Physical constraints on protein structure evolution

Dissertation submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

Cédric Debès

born in: Libourne, France

Oral-examination: July 2013

Referees:

Prof. Dr. Robert Russell

Dr. Frauke Gräter

Zusammenfasung

Die Natur wartet mit einer enormen Vielfalt an dreidimensionalen Proteinstrukturen auf, von denen vermutlich jede für ihre spezifische Funktion optimiert ist. Ein wichtiges Ziel in der Biologie ist es, die treibenden evolutionären Kräfte für die Entdeckung und Optimierung neuer Faltungen von Proteinen zu entdecken. Eine langjährige Hypothese ist, dass die Evolution von Proteinfaltungen bestimmten Randbedingungen gehorcht. Mit dem Ziel der Aufklärung dieser äußeren Bedingungen, die die Entwicklung neuer Proteinfaltungen einschränken, werteten wir einige physikalische Größen wie Flexibilität, Faltbarkeit und Festigkeit für eine Vielzahl von Proteinen aus. Als erstes wurde die Flexibilität mit zwei unabhängigen Methoden abgeschätzt: mit CONCOORD, welches Konformationensembles für atomare Proteinstrukturen durch geometrische Parameter prognostiziert sowie mittels vereinfachter elastischer Netzwerkmodelle. Das Faltungsverhalten wurde durch die sogenannte "contact order" gemessen. Diese kann die Faltungsgeschwindigkeit eines Proteins durch die Messung des Abstands zwischen nativen Kontakten innerhalb des Proteins vorhersagen. Schließlich wurde die mechanische Festigkeit mit Langevin-Dynamik-Simulationen von herkömmlichen Go-Typ-Modellen von Proteinen unter Anwendung von externer Kraft abgeschätzt. Diese grobkörnigen Modelle sind von der Röntgenkristallstruktur abgeleitet. Wir berechneten diese drei physikalische Größen für jede bekannte Proteinstruktur, und bildeten diese auf einem phylogenomischen Baum von ca 3.000 Proteinfamilien ab. Bimodale Trends wurden für die verschiedenen physikalischen Größen beobachtet und deuten auf eine Trendwende in der Proteinevolution vor rund ∼1,5 Milliarden Jahren hin. Diese Wende geht mit einem plötzlichen Erscheinen vieler neuer Proteinstrukturen einher ("big bang") und entspricht dem Erscheinen von vielzel-

ligen Organismen, welche durch veränderte Randbedinungen die Evolution von Proteinstrukturen drastisch verändert haben könnten. Genauer gesagt beobachten wir vor ∼1,5 Milliarden Jahren einen Anstieg der Faltbarkeit und eine Abnahme der mechanischen Stabilität, welche vermutlich das Ergebnis einer Notwendigkeit für schnell und kompakt faltende Proteine aufgrund molekularer Kompartimentierung sind, d.h. dem Erscheinen von Zellen. Im Gegensatz dazu beobachten wir nach ∼1,5 Milliarden Jahren einen Rückgang von Faltbarkeit und eine Erhöhung der mechanischen Stabilität, was auf die Notwendigkeit von mechanischer Stabilität hindeutet. Dieser Trend ist wahrscheinlich auf den Aufstieg von mehrzelligen Organismen mit erhöhten mechanischen Belastungen zwischen den Zellen zurückzuführen. Der Verlust von Faltbarkeit nach dem ”big bang” könnte darin begründet sein, dass Zellen begannen, Proteine wie Chaperone oder andere fortschrittliche Mechanismen zu verwenden, die die Notwendigkeit zur schnellen Faltbarkeit abgeschwächt haben könnten.

Zusammengefasst haben wir in dieser Arbeit physikalische Randbedingungen analysiert, die wahrscheinlich eine Rolle in der Entwicklung von Protein-Strukturen spielen. Unser globaler Ansatz eröffnet Wege für eine umfassendere Analyse von verfügbaren genomischen und strukturellen Daten. Diese neue Sicht auf die Evolution von Proteinstrukturen erlaubt uns, bessere Einblicke in deren Arbeitsweise und Funktion zu bringen. Darüber hinaus kann unser Ansatz helfen, eine netzwerkbasierte Ansicht der Evolution von Proteinstrukturen aufzubauen, um die Klassifizierung der heute bekannten vielfältigen Proteinstrukturen zu verbessern und neue Proteinstrukturen zu entwerfen.

Abstract

Nature has come up with an enormous variety of protein three-dimensional structures, each of which is thought to be optimized for its specific function. A fundamental biological endeavor is to uncover the evolutionary driving forces for discovering and optimizing new folds. A long-standing hypothesis is that fold evolution obeys constraints. Aiming at elucidating those constraints, we evaluated some physical quantities for a large number of biological molecules. Firstly, flexibility was estimated via two independent methods: CONCOORD, which predicts conformational ensembles for atomic protein structures using geometrical constraints, and elastic network models, a simple coarse-grain model. Foldability was measured by Contact Order, which can predict the folding rate of a protein by measuring the distance between native contacts within the protein. Lastly, mechanical strength was predicted with Langevin Dynamics simulations of the conventional Go-type models of proteins, a coarse-grained model based on the X-ray structure, under force. We mapped those physical quantities onto a phylogenomic tree of protein structures resulting from the analysis of the abundance of ∼3,000 protein families. Bimodal trends were observed for the different physical quantities suggesting a turnover at around ∼1.5 billions years ago. This turnover corresponds to the apparition of multicellular organism that could have drastically modified the constraints applied on the evolution of protein structures. More specifically, before ∼1.5 Gya, we observed an increase of foldability and a decrease of mechanical stability that might be the result of a concerted need for fast folders and compact proteins resulting from molecular compartimentalization, i.e. the rise of cells. On the contrary, after ∼1.5 Gya, we observed a decrease of foldability and an increase of mechanical stability that suggest a need for mechanical stability probably related to the rise of multicellular organisms with increased mechanical stresses between cells. The loss in foldability after the big bang might be due to that cells started to make use of proteins such as chaperones or other advanced mechanisms thereby removing, at least partly, the constraint for fast folders.

Taken together, we identified physical constraints that are likely to play a role in the evolution of protein structures. Our global approach opens avenues for a more comprehensive analysis of genomic and structural data available. Improving our view on protein structure evolution is likely to bring more insights into their functioning. Additionally, it could help constructing a network based view of protein structures evolution improving the classification of the known protein catalogue and aiding the design of new protein structures.

Acknowledgments

It would not have been possible to write this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here. The author wishes to express his gratitude to his supervisor, Dr. Frauke Gräter, who was abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to the members of the supervisory committee: Prof. Dr. Rob Russell and Prof. Dr. Ulrich Schwarz. Without their knowledge and assistance this study would not have been successful. I thank Prof. Dr. Harald Herrmann-Lerdon for kindly accepting to be part of the thesis committee. I would like to thank the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences for offering such rich interdisciplinary environment. I thank Dr. Agnieszka Bronowska to kindly accept to be part of my HGS Math-Comp committee, and for her help and support through my time in Heidelberg. I thank Prof. Dr. Gustavo Caetano-Anollès, Dr. Minglei Wang, and others lab members in Urbana to welcome me in their lab and for a fabulous collaboration on "Evolutionary optimization of protein folding". I thank all previous and present Gräter group members for making the group a wonderful place over the years.

I owe a lot to my parents, who encouraged and helped me at every stage of my personal and academic life, and longed to see this achievement come true. For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

谨以这篇论文献于我的女友—张晓敏。感谢她一直以来对我的爱与支持。

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Today, it is generally accepted that proteins appeared on earth between ∼3.9 to ∼3.5 billion years ago. A common theory that could explain the synthesis of the first protein is abiogenesis [1]. In short, primitive earth conditions allow the spontaneous formation of organic molecules. As the result of further transformations, organic compounds assembled into polymers. It is still unclear, however, how protein structures formed from those primitive polymers. Some studies suggested that those polymers first adopted some favorable configurations. Hence, first protein structures could have formed by combination of favorable polypeptide fragments [2]. The way proteins fold into such fragments is encoded into their amino-acid sequences. The process of folding is complex, physical interactions between amino-acids drive polymers to a stable conformation with biological activity. In the course of evolution, in order to evolve from the folding of small polypeptides of only tens of atoms to the folding of a giant molecular machinery of million atoms, nature must have selected protein architectures. Thus, physics and evolution influenced the protein world we can observe nowadays. In this thesis, we would like to understand some of the physical components that impacted protein structure evolution.

The recent accumulation of sequenced genomes is now enabling us to search for evolutionary traces of protein structures. The genomic era makes use of the massive amount of data becoming available combined with new computational methodologies, in order to extract knowledge from genomic data [3, 4]. Furthermore, structural information gathered by experimental techniques such as X-ray crystallography or Nuclear Magnetic Resonance (NMR) can be mapped onto the genomic data to infer evolutionary relationships [5]. To this end, we decided to work at a specific level of protein organization: the domain. Proteins can be divided into domains, which are independent structural and evolutionary units. It was shown that duplications and combinations of different domains are the major evolutionary processes in the acquisition of novel functions [6–8]. Occasionally, domains can also evolve by random mutations, slowly changing their structures inside the fold structure space. Taken together, those evolutionary processes have led to the apparition of protein structure groups sharing sequence or structural similarities. Aiming at characterizing groups of domains according to their similarities, classifications of proteins (SCOP, CATH) were developed [9, 10], allowing further investigations on structural and evolutionary links between domains. Using structural comparison methods, maps of protein structures unveiled a continuous view of the protein universe [11–14], or helped revealing possible paths of evolution between protein structures. In an effort to improve the description of such phylogeny, methods based on sequence alignments or, more recently, whole genome features are now commonly used. These studies lead to a view of the protein universe as a continuous space with bridges inside and within different structural families, thus reinforcing the theory of a divergent evolution of the protein universe. An area of study that would help uncover how the protein universe expanded relates to the factors influencing the selection of structures. A wide range of factors can influence selection, including the genomic position of the encoding genes, expression patterns, the position in biological networks (e.g. high level of contacts), and physical constraints. These factors are thought to influence the evolutionary rate and patterns in proteins, particularly the formation as well as the extinction of protein domain families. Underlying genomic mechanisms responsible for the dynamics of protein family populations are diverse. Duplication, mutation, or recombination of genetic material can lead to jumps between structural spaces, divergence from a structural space, or extinction of a structural space [15,16]. Taken together, all those mechanisms can help to understand how protein structures evolved. This thesis aims at identifying the influence of some physical properties on protein structure evolution. It is divided into two parts. In the first introductory part (Chapters 2-4), we describe concepts and methods used to build our map of physical constraints onto evolutionary history.

In Chapter 2, we present concepts and methods related to the evolution of protein structure. In short, examining protein structures in combination with genomic features allowed to create a tree of protein architectures. This tree unveiled the early history of proteins [17], planet oxygenation [18], and the dynamics of domain organization in proteins [19]. In Chapter 3, we present the fundamental physical principles responsible for the diversity of protein shapes observed. Additionally, we present the dataset of protein structures used for this study. This dataset covers most of the observed folds until now. Interestingly, the number of shapes adopted by proteins is rather limited, suggesting that natural selection played a role in the evolution of protein structures. Aiming at studying the mechanisms underlying this selection, in Chapter 4, we describe how we evaluated physical quantities for our dataset of protein structures. Physical constraints such as folding time, flexibility, and mechanical stability reflect the involvement and function of a protein within its biological environment. By measuring those properties for a set of proteins, one can correlate their composition, topology, or folding propensity with their mechanical performance, and deduce the nature of stress to which the protein is subjected. This result allows classifying proteins depending on their mechanical properties, and therefore helps to identify of protein architectures with outstanding mechanical properties. Those proteins could be used as templates for biomedical or industrial purposes.

