	CONDEL	Cluster	Phenotype
GGC→CGC	⊕G**219**{*198*}R	M141	D	N	🟢	*D*
GGC→TGC	⊕G**219**{*198*}C	M142	N	D	🟢	*D*
*GGC→GAC	⊕G**219**{*198*}D	M143	N	N	🟠	*D*
GGC→GCC	⊕G**219**{*198*}A	M144	N	N	🟢	*D*
GGC→GTC	⊕G**219**{*198*}V	M145	D	D	🟢	*D*
CCC→ACC	P**220**{*199*}T	M146	N	N	🟢	N
CCC→GCC	P**220**{*199*}A	M147	N	N	🟠	D
CCC→TCC	P**220**{*199*}S	M148	N	N	🟢	N
CCC→CAC	P**220**{*199*}H	M149	N	N	🟢	N
CCC→CGC	P**220**{*199*}R	M150	N	N	🟢	N
CCC→CTC	P**220**{*199*}L	M151	N	N	🟣	D
*GAC→AAC	⊕D**221**{*200*}N	M152	N	D	🟢	*D*
GAC→CAC	⊕D**221**{*200*}H	M153	N	D	🔴	*D*
*GAC→TAC	⊕D**221**{*200*}Y	M154	N	D	🟣	*D*
GAC→GCC	⊕D**221**{*200*}A	M155	D	D	🟣	*D*
*GAC→GGC	⊕D**221**{*200*}G	M156	N	D	🟢	*D*
*GAC→GTC	⊕D**221**{*200*}V	M157	N	D	🔴	*D*
GAC→GAG GAC→GAA	⊕D**221**{*200*}E	M158	N	D	🟢	*D*
TGC→TCC TGC→AGC	C**222**{*201*}S	M159	N	D	🟢	N
*TGC→CGC	C**222**{*201*}R	M160	D	D	🟠	D
*TGC→GGC	C**222**{*201*}G	M161	D	D	🟠	D
*TGC→TAC	C**222**{*201*}Y	M162	D	D	🟣	D
*TGC→TTC	C**222**{*201*}F	M163	D	D	🔴	D
TGC→TGG	C**222**{*201*}W	M164	D	D	🔴	D
AAG→CAG	⊕K**223**{*202*}Q	M165	N	N	🟢	*D*
AAG→GAG	⊕K**223**{*202*}E	M166	N	N	🟠	*D*
AAG→ACG	⊕K**223**{*202*}T	M167	D	N	🟠	*D*
AAG→AGG	⊕K**223**{*202*}R	M168	N	N	🔴	*D*
AAG→ATG	⊕K**223**{*202*}M	M169	N	D	🟠	*D*
AAG→AAT AAG→AAC	⊕K**223**{*202*}N	M170	N	N	🟢	*D*
*GAC→AAC	D**224**{*203*}N	M171	N	N	🟠	D
GAC→CAC	D**224**{*203*}H	M172	N	D	🟣	D

Continued on next page. . .

TABLE F.2: (continued)