In the second part comprising the results of this thesis, we study the impact of physical constraints such as folding rates, structural flexibility, or mechanical stability onto the evolution of protein structure. In Chapter 5, we aim at testing if folding rates changed during the evolution of protein structures, and may have constituted an evolutive pressure. Similarly, in Chapter 6, we try to understand how mechanical stability impacted the protein universe. We want to test if mechanical stability was acquired late in evolution, only when multicellular organisms, active transport, and motion developed. Understanding how the protein universe expanded under effects of physical constraints is a first step toward an understanding of how the protein universe was shaped under evolutive pressure. We discuss our results from the view of a relation of the physical features with each other (Chapter 7), with structural features of protein domains and their evolution, in an attempt to link genomics and protein biophysics.

Part I

Physical constraints mapped onto protein structure evolution

Chapter 2

How have protein structures evolved?

Despite our knowledge about nowadays proteins, little is known on their origin and evolutionary history. Theory and experiments suggest that the first proteins originated from short polypeptides arising under specific conditions. Yet, the different steps of molecular evolution of proteins are still unclear. Information gathered from nowadays genomes on the distribution of protein structures could complete our knowledge on the origin and evolution of protein structures, allowing to explain how proteins evolved to the current catalog of existing proteins. In this chapter, we will review the different concepts and computational methods utilized to reconstruct protein structure evolution.

2.1 Principles of molecular evolution

Evolution is a general concept describing a gradual change of a system. In biology, the system of interest is generally a population of organisms or molecules (DNA, RNA, proteins). In this thesis, the system considered is a population of proteins that evolved gradually through mutations, occasionally leading to the apparition or extinction of protein families. In the following sections, we describe the principles of molecular evolution.

2.1.1 Mutations

Mutations represent changes in the genetic material (DNA or RNA coding for proteins). Mutations can affect single or multiple amino-acid depending on the mutation type: copying errors during cell division, exposure to ultraviolet or ionizing radiation, chemicals or viruses. Mutations can affect protein structures in many ways:

- Point mutations correspond to a change of one nucleotide for another. At the protein level they can result (i) in the same amino-acid (silent mutations), without an impact on protein structure, (ii) in a different amino-acid that possesses similar chemical properties as the mutated one (neutral mutation) or different properties (missense mutation), or (iii) in a mutation that codes for a stop, which can lead to a truncated protein (nonsense mutation).

- Insertions represent an addition of one or more extra nucleotides, occasionally resulting in a truncated protein structure (frameshift mutation) by a reading frame shift.

- Deletions remove one or more nucleotides. Similarly to an insertion, deletions can lead to truncated protein structures.

The effect of the mutation on the gene product can be harmful or beneficial depending on the location and nature of the change. Mutations beneficial to the protein function do not seem to occur often (adaptive mutations) constraining proteins to a certain topology [20–23]. However, gene duplication (Section 2.1.2) offer a higher chance of modifying the topology, possibly leading to a change of protein function, as the new duplicate is likely to be under a smaller constraint. Additionally, in a few cases duplication can lead to the insertion of a domain within another domain [24] possibly giving rise to a new type of fold.

2.1.2 Gene duplication

Gene duplication results in an expansion of the genetic material [25]. It may occur during cell meiosis when homologuous chromosomes bind to each together. At this point, exchange of genetic material randomly takes place and may lead to a duplication genetic information in one of the homologuous chromosomes, consequently removing the genetic portion from the other chromosomes. Gene duplication can also result from other events such as retro-transposition events. Transposons are transcripted into RNA and can copy themselves back into the DNA sequence leading to an amplification of genetic material. Gene duplication is often considered as one of the main drivers of evolution [26]. Consequently, evolutionary dynamics of domains might be mostly driven by gene duplication, divergence, and elimination. All together those mechanisms are the basis of “Birth death innovation models” [27,28]. In this thesis, we consider a “Birth death innovation” model where gene duplication expands the protein repertoire linearly (Section 2.2). In the following section, we describe patterns of evolution that result from mutation or gene duplication events.

2.1.3 Patterns of Evolution

Evolution can follow three major patterns, convergent, analoguous, or divergent evolution of the system as specified below.

- Convergent: Proteins evolve simultaneously toward the same function without a common ancestor. It has been reported as a rare event [29].

- Analoguous: Proteins evolve simultaneously toward the same function without a common ancestor and without sequence similarity.

- Divergent: Proteins exhibit a similar function and structure, and derive from a common ancestor, but evolve into a separate function and structure.

The divergent evolution results in homologous proteins referring to their degree of similarity as well as their ancestry relationships. They can result from two events, speciation when a species diverges into two separate species resulting in orthologous sequences, or duplication when a gene is duplicated in one species resulting in paraloguous sequences which subsequently can evolve separately.

2.1.4 Molecular clock

A molecular clock is an evaluation of the elapsed time between events of in an evolutionary model. Geological history is coupled to rates of molecular changes in order to assess the occurrence of divergence, or extinction events. The method originated from the observation of a linear increase in amino-acid mutations in hemoglobin by Emile Zuckerkandl and Linus Pauling [30]. Later, E. Margoliash observed similar patterns in the evolution of Cytochrome C and formulated the term genetic equidistance. However, the model of genetic equidistance is debated with regard to five factors that limit its applicability [31]:

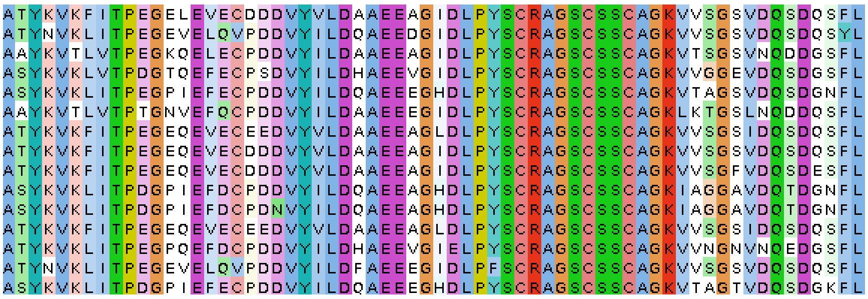
- changes in generation times (if the rate of new mutations depends at least partly on the number of generations rather than the evolutionary time)
- population size (genetic drift is stronger in small populations, so that more mutations are effectively neutral) [32]
- species-specific differences (due to different metabolism, environment, evolutionary history, or others)
- change in function of the protein studied
- change in the intensity of natural selection

Notwithstanding these crucial concerns, a molecular clock based on changes in protein structure abundance calibrated using fossils and crucial events in evolution allowed to obtain a time line of protein structural evolution and was used in this thesis (Section 2.2).

2.1.5 Common computational methods used for phylogeny

Multiple sequence alignments

A multiple sequence alignment consists of three or more biological sequences, which are aligned according to a scoring scheme in order to find the best match between them (Figure 2.1). Alignments are generally conducted on homologous sequences in order to identify conserved amino-acids that usually represent positions or regions which are key to the function, structure, or evolution of the protein of interest. Sequence alignment methods match amino-acids according to a degree of identity related to their chemical properties and the evolutionary probability of mutation between then (substitution matrix), and uses gaps to optimize the alignment of similar or identical amino-acids. Dynamic programming is used to find the globally optimal alignment solution, which consists in the lowest score being calculated from the sum of all of the pairs of characters at each position in the



```
ATYKVKFITPEGELEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSFL
ATYNVKLITPEGEVELQVPDDVYILDQAEEDGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSYL
AAYKVTLVTPEGKQELECPDDVYILDAAEEAGIDLPYSCRAGSCSSCAGKVTSGSVNQDDGSFL
ASYKVKLVTPDGTQEFECPSDVYILDHAEEVGIDLPYSCRAGSCSSCAGKVVGGEVDQSDGSFL
ASYKVKLITPEGPIEFECPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGSVDQSDGNFL
AAYKVTLVTPTGNVEFQCPDDVYILDAAEEEGIDLPYSCRAGSCSSCAGKLKTGSLNQDDQSFL
ATYKVKFITPEGEQEVECEEDVYVLDAAEEAGLDLPYSCRAGSCSSCAGKVVSGSIDQSDQSFL
ATYKVKFITPEGEQEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSFL
ATYKVKFITPEGEQEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGFVDQSDESFL
ASYKVKLITPDGPIEFDCPDDVYILDQAEEAGHDLPYSCRAGSCSSCAGKIAGGAVDQTDGNFL
ASYKVKLITPDGPIEFDCPDNVYILDQAEEAGHDLPYSCRAGSCSSCAGKIAGGAVDQTDGNFL
ATYKVKFITPEGEQEVECEEDVYVLDAAEEAGLDLPYSCRAGSCSSCAGKVVSGSIDQSDQSFL
ATYKVKLITPEGPQEFDCPDDVYILDHAEEVGIELPYSCRAGSCSSCAGKVVNGNVNQEDGSFL
ATYNVKLITPEGEVELQVPDDVYILDFAEEEGIDLPFSCRAGSCSSCAGKVVSGSVDQSDQSFL
ASYKVKLITPDGPIEFECPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGKFL
```

Figure 2.1: Example of a multiple sequence alignment using Jalview [33]. Amino-acids are colored according to their type.

alignment. Such score can be used as evolutionary distance, consequently allowing the reconstruction of a phylogenetic tree using methods presented in Section 2.1.6. Typically, aligned amino-acids also share the same position in a structural alignment of homologuous structures. A multiple sequence alignments can be based on Hidden Markov Models (HMM, see below) enabling the detection of more distant evolutionary relations which are relevant for classification or sequence assignment of homologous protein.

Hidden Markov models

Hidden Markov models (HMMs) are sophisticated and powerful statistical models allowing the assignment of protein sequences to their respective family [34]. They use family multiple sequence alignments (MSA) to build profiles based on insertion, deletion, and transition probabilities (i.e. the likelihood that one particular amino acid follows another particular amino acid). HMMs are composed of several layers (Figure 2.2), each of which represents one position in the sequence.

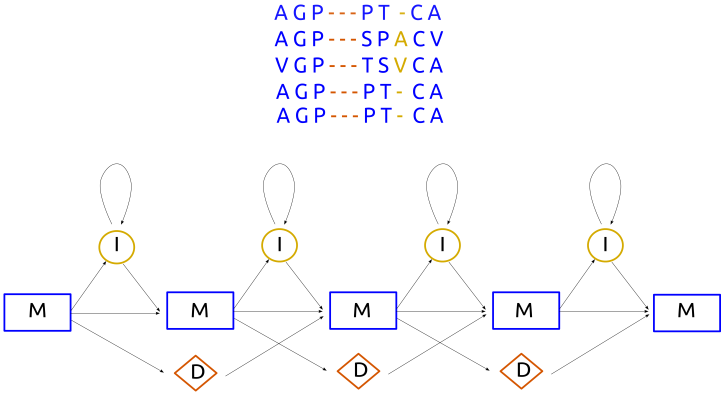


Figure 2.2: Hidden Markov Models are modeling each position of a multiple sequence alignment as a state: Matched (M), Inserted (I), Deleted (D). The states possess frequencies of amino-acids extracted from the MSA, that are used to produce a signature of a given alignment.