wt → mt	wt → mt	Index	PMUT	CONDEL	Cluster	Phenotype
GAC→TAC	D**224**{*203*}Y	M173	N	D	🔴	D
*GAC→GCC	D**224**{*203*}A	M174	N	D	🟠	D
*GAC→GGC	D**224**{*203*}G	M175	N	N	🟠	D
*GAC→GTC	D**224**{*203*}V	M176	N	D	🟠	D
GAC→GAG GAC→GAA	D**224**{*203*}E	M177	N	D	🔴	D
AAA→CAA	K**225**{*204*}Q	M178	N	N	🟢	N
AAA→GAA	K**225**{*204*}E	M179	N	N	🟢	N
AAA→ACA	K**225**{*204*}T	M180	D	D	🟣	D
AAA→AGA	K**225**{*204*}R	M181	N	N	🟢	N
AAA→ATA	K**225**{*204*}I	M182	N	N	🟢	N
AAA→AAT AAA→AAC	K**225**{*204*}N	M183	N	N	🟠	D
TCT→ACT	S**226**{*205*}T	M184	N	N	🟢	N
*TCT→CCT	S**226**{*205*}P	M185	N	D	🟣	D
TCT→GCT	S**226**{*205*}A	M186	N	N	🟠	D
TCT→TAT	S**226**{*205*}Y	M187	N	D	🟠	D
*TCT→TGT	S**226**{*205*}C	M188	N	D	🟠	D
TCT→TTT	S**226**{*205*}F	M189	N	D	🟠	D
GAC→AAC	⊕D**227**{*206*}N	M190	N	D	🟠	*D*
GAC→CAC	⊕D**227**{*206*}H	M191	N	D	🔴	*D*
GAC→TAC	⊕D**227**{*206*}Y	M192	N	D	🟣	*D*
GAC→GCC	⊕D**227**{*206*}A	M193	N	D	🟣	*D*
GAC→GGC	⊕D**227**{*206*}G	M194	N	D	🟣	*D*
*GAC→GTC	⊕D**227**{*206*}V	M195	N	D	🟣	*D*
GAC→GAG *GAC→GAA	⊕D**227**{*206*}E	M196	N	D	🟢	*D*
*GAG→AAG	⊕E**228**{*207*}K	M197	N	D	🔴	*D*
*GAG→CAG	⊕E**228**{*207*}Q	M198	N	D	🟢	*D*
*GAG→GCG	⊕E**228**{*207*}A	M199	D	D	🟠	*D*
GAG→GGG	⊕E**228**{*207*}G	M200	N	D	🟠	*D*
GAG→GTG	⊕E**228**{*207*}V	M201	N	D	🟠	*D*
GAG→GAT GAG→GAC	⊕E**228**{*207*}D	M202	N	N	🔴	*D*
GAA→AAA	E**229**{*208*}K	M203	D	N	🟢	N

Continued on next page. . .

Table F.2: (continued)

wt → mt	**wt → mt**	**Index**	**PMUT**	**CONDEL**	**Cluster**	**Phenotype**
GAA→CAA	E**229**{*208*}Q	M204	N	N	🟠	D
GAA→GCA	E**229**{*208*}A	M205	D	N	🟢	N
GAA→GGA	E**229**{*208*}G	M206	N	N	🟠	D
GAA→GTA	E**229**{*208*}V	M207	N	N	🟢	N
GAA→GAT	E**229**{*208*}D	M208	N	N	🟢	N
GAA→GAC						
AAC→CAC	N**230**{*209*}H	M209	N	N	🟢	N
AAC→GAC	N**230**{*209*}D	M210	N	N	🟢	N
AAC→TAC	N**230**{*209*}Y	M211	D	N	🟢	N
AAC→ACC	N**230**{*209*}T	M212	N	N	🟢	N
AAC→AGC	N**230**{*209*}S	M213	N	N	🟢	N
AAC→ATC	N**230**{*209*}I	M214	D	N	🟢	N
AAC→AAG	N**230**{*209*}K	M215	D	N	🟢	N
AAC→AAA						
TGC→TCC	C**231**{*210*}S	M216	N	N	🟣	D
TGC→AGC						
*TGC→CGC	C**231**{*210*}R	M217	D	D	🟢	N
*TGC→GGC	C**231**{*210*}G	M218	N	D	🟢	N
*TGC→TAC	C**231**{*210*}Y	M219	D	D	🟢	N
TGC→TTC	C**231**{*210*}F	M220	D	D	🟢	N
*TGC→TGG	C**231**{*210*}W	M221	D	D	🟠	D
GCT→ACT	A**232**{*211*}T	M222	N	-	🟢	
GCT→CCT	A**232**{*211*}P	M223	N	-	🟢	
GCT→TCT	A**232**{*211*}S	M224	N	-	🟢	
GCT→GAT	A**232**{*211*}D	M225	N	-	🟢	
GCT→GGT	A**232**{*211*}G	M226	N	-	🟢	
GCT→GTT	A**232**{*211*}V	M227	N	-	🟢	

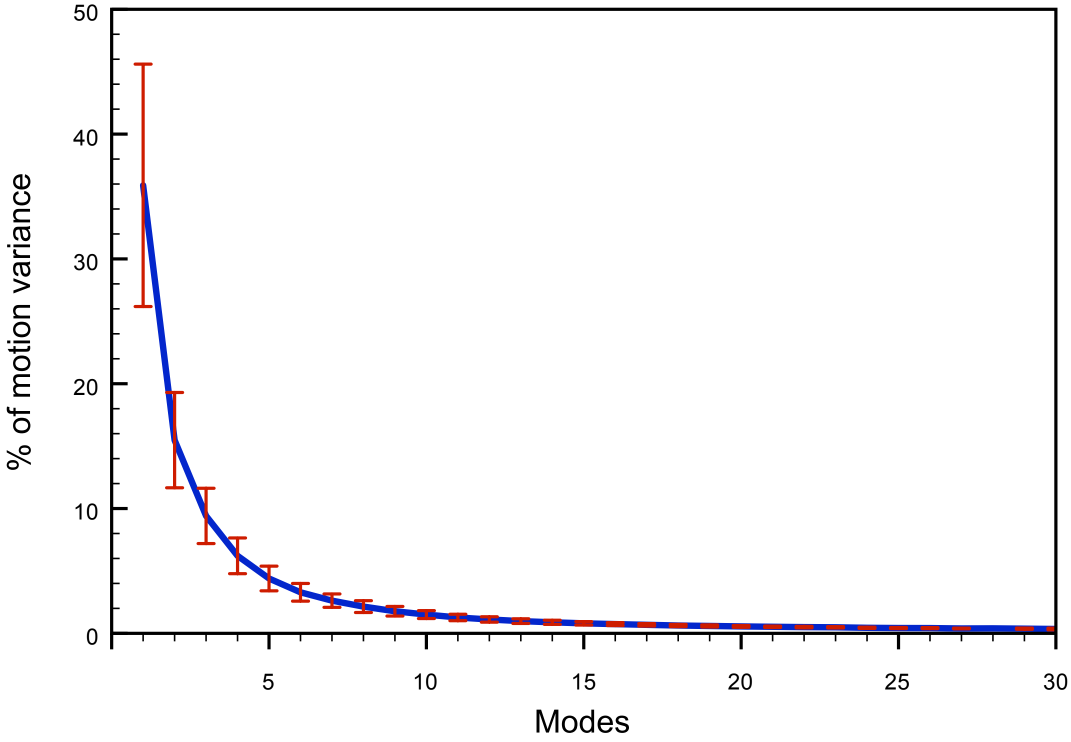
TABLE F.3: We include a table with the link addresses to the online videos for the dynamical evolution of the PC projection density plots of some of the LDL-r mutants referenced in the text. In the first column are the amino acid changes, and in the second one the URL addresses for the online videos. This table is meant to serve as guide for accessing the videos in case of being reading the printed version of this document, and includes the examples described in Figures 4.5, 4.7 and 4.12