At each step, according to the frequency obtained from the MSA and the state in the model (insertion or match) a residue type is added to the signature sequence. The resulting sequence profile corresponds to a given family and can be used to help the assignment of other sequences.

2.1.6 Phylogeny reconstruction

Phylogenetics describe the evolution of a population of proteins via a tree, where branches can diverge, converge, or terminate corresponding, respectively, to the apparition, the parallel evolution, or the extinction of a protein. In order to infer the evolution between proteins, a number of different computational methods have been developed, building trees on the basis of similarities and differences of protein sequences, structures, or genomic distribution.

Reconstruction based on distance

One class of computational techniques for tree reconstruction considers distances between sequences as the main ingredient. The different methods take as input a genetic distance that can be calculated from a multiple sequence alignment (Section 2.1.5). We shortly describe two major methods that are based on distance matrices of an MSA.

1. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) regroups sequences according to distance values. The algorithm associates sequences starting from the closest to the most distant. Sequences are merged into clusters at each step. Therefore, the distance taken into consideration for the next step is an average distance between two merged sequences. Additionally, UPGMA considers a constant rate of evolution consistent with a molecular clock (Section 2.1.4), such that the same evolutionary time is considered for every branch of the resulting tree.

2. The Neighbor Joining method clusters a set of taxa (e.g. species, sequences) on the basis of the distance between each pair. In contrast to UPGMA, Neighbor joining uses a modified matrix, hereby agglomerating information of all pairs. Subsequently, the rate of evolution is modified allowing insight into the evolutionary distance between divergence events.

Reconstruction based on character

Tree reconstruction techniques using character as input again require a multiple sequence alignment, from which they, however, deduce characters instead of distances. A character corresponds to an attribute that varies between sequences, organisms, or genomic features as used in this thesis, see Section 2.2. Hence, characters represent the evolution of heritable variation. Each character can have two or more discrete states. For instance, the character "hair color" might have the states "brown" and "black", or the character "weight" might have states on a 0-10 scale coded from the distribution of measured weights. When a character exhibits more than two discrete states, it can be treated as unordered or ordered. Unordered characters have an equal "cost" (in terms of number of the "evolutionary events") to change from any one state to any other; complementary, they do not require passing through intermediate states. On the contrary, ordered characters follow a sequence that must occur through specific intermediates. Thus, the cost of evolutionary variation is to be considered between different pairs of states.

Maximum parsimony

Trees generated by maximum parsimony [35] are optimized toward a minimum of the total number of changes. More precisely, the most parsimonious tree is the preferred hypothesis of relationships among taxa, sequences, or protein architectures. A simple algorithm determines how many evolutionary transitions are required to explain the distribution of each character. As the result of the high number of evolutionary transitions, plenty of possible phylogenetic trees can be produced from the same dataset. Therefore, most algorithms attempt to increase the tree score by applying perturbations onto it until convergence of the score is achieved. Maximum parsimony was used for the construction of protein structure trees within this thesis (Section 2.2), as we assume a linear rate of evolution between lineages and characters.

Maximum likelihood

The maximum likelihood method, similar to maximum parsimony, requires a substitution model to assess the probability of particular mutations. In short, the number of mutations possible between node is constrained to obtain the best tree. In contrast to maximum parsimony, maximum likelihood permits varying rates of evolution between lineages and characters. As a result, maximum likelihood is well suited for the analysis of distantly related sequences, for which the total number of possible tree topologies and the branch length is high. In order to reduce the search space for the optimal tree, algorithms such as the pruning algorithm are used.

Bayesian inference

Bayesian inference assumes a predefined probability distribution for possible trees. This distribution can consider a more sophisticated estimate that takes into account a stochastic process for divergence events. Similar to maximum likelihood, Bayesian inference decreases the changes required between nodes and leaves of a given tree.

2.2 Phylogenomic tree of protein families

Comparative genomics approaches to study similarities of organisms have been developed recently. They rely on genomic data as fossils to reconstruct trees over a large evolutionary period. Genomic structural information was first used by Gerstein et al [36, 37], at a time when only one species from each of the three superkingdoms was sequenced. Phylogenomic trees based on structural information, which is more conserved than sequence information, allow to track the evolutionary link between distant proteins [38, 39]. Protein domains are central units of protein organization and evolution, as the duplication and shuffling of domains are fundamental for diversity [6, 7]. This section describes the work flow used to generate a phylogenomic tree of protein fold architecture [5].

The character-based reconstruction of protein structure tree uses data from more than 1000 genomes (Figure 2.3), for which the presence and abundance of protein structures was obtain as follows. Three-dimensional structures of protein domains were matched via HMMs (Section 2.1.5, Figure 2.3) to more than 60 % of the open reading frames in those complete genome sequences. This census of protein architectures results in abundance values for each superfamily or family (for SCOP classifications of superfamilies or families, see Section 3.6.1). Abundance values together with the presence of a superfamily or family in each genome constitute the basis of the method. Abundance values scale from zero to thousands. Matrices were constructed from the abundance values (G) of domains at different levels of the SCOP classification.

The abundance is coded as a character (Section 2.1.6) as follows: each level of abundance corresponds to one states (standardized between 0 to 20 scale), and state are linearly ordered. Thus, the model assumes a constant successive addition of homologous genes that leads to an increase of

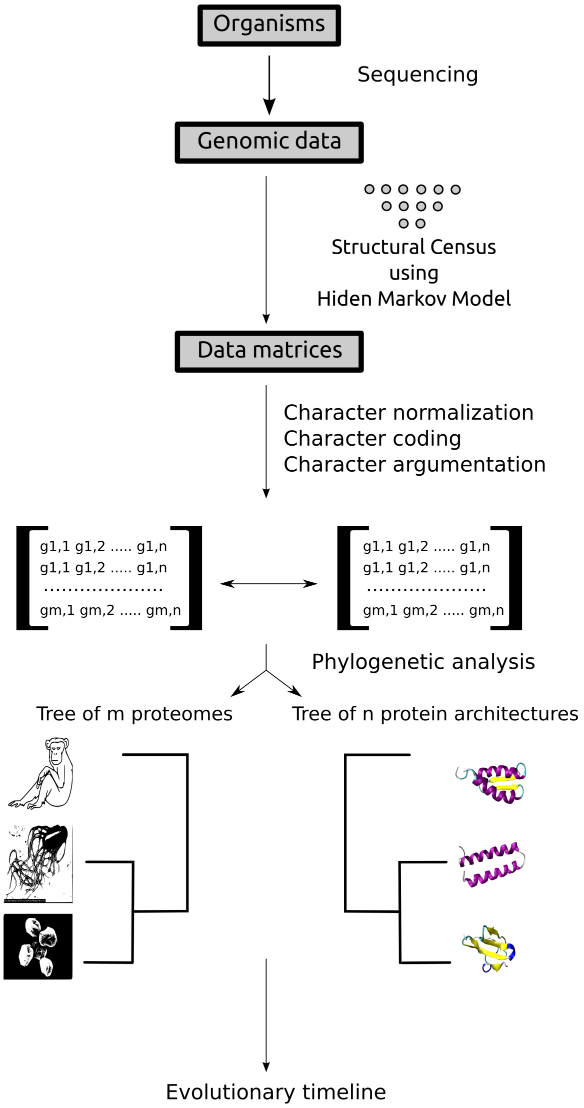


Figure 2.3: Workflow describing the different steps of the phylogenomic reconstruction of proteome trees and protein structure trees.

the given population. Therefore, families that appeared earlier in evolution are prominent in many genomes. In this model, duplication is considered to occur more frequently than gene loss leading to an increase of the gene family copy number. Consequently, ancient structures are more abundant and more widely present than younger ones. This model follow principles such as preferential attachement [40] where large domain families are more prone to expand compared to small domain families.

Phylogenetic trees of protein architectures were reconstructed from the abundance of protein structures in genomes as characters using maximum parsimony (Section 2.1.6). The data matrix can be used to construct either proteome trees or domain structures trees. Using a molecular clock to map events in protein structure evolution to molecular fossils, a domain structure tree (Figure 2.4) describing the evolution of protein families over ∼3.8 billions years was obtained.

Relative evolutionary ages are mapped according to a node distance (nd). A node distance is calculated by counting internal nodes along a lineage from the root to a terminal node (a leaf) of the tree. Hence, nd measures evolutionary speciation between protein structures with the most ancestral taxon having 0 as nd, and the most recent one 1.

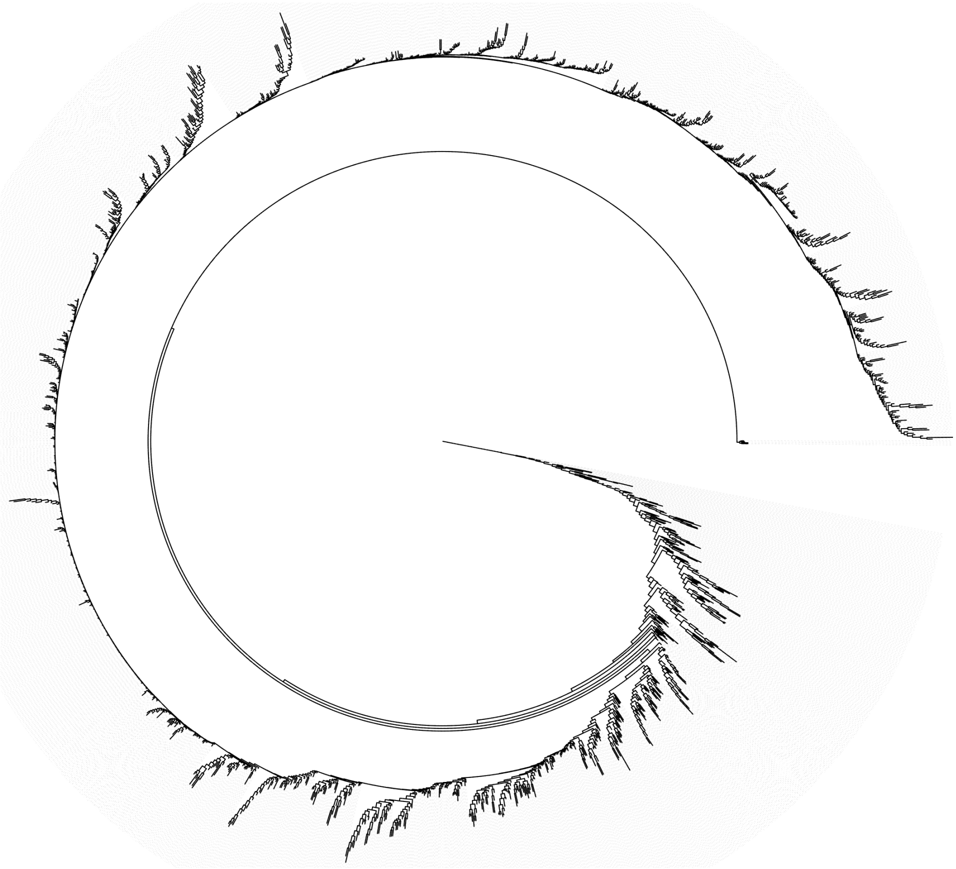


Figure 2.4: Phylogenomic tree of protein architectures at the family level of SCOP organization (Section 3.6.1) [41].