wt → mt	**URL**
C197G	https://drive.google.com/file/d/0B2oR8_NjxhbUOUNzaW5tVWYtSTg/edit?usp=sharing
F200C	https://drive.google.com/file/d/0B2oR8_NjxhbULWtadnJPckhGVjA/edit?usp=sharing
D221H	https://drive.google.com/file/d/0B2oR8_NjxhbUUjY4QW9BaC1ncjg/edit?usp=sharing
C222F	https://drive.google.com/file/d/0B2oR8_NjxhbUOTFYQTNSSXAxZ00/edit?usp=sharing
E228K	https://drive.google.com/file/d/0B2oR8_NjxhbUSXYwZ2tRVFlwVE0/edit?usp=sharing
A199G	https://drive.google.com/file/d/0B2oR8_NjxhbUWlBBVktXNE5lZlE/edit?usp=sharing
L205V	https://drive.google.com/file/d/0B2oR8_NjxhbUWFJtRXBIQi1NRVU/edit?usp=sharing
S206N	https://drive.google.com/file/d/0B2oR8_NjxhbUcFZJQ1VJMlNucU0/edit?usp=sharing
C209W	https://drive.google.com/file/d/0B2oR8_NjxhbUcXR1MzA5b2NoTkE/edit?usp=sharing
A232G	https://drive.google.com/file/d/0B2oR8_NjxhbUYVpjci1ydjFBTnM/edit?usp=sharing
C209Y	https://drive.google.com/file/d/0B2oR8_NjxhbUUmRJdG9Ya3hEY2M/edit?usp=sharing
W214S	https://drive.google.com/file/d/0B2oR8_NjxhbUalg3Z2liZFFOS2c/edit?usp=sharing
C216Y	https://drive.google.com/file/d/0B2oR8_NjxhbUeHRFanBFSm5lT0U/edit?usp=sharing
E228Q	https://drive.google.com/file/d/0B2oR8_NjxhbUakZ2ZzY1aDV2MlU/edit?usp=sharing
C231Y	https://drive.google.com/file/d/0B2oR8_NjxhbUenFNSEFaSU5mWU0/edit?usp=sharing

Appendix G

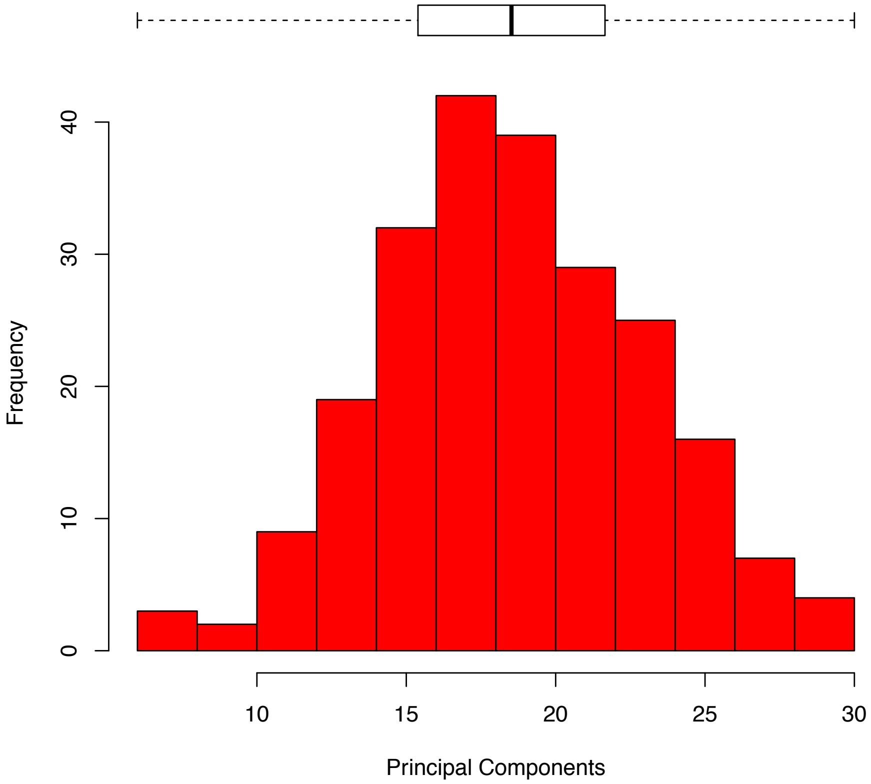
PCA Summary of all LDL-r LA5 Mutants MD Trajectories

Figure G.1: Motion Variance Described by Each Principal Component in LDL-r LA5 Mutants MD Trajectories



'Scree Plot' of the percentage of the motion variance that is described by each Principal Component. For all the 227 MD simulations we plot the mean and standard deviation of the motion variance –*i.e.* the eigenvalues– that is described by each mode –*i.e.* the eigenvectors– between the first and the thirtieth