Chapter 3

What are the determinants of protein structure?

Encoded by the amino-acid sequence, proteins can adopt a wide variety of structures. In this chapter, we will discuss the determinants of the shape of protein structures. More precisely, we are interested in the underlying physical mechanisms that influence shapes adopted by proteins. We note that some proteins such as intrinsically disordered protein transiently adopt various configurations to fulfill their functions. Hence, they do not possess any defined shape. This implies that structure is not always required for a protein to fulfill its role. In this thesis, we focus on ordered, well structured globular proteins.

3.1 The geometry and size of proteins

Proteins are polymers formed from a mix of 20 different monomers called amino-acids. An important factor that distinguishes proteins from one another is their specific three-dimensional structure, which is determined only by the sequence of the monomers themselves. These monomers or amino-acids assemble via the formation of a peptide bond leading to a polypeptide. Amino-acids in proteins are referred to as residues. Their side chains are responsible for the structure of the protein as they represent the only variation, while the other part, referred to as backbone, is shared by all aminoacids. This backbone possesses a certain flexibility conferred by the rotation of two flanking bonds of a peptide link. Protein structures can be classified in different categories according to their general shape (Figure 3.1):

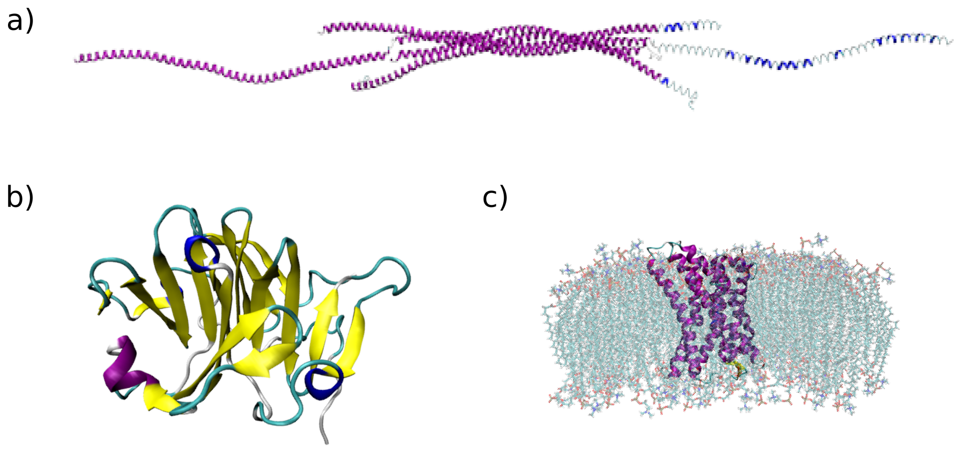


Figure 3.1: Proteins feature a high structural and functional variety. Three main categories are, from top to bottom; a) fibrous, b) globular, and c) membrane proteins.

- Fibrous proteins (Figure 3.1a) are long (generally less than 1000 residues) filamentous or fibrous proteins in the shape of a rod or a wire. They are found in two forms, namely either a helix composed of repetitive motifs, or a string of domains with higher sequence variations. They usually fulfill structural or storage functions. Examples are keratins, collagens, and elastins.

- Globular proteins (Figure 3.1b) are compact spherical-shaped proteins. Their size ranges from a hundred to several hundred residues. They can act as enzymes, messengers, transporters, regulatory, and structural proteins. As a consequence of the variety of functions that globular proteins can handle, a variety of structures is adopted. Hence, globular proteins are naturally relevant for studying structure function space relationships.

- Membrane proteins (Figure 3.1c) are a class of proteins that interact with a membrane. Their shape and size is closely related to those of globular proteins, but they predominantly adopt bundle or barrel shapes when present inside a membrane. Additionally, they form large protein complexes often featuring a high structural flexibility. Due to their preference of a hydrophobic environment instead of water as a solvent, like other globular proteins, membrane proteins pose a challenge for structure determination.

3.2 Hydrophobic core and secondary structure

The structure of a protein is generally acquired during a physical process called folding, during which the protein hydrophobic residues turn towards the inside of the structure, whereas hydrophilic residues turn outside towards the aqueous solvent or membrane environment. The formation of weak, non-covalent interactions occurs inside the protein leading to a hydrophobic core. Other non-covalent interactions within the core and at the protein surface are hydrogen bonds and salt bridges between polar atoms of sidechains and backbone. Backbone hydrogen bonds form regular patterns of mainly two types [42]. When the carbonyl group of residue i is connected to the amide group of residue i+4 repetitively, an α-helix is formed. When a ladder of hydrogen bonds is formed between two segments of the backbone chain, a β-sheet is formed [43]. Physical properties of proteins differ due to the relative composition of the two main secondary structures. For instance, α-helices are known to confer a higher flexibility, while β-sheet can increase stability. Consequently, the composition of protein secondary structure is relevant for the classification of proteins (Section 3.6). In the following section, we present how proteins can be grouped according to their secondary structure.

3.3 Structural classes

From the two secondary structure combinations, proteins fall into three major structural classes: (1) all-α, (2) α combined with β, and (3) all-β proteins [44].

- All-α proteins: The high content of helices makes this class rich in inter-atomic contacts. They are on average of a relatively small size due to the fact that α-helices can form in smaller segment.

- All-β proteins: The arrangement of β-strands in an anti-parallel or parallel fashion gives a relatively high rigidity. The variation in the number of β-strands and β-sheets and their orientation is the source of the diversity of this class.

- α-β proteins: The α-β class can be split into two categories with either clearly separated α and β (α+β) or with mixed α and β (α/β). A well-studied example of the latter category is the TIM-barrel, an α/β topology that several unrelated proteins possess.

3.4 Protein topology

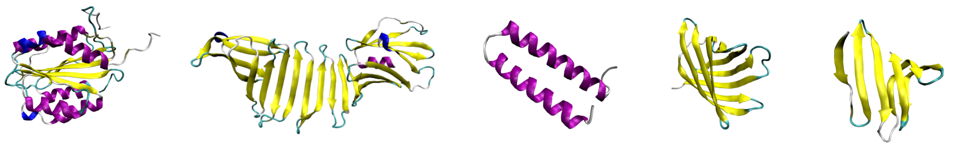


Figure 3.2: Structural motifs from left to right : β-α-β units, β-meander, α-α unit, β-barrel, Greek key.

A protein topology refers to the tree-dimensional path of the amino-acid chain leading to the network of interactions responsible for the general shape of a given fold. Some typical structural topologies (Figure 3.2) are found with high frequency among the protein structures known to date, such as the β-α-β unit, β-meander, α-α unit, β-barrel, and Greek key.

Similar structural motifs do not necessarily involve a similarity in the amino-acid sequence of the protein chain. On the contrary, they represent favorable physical conformations that are found for several unrelated sequences. They constitute repeating units that are basis of numerous protein structures. The fact that structural topologies are more conserved than sequences rationalizes the phylogenetic approach based on structure instead of sequence (Section 2.2). The next section presents protein domains as a type of repeating units of structural topologies central for protein evolution and organization.

3.5 The protein domain: A fundamental unit of organization

As seen above, proteins possess different levels of organization. The primary structure corresponds to the sequence of amino-acids, which in turn fold into a secondary structure. Secondary structures pack into a topology corresponding to a tertiary structure. If a protein molecule is formed from more than one polypeptide chain, the complete structure is designated as the quaternary structure.

A level of organization embedded into the tertiary structure and particularly relevant for evolutionary analyses is the domain. Domains are independent evolutionary and folding units that can be detected using conformation, function, or sequence similarity. Studies using these similarity measures revealed that the same domain can be present in different proteins [45–47]. Thus, domains act as modules that can be combined to form new proteins. Their occurrence in various proteins is due to genomic mechanisms such as gene duplication, resulting in an independent evolution of one copy toward a new function. Thus, successive duplications lead to an increase of domain populations resulting in protein families, with each family member having a sequence and a three-dimensional structure that resemble those of the other family members. Taken together, domains are a unit of organization relevant for the classification of proteins as they represent unique modular units that can be identified and classified. Domains also were used as structural units throughout all analyses presented in this thesis. In the next section, we review two different databases of protein structure classification, both of which are based on domains.

3.6 Classifications of proteins

Structural classifications of proteins group domains according to their topologies, structures, and sequences similarities. Several classification schemes have been developed. Here, we present two main classification databases that share a hierarchical organization but differ with respect to the definition of domains and the methods used to group protein structures (e.g manual, semi-automatic, or fully-automatic methods).

3.6.1 SCOP

The structural classification of proteins (SCOP) [48] defines four hierarchical levels (Figure 3.3) :

- Class (C): types of folds according to secondary structure (Section 3.2)
- Fold (FO): according to the general arrangement of secondary structure
- Superfamily (SF): based on structure similarities
- Family (F): based on sequence similarities

SCOP is essentially manually annotated. Hence, similarities are partly detected on the basis of visual inspection which may result in a bias of the assignment. Recent studies reported that SCOP already covers the majority of fold space [49]. However, SCOP is based on structures that can be crystallized or have been determined by NMR, and is only poorly representing (Section 4.1.2) membrane proteins (Section 3.1) or does not include intrinsically disordered proteins.

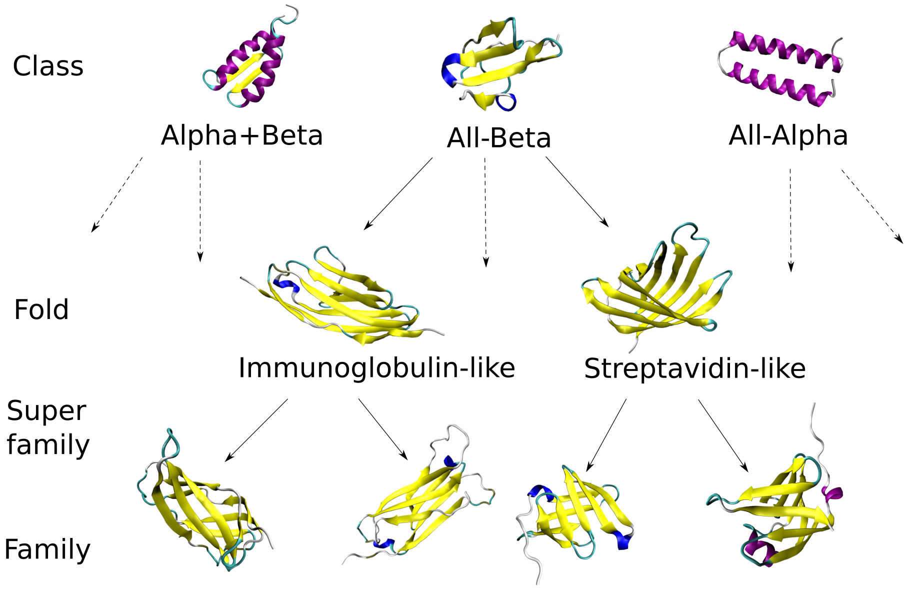


Figure 3.3: Graphical representation of the hierarchical organization of SCOP.

3.6.2 CATH

Class Architecture Topology Homology [10] (CATH) uses different layers to classify protein structures.

- Class: specified by the secondary-structure content of the domain

- Architecture: high structural similarity, but no evidence of homology. Equivalent to a fold in SCOP
- Topology: sharing particular structural features.
- Homology: presence of an evolutionary relationship. Equivalent to the superfamily level of SCOP.

Additionally, the way domains are selected differs from the SCOP database, as CATH mostly utilizes automatic methods based on sequence similarities for dissecting proteins into domains, for instance at the homologous level of classification. Further, at the topology level, structural similarity is used. Only the architecture level is manually assigned based on visual inspection. While most of the analysis of this thesis is based on SCOP (Section 3.6.1, results have been validated by comparisons to CATH.)

Chapter 4

How to evaluate physical properties of protein structures?

Physical properties of protein are numerous. For this thesis, we selected physical properties relevant for the formation and function of proteins and likely to play a role in their evolution. The set of protein molecules investigated here was obtained from the SCOP (Structural Classification Of Protein, Section 3.6.1) database [48]. This classification scheme groups protein domains into super-families and families depending on their general composition and their structural fold. For this set, we measured physical values that potentially played the role of a constraints during the evolution of protein structures. More precisely, we focused on three different but inter-dependant measures, namely protein flexibility, foldability, and mechanical stability, each of which is defined in further detail in the next sections. The generated data required to be stored in a data model that allowed exploration of different variables. Analysis of this data using comparative, statistical, and phylogenetic methods allowed insight into changes of protein fold apparition during evolution.

4.1 Flexibility

Flexibility corresponds to the capacity of a protein to deform. Flexibility is of great importance to a protein's biological function. Flexibility analysis also enables the possibility to assess protein stability [50], which in turn again is a requisite for a protein to play its biological role. Evaluation of the fluctuation of the atoms around their mean position in the protein is commonly used to measure the global flexibility of a protein. Atomic fluctuations can be evaluated using experimental methods such as X-ray crystallography (via the Debye-Waller factor) or specific Nuclear Magnetic Resonance techniques (Section 4.1.2). Within this project, we used two alternative computational methods to assess protein flexibility, the Gaussian Network model (GNM), and CONCOORD, a method based on geometrical constraints.

4.1.1 Computational methods

Computational methods allow to predict the favorable motions of a proteins, using energetic or geometric descriptions of the protein structural space. From the obtained set of motions or structures, i.e. the conformational ensemble, a Root Mean Square Deviation (RMSD) can be obtained as a measure for the overall protein flexibility. The RMSD is the measure of the average distance between atoms of superimposed proteins. In the study of globular protein conformations, one customarily measures the similarity in three-dimensional structure by the RMSD of the C-alpha atomic coordinates after optimal rigid body superposition. When a dynamical system fluctuates around some well-defined average position, the RMSD can be calculated over time.

Gaussian Network model

A Gaussian Network Model (GNM) [51] is created from a protein structure by connecting its neighboring residues by springs with a uniform force constant. Residues are represented by a single particle at the position of the C-alpha atom that represents the magnitude of the residue's positional fluctuations. The correlation matrix of these fluctuations is given by the inverse of a contact matrix in which each atom pair within a given cut off has the value 1 and all other atom pairs the value 0. Harmonic modes of motion are obtained from the correlation matrix, which in turn give the overall flexibility in terms of an RMSD.

CONCOORD

CONCOORD [52] is a Monte-Carlo method which generates a set of conformations. Those conformations are produced following distance constraints between all atoms. A conformation is created starting from random positions for all atoms. Positions are subsequently corrected to obey the geometric constraints such as hydrogen bonds or local contacts. This method is faster than a Molecular Dynamics (MD) simulation but is able to reproduce the atomic fluctuations very well.

4.1.2 Experimental techniques

The two major techniques to determine the structure of a protein at atomic detail are Nuclear Magnetic Resonance (NMR) and X-Ray crystallography. They allow limited insight into the flexibility of the protein, and also are the starting points for the computational methods described in Section 4.1.

The NMR spectrum provides information about the chemical environment of the nuclei. Applying an electromagnetic field to a certain atom give rise to a resonance phenomenon caused by the nuclear spin. This resonance occurs upon absorption of energy at a precise frequency which depends on the electromagnetic environment of the atom. Hence, measuring spin frequencies of macromolecules in solution enables the possibility to determine relative atomic positions in a given molecule, allowing the reconstruction of the three-dimensional structure of proteins.

Novels developements such as dipolar residual couplings also allow insights into protein dynamics at shorter to larger time scales [53]. X-Ray crystallography, in contrast, requires a protein in crystalline form, as the periodic arrangement of the molecules in a crystal is required for a valuable diffraction pattern of the X-ray light. The diffraction pattern can be used to infer the three-dimensional arrangement of the diffracting electrons, and therefore of the atoms of the molecule. Flexibility is only partly, within the low temperature crystal, embedded into the B-factors of the atoms.

Resulting three-dimensional structures from NMR and X-Ray crystallography are stored in databases such as the Protein Data Bank (PDB) [54] and further classified into domains in SCOP and CATH (Section 3.6).

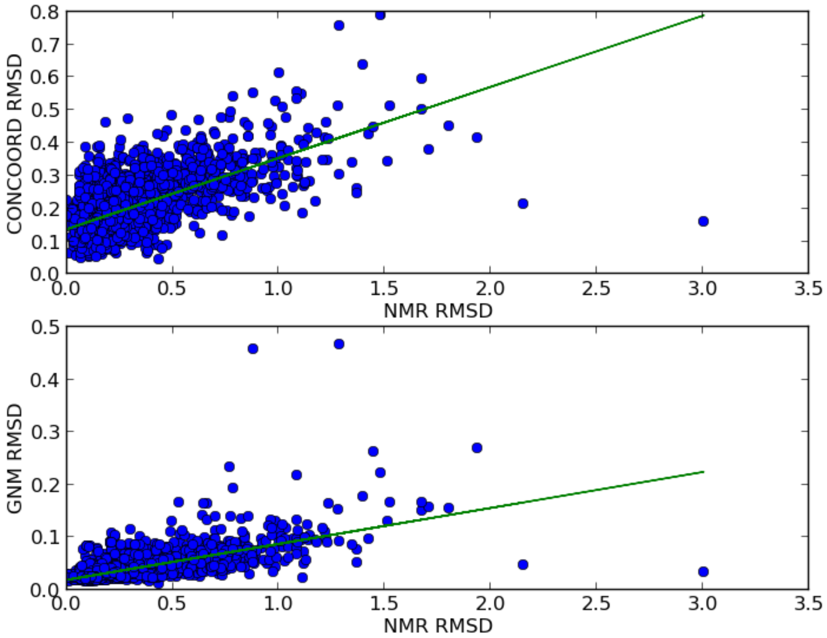


Figure 4.1: Correlation of the flexibility predicted by theoretical methods (CONCOORD and GNM) with an experimental method (NMR). The sample set includes 600 protein domains. The RMSD (Root Mean Square deviation) within the Cα-atoms of the computed or measured structural ensemble is given in nm. The green line is a linear fit to the data.

We could show that the flexibility predicted by CONCOORD and GNM correlates well with those measured in NMR ensembles (Figure 4.1). The correlation coefficients are 0.63 and 0.66, respectively. Correlation between GNM and NMR ensemble was previously observed by Bahar et al [55].

4.2 Foldability

4.2.1 Computational methods

The folding of a protein is a physical process by which a polypeptide adopts a characteristic three-dimensional structure, which is functional. Every protein is trans-coded from an mRNA sequence into a linear amino-acid chain. This polypeptide does not possess a three-dimensional structure at this time. However, each amino-acid of the chain possesses some essential chemical characteristics. This could be hydrophobicity, hydrophility, or electric charge. They interact with each other, leading to a well-defined three-dimensional structure, the folded protein or so called native state. The three-dimensional structure is determined by the amino-acid sequence. In the present work, foldability is measured by the folding time, the time from the unfolded to folded state, assuming that efficient folding without the risk of misfolding requires, among others, a short folding time, i.e. a short life time of the unstable and aggregation-prone unfolded state. The folding time was assessed by a method called Contact Order. Contact Order [56] is a value used in order to estimate the folding time of a protein. It is measured from the average number of amino-acids between all contact points (commonly defined with a cutoff values of around 7 Å between Cα-atoms) within a protein. It correlates with the number of long-range contacts, i.e. contacts distant in sequence but close in space. A high value indicates many long-distance contacts, which will result in a longer folding time. Contact order was found to be in good correlation with folding times of two state folders but not multi-state proteins. Subsequent studies with extended comparison to experiments led to the definition of the Size-Modified Contact Order [57] (SMCO),

$$\text{SMCO} = (\frac{1}{L}\sum^{N}\Delta L_{ij}) \cdot L^{0.7}, \tag{4.1}$$

where N is the number of contacts, L is the total number of aminoacids, and ΔL_{ij} is the number of aminoacids along the chain between residues i and j forming a native contact. By correcting for protein size L, the SMCO showed an improved correlation with experimental folding times, with a correlation coefficient of 0.74 [57]

4.2.2 Experimental techniques

Various experimental techniques which measure quantities varying during the folding process have been developed to assess folding times and mechanisms, including fluorescence or absorption spectroscopy. One conventional method is circular dichroism, which depending on the secondary structure measures the absorbtion of circular polarized light. Quantifying the absorption of light allows to evaluate the degree of nativeness of the protein. The degree of nativeness of a given protein is artificially modified under the action of a chemical or physical denaturant. Most commonly, the protein is denatured by heat, light or solvent (such as denaturant), and then allowed to relax into the folded state upon removal of the denaturing conditions. Variation of the denaturant concentration allows to monitor the folding or unfolding of a protein.

Folding experiments are typically represented by a Chevron plot (Figure 4.2), which allows insights into the number of states involved, such as two-state (unfolded and folded) or three-state folding (comprising also an intermediate state). A denaturation midpoint corresponds to the temperature (Tm) or denaturant concentration (Cm) at which half of the protein is folded and the other half is unfolded.

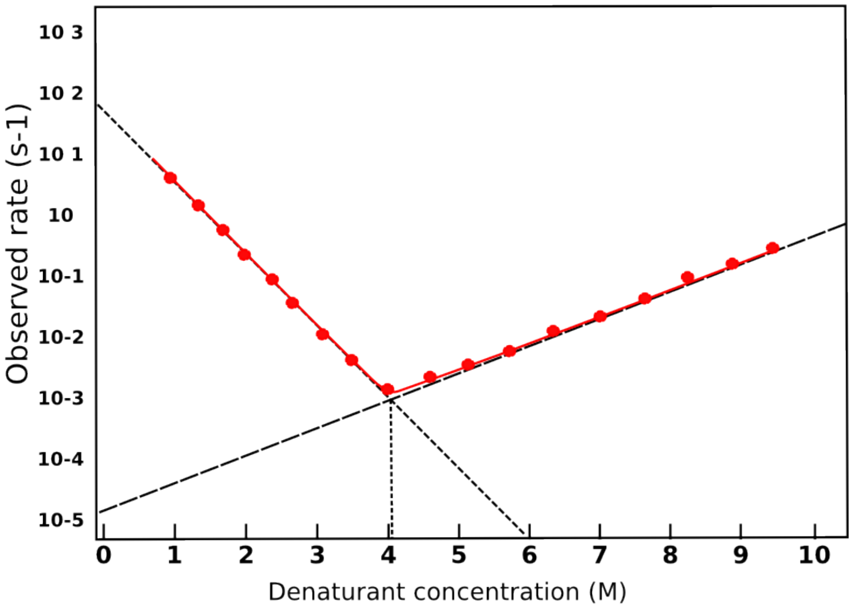


Figure 4.2: Chevron plot obtained from relaxation rates as a function of the denaturant concentration. A linear change in rate as shown in the schematic example here indicates two-state folding.