FIGURE G.2: Distribution of Principal Components in LDL-r LA5 Mutants MD Trajectories



The number of Principal Components extracted from the PCA of the 227 LDL-r LA5 mutants MD trajectories to describe 95% of the motion variance

Appendix H

Curriculum Vitae

Personal Information

- **Date of Birth:** November the 6th, 1978
- **Place of Birth:** Havana, Cuba

Education

- **2009–2014: Ph.D. in Biochemistry**, Department of Biochemistry and Molecular and Cellular Biology, University of Zaragoza, Spain
- **2007–2008: M.A.S. in Molecular and Cellular Biology**, Department of Biochemistry and Molecular and Cellular Biology, University of Zaragoza, Spain (*cum laude*)
- **2001–2002: Bioinformatics Specialist Training**, Center for Genetic Engineering and Biotechnology (CIGB), Havana, Cuba
- **1997–2002: B.Sc. in Biochemistry**, Biology Faculty, University of Havana, Cuba (*summa cum laude*)
- **1993–1996: 3-year Bachelor**, Exact Sciences Vocational Institute "Vladimir Ilich Lenin", Havana, Cuba (*summa cum laude*)

Postgraduate Courses and Trainings

- **2013: Postgraduate Course "Translational and Integrative Bioinformatics"**, organized by INBIOMEDvision in collaboration with Pompeu Fabra University, Bioinformatics Barcelona, Barcelona Supercomputing Center, Spanish Technological Platform for Innovative Medicines, and The Spanish Institute of Bioinformatics, Barcelona, Spain

- **2008: Workshop "Getting the most from biomolecular simulations"**, University of Barcelona/Institute for Research in Biomedicine/Barcelona Supercomputer Center, Barcelona, Spain

- **2007: Trends in Transient Interactions between Biological Molecules**, Institute of Plant Biochemistry and Photosynthesis, University of Sevilla, Spain

- **2007: IInd Meeting of the Spanish Group of Protein Structure and Folding**, University Miguel Hernández, San Juan de Alicante, Spain

- **2003: International course New Frontiers of Bioinformatics in Latin America**, University of "Los Andes", Mérida, Venezuela

- **2002: Introduction to Logics**, Mathematics Faculty, University of Havana, Cuba

- **2002: French-Cuban Doctoral School (on the subject of Data Mining)**, Mathematics Faculty, University of Havana, Cuba

- **2002: Postgraduate Course on Bioinformatics**, Mathematics Faculty, University of Havana, Cuba

- **2001: Postgraduate course on Introduction to Protein Computational Modeling**, Physics Faculty, University of Havana, Cuba

- **2001: IIIrd International Workshop on Nucleic Acids and Proteins sequence analysis**, Center for Genetic Engineering and Biotechnology, Havana, Cuba

- **2001: Post-Congress course Interaction of Peptides and Proteins with Mimetic Membrane Systems**, Ist International Symposium on Biochemistry and Molecular Biology, Havana, Cuba

Peer-reviewed Publications

Articles Published While at the University of Zaragoza Working in my PhD Thesis:

1. Gonzalez, A., Angarica, V.E., Sancho, J., and Fillat, M.F. (2014). **The FurA regulon in Anabaena sp. PCC 7120: in silico prediction and experimental validation of novel target genes**. *Nucleic Acids Res.*, **accepted**

2. Angarica, V.E., Angulo, A., Giner, A., Losilla, G., Ventura, S., and Sancho, J. (2014). **PrionScan: an online database of predicted prion domains in complete proteomes**. *BMC Genomics*, **15:** 102

3. Angarica, V.E., Ventura, S. and Sancho, J. (2013). **Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains**. *BMC Genomics*, **14:** 316

4. Angarica, V.E., Sancho, J. (2012). **Protein dynamics governed by interfaces of high polarity and low packing density**. *PLoS One*, **7(10):** e48212

5. Martinez-Julvez, M., Rojas, A., Olekhnovich, I., Angarica, V.E., Hoffman, P.S. and Sancho, J. (2012). **Structure of RdxA: an oxygen insensitive nitroreductase essential for metronidazole activation in Helicobacter pylori**. *FEBS J*, **279(23):** 4306–17

6. Ayuso-Tejedor, S., Angarica, V.E., Bueno, M., Campos, L.A., Abian, O., Bernado, P., Sancho, J., Jimenez, M.A. (2010). **Design and structure of a protein folding inter- mediate. A hint into dynamical regions of proteins**. *J. Mol. Biol.*, **400(4):** 922–34

7. Contreras-Moreira, B., Sancho, J. & Angarica, V.E. (2009). **Comparison of DNA-binding across protein superfamilies**. *Proteins*, **78(1):** 52–62

8. Perez, A.G., Angarica, V.E., Collado-Vides, J., Vasconcelos, A.T. (2009). **From sequence to dynamics: The effects of transcription factor and polymerase concentration changes on activated and repressed promoters**. *BMC Mol. Biol.*, **10:** 92

9. Angarica, V.E., Perez, A.G., Vasconcelos, A.T., Collado-Vides, J., Contreras-Moreira, B. (2008). **Prediction of TF target sites based on atomistic models of protein-DNA complexes**. *BMC Bioinformatics*, **9:** 436

10. Lozada-Chavez, I., Angarica, V.E., Collado-Vides, J., Contreras-Moreira, B. (2008). **The role of DNA-binding specificity in the evolution of bacterial regulatory networks**. *J. Mol. Biol.*, **379(3):** 627–43

11. Gonzalez Perez, A.D., Gonzalez Gonzalez, E., Espinosa Angarica, V., Vasconcelos, A.T., Collado-Vides, J. (2008). **Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria**. *BMC Genomics*, **9:** 128

12. Lopez-Gomollon, S., Hernandez, J.A., Pellicer, S., Angarica, V.E., Peleato, M.L. & Fillat, M.F. (2007). **Cross-talk between iron and nitrogen regulatory networks in Anabaena (Nostoc) sp. PCC 7120: Identification of overlapping genes from/in FurA and NtcA regulons**. *J. Mol. Biol.*, **374(1):** 267–81

Articles Published While Working at Other Institutions Previous to Starting this PhD Thesis:

1. Perez, A.G., Angarica,V.E., Vasconcelos, A.T., Collado-Vides, J. (2007). **Tractor_DB (version 2.0): a database of regulatory interactions in gamma- proteobacterial genomes**. *Nucleic Acids Res.*, **35:** D132–6

2. Espinosa, V., Gonzalez, A.D., Vasconcelos, A.T., Huerta, A.M & Collado-Vides, J. (2005). **Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes**. *J. Mol. Biol.*, **354(1):** 184–99

3. Guia, M.H., Perez, A.G., Angarica, V.E., Vasconcelos, A.T., Collado-Vides, J. (2005). **Complementing computationally predicted regulatory sites in Tractor_DB using a pattern matching approach**. *In Silico Biol.*, **5(2):** 209–19

4. Gonzalez, A.D., Espinosa, V., Vasconcelos, A.T., Perez-Rueda, E., Collado- Vides, J. (2005). **TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes**. *Nucleic Acids Res.*, **33:** D98–102

Research Meetings Talks and Poster Presentations

- Angarica, V.E., Ventura, S. and Sancho, J. (2011). **Prediction of prion proteins in complete proteomes using probabilistic representations of prionogenic domains**. *XXXIVth Congress of the Spanish Society for Biochemistry and Molecular Biology*, Barcelona, Spain (***Poster Presentation***)

- Angarica, V.E., and Sancho, J. (2011). **Protein buried interfaces in protein stability and dynamics**. *XIst Congress of the Spanish Biophysical Society*, Murcia, Spain (***Oral Presentation***)

- Angarica, V.E., and Sancho, J. (2010). **A computational study of protein intrafaces: prediction of unstructured regions in protein folding intermediates**. *XXXIIIrd Congress of the Spanish Society for Biochemistry and Molecular Biology*, Córdoba, Spain (***Oral Presentation***)