4.3 Mechanical stability

4.3.1 Computational methods

Go-model

Mechanical strength is experimentally assessed by measuring rupture forces. A simple and computationally efficient method to predict rupture forces from simulations is the Go-model [58], a coarse-grained model based on the experimental X-ray structure. The Go-model treats each amino acid by a single bead connected to its neighbors by springs. Contacts close in space in the X-ray structure are assigned favorable potentials to stabilize the protein in its native state. Here, we will use this model to predict the force to unfold each SCOP domain, and use forces as a measure for mechanical strength (Chapter 5).

The peak of the force curve (Figure 4.3) is used as a reference for the mechanical strength of a given domain. The mechanical strength values obtained by this method are in good agreement with experimental data [59]. The comparison of experimental force peaks with simulations is shown in Figure 4.4. The dataset is composed of 13 domains and the correlation coefficient is 0.90. Thus, our computational results agree very well with experimental measurements of mechanical strength.

We used a specific implementation of the Go-model called Self-Organized Polymer (SOP) [60], which describes a protein in terms of beads on the position of Cα-atoms representing amino-acids. SOP uses the Langevin equation to describe the dynamics of the i-th Cα-atom:

$$\xi \frac{dR_i}{dt} = \int (R_i) + G_i(t) \tag{4.2}$$

where ξ is the friction coefficient, $G_i(t)$ is the Gaussian distributed random force with zero mean and delta function correlations (white noise). The random force mimics hits of protein residues with the solvent (water) molecules.

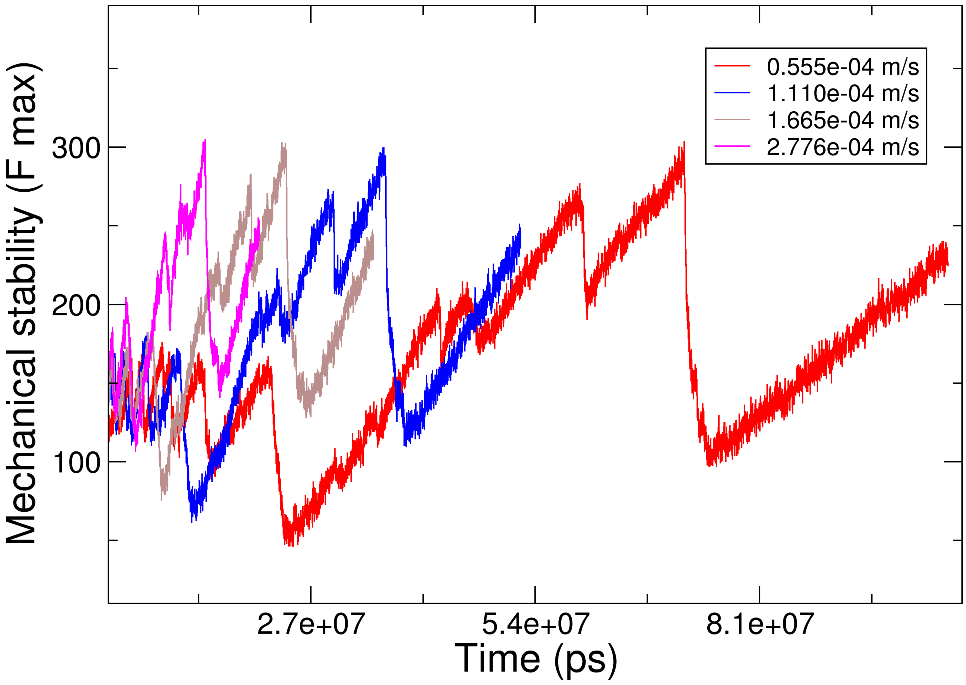


Figure 4.3: Force curves obtained for unfolding the coronavirus main proteinase for different pulling velocities obtained from pulling simulations using the Go-model. Mechanical stability is measured by the maximal force, F_{max}, here approximately 300 pN.

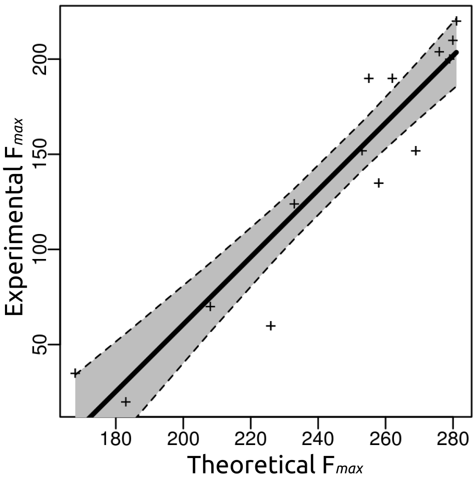


Figure 4.4: Correlation between the experimental and theoretical F_{max}. The solid line is a linear fit, the grey shade a 95% confidence interval.

$\int(R_i) = -\partial V/\partial R_i$ is the molecular force exerted on the i-th particle due to the potential energy V. The force field (potential energy function) of a protein conformation is given by:

$$\begin{aligned} V &= V_{FENE} + V_{NB}^{ATT} + V_{NB}^{REP} = && (4.3)\\ &- \sum_{i=1}^{N-1} \frac{k}{2} R_0^2 \log\left(1 - \frac{(r_{i,i+1} - r_{i,i+1}^0)^2}{R_0^2}\right) && (4.4)\\ &+ \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \varepsilon_h \left[\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6}\right] \triangle_{ij} && (4.5)\\ &+ \sum_{i=1;j=i+2}^{N-2} \varepsilon_l \left(\frac{\sigma_{ij}}{r_{ij}}\right)^6 + \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \varepsilon_l \left(\frac{\sigma}{r_{ij}}\right)^6 (1 - \triangle_{ij}) && (4.6) \end{aligned}$$

The first term in the equation describes the backbone chain connectivity using the finite extensible nonlinear elastic (FENE) potential with a tolerance in the change of a covalent bond of $R_0 = 2$ Å and a force constant of k=1.4 N/m. The distance between any two interacting residues i and $i+1$ is $r_{i,i+1}$, whereas $r_{i,i+1}^0$ is the value in the native (PDB) structure. The second term (4.5) in the equation represents forces such as hydrophilic and hydrophobic interactions (only between non-covalently linked residues i and j, i.e. |i - j|> 2) via attractive and repulsive forces defined by a cutoff distance R_c in the native state, i.e., $r_{ij} < RC$, then $\triangle_{ij}$=1 (for native contacts), and zero otherwise (for non-native contacts). The strength of the non-bonded interactions is given by ε_h term. Additional constraints are imposed on the bond angles formed by residues i, i+1, and i+2 by including a repulsive potential with parameters ϵ_l=1 kcal/mol and $\sigma = 3.8$ Å, which quantify, respectively, the strength and the range of the repulsion. To ensure self-avoidance of the protein chain between beads of non-native contacts ($r_{ij} > RC$), a repulsive term (last term in Eq.4.6) is introduced, with $\sigma = 3.8$ Å. To induce mechanical unfolding, we pulled each protein structure from the N to C terminal residue with a constant velocity of 2,776*10^{-4} nm/ns and a spring constant of 700 pN/nm. The simulation was stopped, when the N- to C-distance represented 90 % of the maximum distance between the N and C termini (calculated by the number of residues times 1.4 Å). We defined the topologies according to the SCOP database (Section 3.6.1). The timescale of the simulations was defined according to the following relation

$$\tau_L = (\frac{ma^2}{\epsilon_h})^{1/2}, \tag{4.7}$$

considering inertia in the Langevin paradigm with a unit-less mass of $m = 3 * 10^{-22}$, a distance of $a = 5 * 10^{-8}$ and $\epsilon_h = 1.4$ kcal/mol^{-1}. When no inertia are considered, the timescale becomes

$$\tau_H = \frac{\zeta a^2}{k_B T_s} = \frac{6\Pi\eta a^3}{k_B T_s} = \alpha\tau_L \frac{\epsilon_h}{k_B T_s} \tag{4.8}$$

and corresponds then to a Langevin overdamped limit or Brownian dynamics simulation. The real time can then be obtained from $\triangle T * \tau_H$, which is the relation used in this thesis.

4.3.2 Experimental techniques

Atomic force microscopy (AFM) and optical tweezer experiments have enabled the induction and monitoring of large conformational changes in biomolecules, including protein denaturation and refolding under mechanical force. In such a study, a pulling force is applied on given points of the protein of interest - often the termini. Optical tweezers use a focused laser beam to move a nano-element. The electric gradient generated by the laser beam attracts the nano-element to the center of the beam, allowing a controlled displacement of the nano-element. The nano-element is attached to one side of the protein while the other side is fixed (Figure 4.5a). In an AFM, the laser beam is replaced by a nano-stick attached to the protein called cantilever (Figure 4.5b). Both methods allow to evaluate the force needed to unfold a protein, but lack a description at the atomic level. Computational methods can help to complement these experiments by suggesting pathways of the protein. Also, in this thesis, they allow predicting unfolding forces of $\sim$100.000 protein domains, which is currently unfeasible by experimental means.

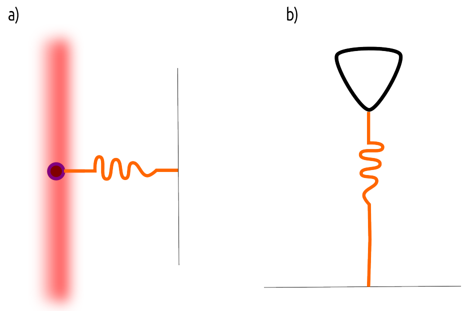


Figure 4.5: Schematic representation of experimental techniques. a) Optical tweezer experiments with the protein attached to a surface and a bead, which is optically trapped in a laser beam. b) Atomic Force Microscopy with the protein attached with a sharp tip and a surface.