- Angarica, V.E., and Sancho, J. (2010). **A computational study of protein intrafaces and their implication in protein folding**. *IVth Spanish-Portuguese Biophysical Congress*, Zaragoza, Spain (***Poster Presentation***)

- Angarica, V.E., and Sancho, J. (2010). **Intrafaces and cavities in protein structures: implications in protein folding and dynamics**. *VIth Meeting of the Spanish Protein Structure and Function Network*, Madrid, Spain (***Oral Presentation***)

- Angarica, V.E., Cuesta-Lopez, S., and Sancho, J. (2009). **Exploring the mutational space of the LDL-r module using molecular dynamics: connecting SNPs to abnormal phenotypes in a conformational disease**. *XXXIInd Congress of the Spanish Society for Biochemistry and Molecular Biology*, Oviedo, Spain (***Poster Presentation***)

- Contreras-Moreira, B., Angarica, V.E. (2009). **Comparison of DNA-binding across protein superfamilies**. *Congress BIFI 2009*, Zaragoza, Spain (***Oral Presentation***)

- Angarica, V.E., Cuesta-Lopez, S., Estrada, J. and Sancho, J. (2009). **Using Molecular Dynamics to study the conformational changes of the LDL-r LA5 module upon mutation**. *Congress BIFI 2009*, Zaragoza, Spain (***Oral Presentation***)

- Contreras-Moreira, B., Angarica, V.E. (2009). **3D-footprint: structural analysis of protein-DNA complexes**. *International Workshop Angel Ramirez Ortiz in memoriam*, Madrid, Spain (***Oral Presentation***)

- Angarica, V.E., Cuesta-Lopez, S., Estrada, J. and Sancho, J. (2009). **A computational study of the mutational space of the LDL-r LA5 module: in silico prediction of disease-like phenotypes in a conformational disease**. *International Workshop Angel Ramirez Ortiz in memoriam*, Madrid, Spain (***Poster Presentation***)

- Angarica, V.E., Cuesta-Lopez, S., Estrada, J. and Sancho, J. (2008). **Conformational Changes of the LDL-r LA5 Module Upon Mutation: a Computational Approach**. *XXXIst Congress of the Spanish Society for Biochemistry and Molecular Biology*, Bilbao, Spain (***Poster Presentation***)

- Contreras-Moreira, B., Lozada-Chavez, I. and Angarica, V.E. (2008). **Structural (and sequence-based) analysis of transcriptional regulation**. *VIIIth Spanish Symposium on Bioinformatics and Computational Biology*, Valencia, Spain (***Oral Presentation***)

- Angarica, V.E., Perez, A.G., Vasconcelos, A.T.R, Collado-Vides, J., Contreras-Moreira, B. (2007). **Prediction of Transcription Factor Binding Sites using Structural Information**. *XXXth Congress of the Spanish Society for Biochemistry and Molecular Biology*, Málaga, Spain (***Poster Presentation***)

Grants

- **2012:** Grant from the “Consejo Superior de Investigaciones Científicas” (CSIC) for a 3-month research stay at the Molecular Recognition & Bioinformatics Group, Institute for Research in Biomedicine (IRB), Barcelona, Spain

- **2009–2013:** Ph.D. grant from the "Consejo Superior de Investigaciones Científicas" (CSIC), JAE Program

- **2007–2009:** M.A.S. Grant from the Banco Santander Central Hispano, Fundación Carolina and Universidad de Zaragoza

- **2006:** Grant from the Iberoamerican Bioinformatics Network (Red Iberoamericana de Bioinformática, RIBIOrt-VII.L CYTED) for a 6-month research stay at the Center for Genomic Sciences (CCG), Cuernavaca, México

- **2005:** Travel Award from the Iberoamerican Bioinformatics Network (Red Iberoamericana de Bioinformática, RIBIO VII.L CYTED) to assist to a 3-month research stay at the National Laboratory for Scientific Computing (LNCC), Petrópolis, Brasil

Keyword Index