Part II

Physical constraints in protein structure evolution

Chapter 5

Evolutionary optimization of protein folding

5.1 Introduction

The catalog of naturally occurring protein structures [61] exhibits a large disparity of folding times (from microseconds [62], to hours [63]). This disparity is the result of roughly ∼3.8 billion years of evolution during which new protein structures were created and optimized. The evolutionary processes driving the discovery and optimization of protein topologies is complex and remains to be fully understood. Nature probably uncovers new topologies in order to fulfill new functions, and optimizes existing topologies to increase their performance. Various physical and chemical requirements, from foldability to structural stability, are likely to be additional players shaping protein structure evolution. One indicator for foldability, i.e. the ease of taking up the native protein fold, is a short folding time. Here we propose that foldability is a constraint that crucially contributes to evolutionary history. Optimization of foldability during evolution could explain the existence of a folding funnel [64, 65], into which a defined set of folding pathways lead to the native state, as postulated early on by Levinthal [66]. While the biological relevance of efficient folding still needs to be explored, an obvious advantage is the increase of protein availability to the cell. For instance, folding could decrease the time between an external stimulus and the organismal response. However, this increase of accessibility is probably limited by other factors such as protein synthesis, proline isomerization and disulfide formation. A probably more important point to support folding speed as an evolutionary constraint is that fast folding avoids proteins aggregation in the cell [67]. Aggregation avoidance could lead to a selection of topologically simple structures that fold rapidly or exclusion of a large number of geometrically feasible structures that compromise accessiblity. This could have reduced the catalog of naturally occurring folds [60, 68, 69]. The balance between the need for new structural designs and functions in evolution and the physical requirements imposing pressure on folding has remained elusive. The increasing number of organisms with completely sequenced genomes and experimentally acquired models of protein structures, combined with new techniques to study the folding behavior of proteins now open new avenues of inquiry. A common approach for such studies has been the use of molecular simulations such as lattice or coarse-grained techniques, which are efficient enough to scan sequence space. Simulations generally involve an algorithm that mimics the evolutionary accumulation of mutations. This allows to monitor how proteins are selected and evolve towards specific features that are optimized, including those linked to folding, structure and function [70–72]. In contrast, we have uncovered phylogenetic signal in the genomic abundance of protein sequences that match known protein structures. Specifically, phylogenomic trees that describe the history of the protein world are built from a genomic census of known protein domains defined by the Structural Classification of Proteins (SCOP) [48] and used to build timelines of domain appearance [5, 73] that obey a molecular clock [18]. This information revealed for example the early history of proteins [17], planet oxygenation [18], and the dynamics of domain organization in proteins [19]. All-atom simulations of denatured proteins folding into their native state [74, 75] are computationally too demanding to systematically evaluate the folding times of the available structural models of protein domains, currently ∼100,000 in total. A decade ago, Baker and co-workers [56] introduced the concept of contact order, a measure of the non-locality of intermolecular contacts in proteins. Contact order was found to be in good correlation with folding times of two state folders but not multistate proteins. Subsequent studies with extended comparison to experiments led to the definition of the Size-Modified Contact Order (SMCO),

$$\text{SMCO} = (\frac{1}{L}\sum^{N} \Delta L_{ij}) \cdot L^{0.7}, \tag{5.1}$$

where N is the number of contacts, L is the total number of aminoacids, and ΔL_{ij} is the number of aminoacids along the chain between residues i and j forming a native contact. By correcting for protein size L, the SMCO showed an improved correlation with experimental folding times, with a correlation coefficient of 0.74 [57].

Here, we reveal evolutionary patterns of foldability by mapping the SMCO and thus the folding time onto timelines derived from phylogenomic trees of domain structures (Figure 5.1).

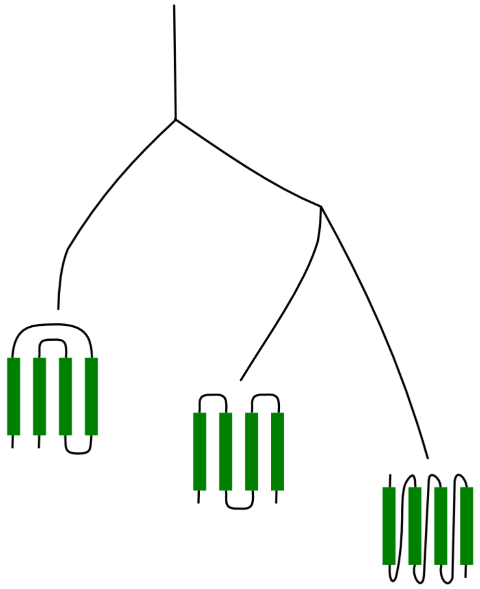


Figure 5.1: Protein topologies that favor short range inter-aminoacid contacts might be the result of an evolutionary optimization of foldability and thus would have likely appeared late in evolution.

Remarkably, we find there is selection pressure to improve overall foldability, i.e reduce folding times, during protein history. Interestingly, different topologies such as all-β and all-α folds show distinct patterns, suggesting folding impacts the evolution of some classes of protein structures more than others.

5.2 Results

5.2.1 Change in foldability during evolution

To trace protein folding in evolution, we determined the SMCO of protein domain structures at the Family (F) level of structural organization. Figure 5.2a shows the folding rate of each F, as measured by its average SMCO, as a function of evolutionary time. Using polynomial regression, we observed a significant decrease (p-value = 9.5e-15) in SMCO in proteins appearing between ∼3.8 and ∼1.5 billion years ago (Gya). Trends were maintained when excluding domains from the analysis solved in multi-domain proteins, and also when studying domain evolution at more or less conserved levels of structural abstraction of the SCOP hierarchy. Namely, we find a significant decrease of SMCO at the level of Superfamily (SF), p-value = 2.6e-15), and at the level of domains with less than 95 % sequence identity (p-value <= 2.0e-16). Similarly, consistent results were obtained at the F level using linear regression (p-value = 1.0e-06). Remarkably, even within a smaller data set of only 87 proteins for which folding times have been measured [76], we find that the experimental folding times exhibit a tendency to decrease early in protein evolution (Figure 5.4). As an additional way of validation, we repeated the analysis for ∼3 million single domain sequences with predicted SMCO [77], and obtained a decrease again of SMCO up to ∼1.5 Gya (p-value <= 2.0e-16). Thus, in this initial evolutionary period, proteins tended to fold faster on average. As suggested by the decrease

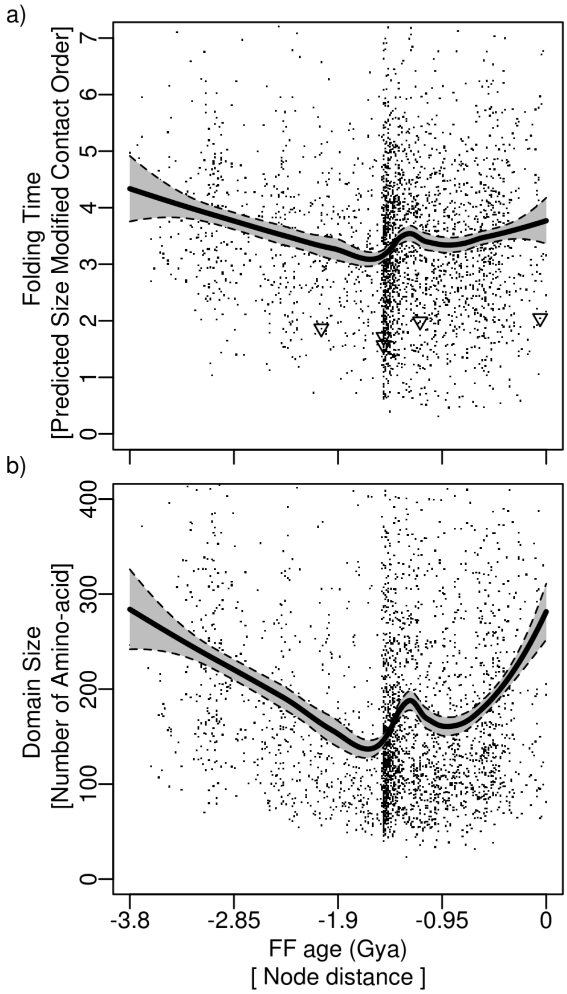


Figure 5.2: Change in length and foldability during evolution : a) Size Modified Contact Order (SMCO) versus approximative F domain age in billion of years (Gya). Each data point represents an SMCO average of domain belonging to the same F. Triangles show SMCO averages for domains belonging to the same F and experimentally known to be ultra-fast folders [78]. b) Average amino-acid chain length for domains belonging to the same F versus F domain age in Gya. The solid line shows a LOESS polynomial regression [79], and the grey shade the 95% confidence interval.

in SMCO, during evolution, domains diminish long-range and favor short-range interactions, thereby becoming more strongly connected locally. This picture was further corroborated by an analogous analysis of evolutionary trend in tightness, measured by shortest paths in the network of protein contacts [80]. Tightness, and thus the lengths of paths in the interaction network, decreased in evolution until ∼1.5 Gya, followed by an increase, just like the SMCO (Figure 5.3). Our results support the hypothesis that folding speed acts as an evolutionary constraint in protein structural evolution. In contrast, we observed an increase in SMCO between ∼1.5 Gya and the present (Figure 5.2a). Thus, the appearance of many new structures by domain rearrangement ∼1.5 Gya, also refered to as the "big bang" [19] of the protein world, affected the evolutionary optimization of protein folding. While a linear regression supports the SMCO increase (p-value = 2e-16), it was not as observed at the SF level or at the level of domains, and for the analysis of experimentally determined rates (Figure 5.4). Given the observed overall evolutionary speed-up of protein folding, we would expect a late evolutionary appearance of so-called downhill proteins, which feature ultra-short folding times on the microsecond scale. We annotated 11 downhill folders [78] by their Fs, namely a.35.1.2, a.4.1.1, a.8.1.2, b.72.1.1, and

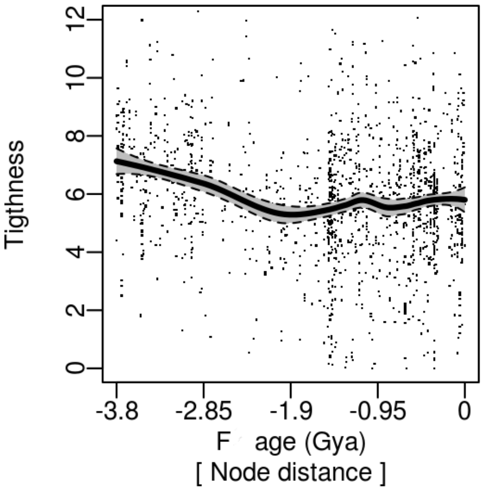


Figure 5.3: Tigthness versus approximate domain age (Gya). A polynomial regression is shown as black solid line. The gray area indicates the 95% confidence interval.

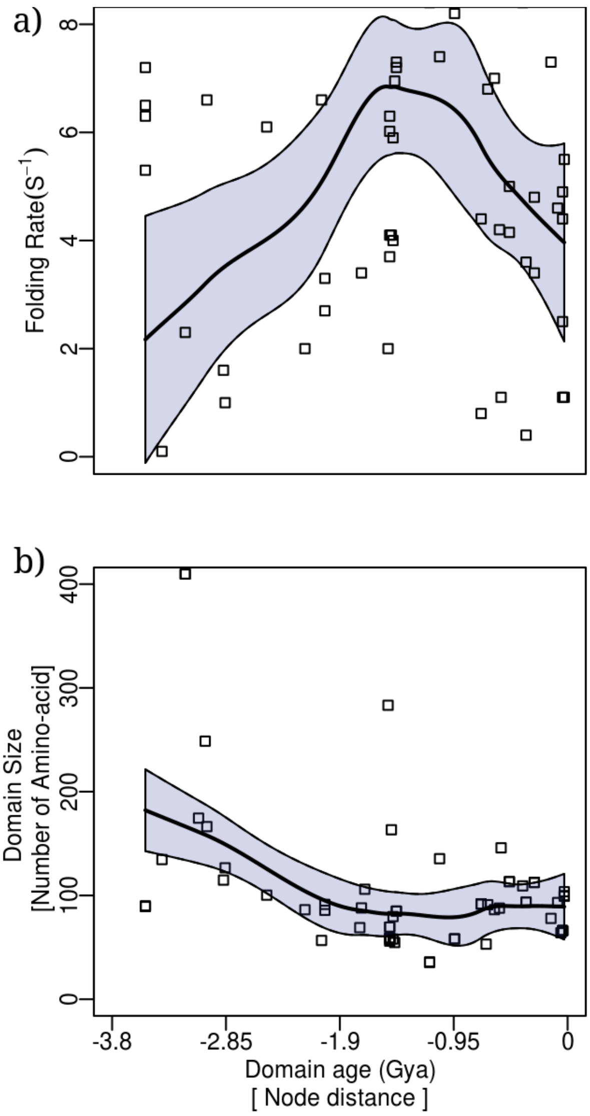


Figure 5.4: Evolutionary changes for an experimental dataset. a) Experimental folding rates versus approximate domain age in billion of years ago (Gya). b) Domain size of the same set of 87 proteins versus approximate domain age. A polynomial regression is shown as black line, and the 95 % confidence interval as grey shade.

d.100.1.1, and show their average SMCO per family as black triangles in the timeline of Figure 5.2a. All of them, unsurprisingly, have an SMCO < 2, and thus fold significantly faster on average than other structures. We find 7% of families to have a lower SMCO (SMCO < 1.5) than the experimentally identified downhill folders. We predict these Fs will fold even faster than the known downhill folders, rendering them interesting candidates for folding assays. The five Fs containing the fast folders have all appeared no earlier than ∼2.5 Gya, suggesting that they are a result of lengthy evolutionary optimization. According to our predictions, the first fast-folding proteins appeared already ∼3.4 Gya. However, their frequency and optimization of folding speed continue to increase until ∼1.5 Gya.

5.2.2 Protein length and evolution of foldability

The length of the amino acid chain has been reported to influence the folding kinetics of a protein, with longer chains folding more slowly [57, 78, 81, 82]. We therefore ask if the decrease in SMCO we observed from ∼3.8 to ∼1.5 Gya can be explained by a decrease in the chain length of proteins. Figure 5.2b shows how domain size has varied in evolution. Folding time measured by SMCO and domain size follow a very similar bimodal trend, with a clear decrease occuring prior to ∼1.5 Gya and a slight increase after the

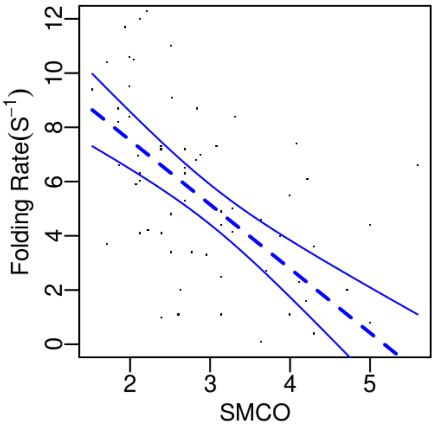


Figure 5.5: Size Modified Contact Order (SMCO) versus folding rate for 87 proteins with experimentally known folding rates. A linear regression is shown as blue dashed line. The solid lines indicates the 95% confidence interval.

"big bang". As expected, we find domain size, which equals L in Equation 1, and SMCO to be correlated with folding rate in agreement with other studies [57, 60] (Figure 5.5). In line with this correlation, the downhill folders discussed above and shown in Figure 5.2a as triangles, have a small domain size of less than 100 residues in common.

We next eliminated the effect of domain size on the evolutionary trends observed in folding rate to analyze factors other than domain size. To this end, we dissected our dataset according to the amino acid chain length. This analysis was done with all ~92,000 domains to ensure enough data points for each length. The distributions of chain length are shown in Figure 5.6a, b for the two time periods before and after the "big bang" (~1.5 Gya). The length distribution for proteins appearing before the "big bang" exhibited a peak at around ~150 amino acids, and shifted later (~1.5 Gya to the present) to shorter chains with a peak at around 100 aminoacids, underlining the tendency for a decrease of domain size. We note that the resulting average chain length of three-dimensional structures in SCOP, which have been obtained from X-ray or NMR measurements, is smaller than the average length of sequences in genomes [83], apparently due to the increasing experimental difficulties when working with large proteins. We then analyzed evolutionary tendencies for every domain length subset by measuring the variation in the end points of a polynomial regression. The color mapping in Figure 5.6a indicates an increase (blue), a decrease (yellow-red), or a non-significant change (green) of SMCO. Overall, 85 % of the data returned a significant result according to the F-test. During early protein evolution (3.8-1.5 Gya), we found that 54 % ± 0.3 % of all domains in each size subset optimized their foldability during evolution by decreasing their SMCO. Conversely, 37 % ± 0.4 % of domains showed a slow-down in folding, i.e. a significant increase in SMCO. These results confirm the tendencies observed for the full data set (Figure 5.2a), and hold for different tresholds of identity,

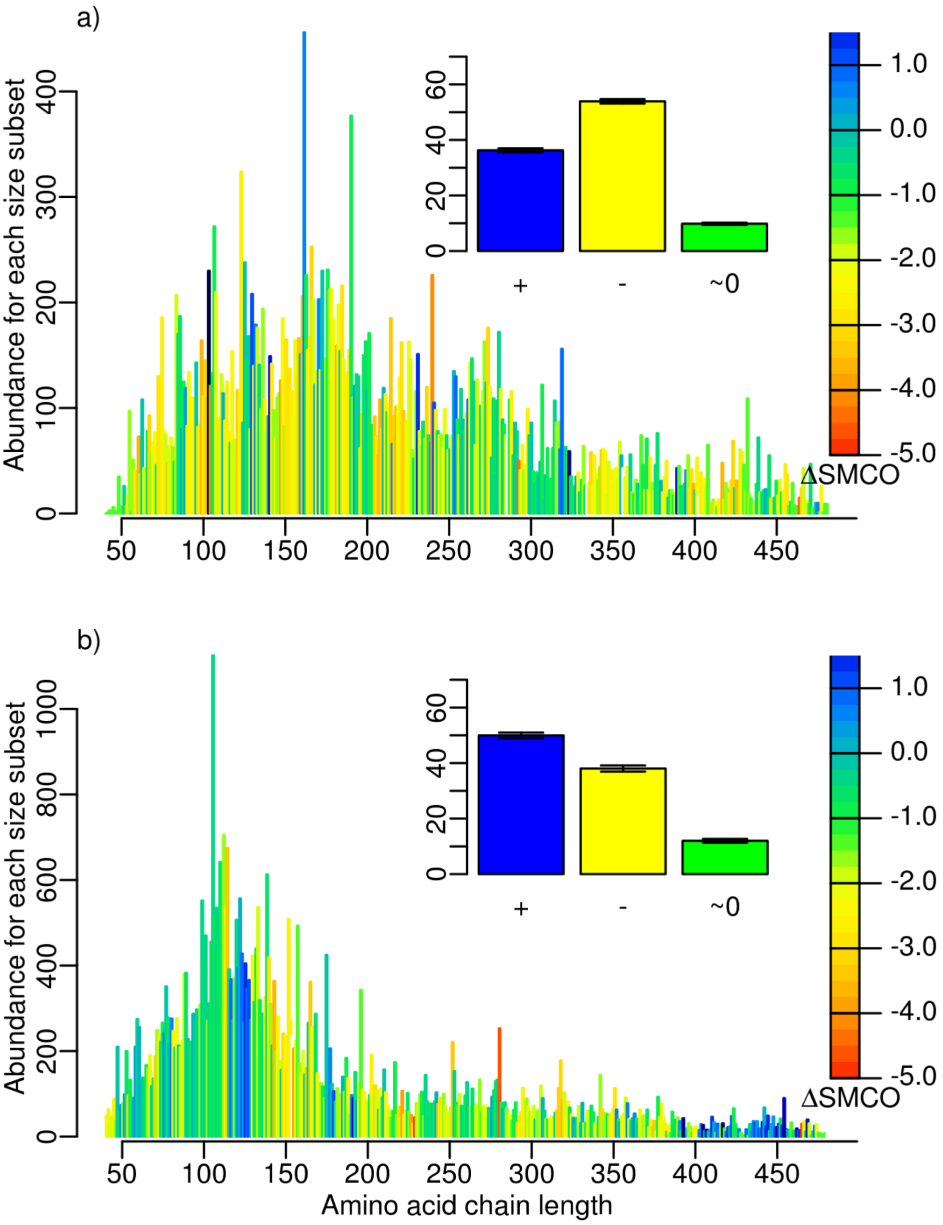


Figure 5.6: Change in foldability during evolution for subsets of chain size : Distribution of domain length for domains appearing a) 3.8-∼1.5 Gya and b) ∼1.5-0 Gya. Abundancies were colored according to the average ΔSMCO, the difference between the end points of the polynomial regression of SMCO in this dataset, for the specified initial (a) and later (b) time period. Yellow to red indicates a decrease, and blue an increase in SMCO. The barplots (inset) show the percentage of domains with positive (blue), negative (yellow), and insignificant (green) ΔSMCO.

namely 95 % and 40 %. As expected, due to the smaller data set, partitioning domains defined at F and SF levels according to size yielded results that were statistically not significant. In summary, even after dissecting the effect of chain length on changes in SMCO, the tendency of proteins to fold faster during evolution is confirmed.

After the "big bang", the SMCO and thus foldability showed a overall increase in evolution (Figure 5.6b), in agreement with results from the total set (Figure 5.2a). Apparently, fast folding did not represent a major evolutionary constraint during this period. Instead, other constraints must have been optimized at the expense of foldability. We next discuss secondary structure as one factor influencing the impact of foldability on protein structure evolution.

5.2.3 Secondary structure and evolution of foldability

Secondary structure composition is another factor reported to have an influence on folding kinetics [57, 78, 81]. We repeated the analysis of domains partitioned by size that was described above for domains in each secondary structure class of SCOP (all-α, all-β, α/β, and $\alpha+\beta$ domains) and thereby revealed differences in the evolution of foldability. As shown in Figure 5.7a, the tendency of a decreasing SMCO before the "big bang" is reproduced for all classes. This result was confirmed at the level 95 % identity and 40 % identity, though with a significant decrease only for the $\alpha+\beta$ and α classes at the 40 % identity level, i.e. for a much smaller data set. Again, our analysis strongly supports an evolutionary constraint for fast folding of proteins appearing early in evolution, 3.8-1.5 Gya.

Interestingly, we here observe a specialization of protein classes, with all-α proteins tending to fold faster and all-β proteins tending to fold more slowly, all of which was supported at the 95 % domain level (Figure 5.7b). Why should the all-α class be under a stronger fast folding constraint than the all-β class? Figure 5.8 shows the average SMCO for each secondary structure class. The all-β and all-α class show the highest and lowest SMCO, respectively, suggesting that all-β proteins in general fold slower than all-α proteins. This is in line with previous findings that containing all-β proteins fold more slowly than all-α proteins due to long range interactions between all-β strands that increase contact order [57, 78, 81].

5.3 Discussion

Protein aggregation damages cellular components and can lead to a variety of neuronal diseases [84–86]. A way of reducing aggregation is to enhance the kinetic and thermodynamic accessibility of the native fold of a protein. Incremental increases in kinetic or thermodynamic stability of a protein might therefore represent an evolutionary trace reflecting optimization of protein foldability [87].

Here, we confirm the hypothesis that foldability exerts a constraint in the evolution of protein domain structures, as we find a tendency of proteins to on average fold faster than their structural ancestors. As expected, shortening of protein chain length during evolution is an important factor leading to faster folding. However, the exclusion of this protein-size effect preserved the trend of decreasing folding times. Thus, faster folding is not a side effect of chain shortening, but likely acts as an evolutionary constraint in itself. An alternative reason for the decrease of folding times in evolution is the need of proteins for flexibility in order to optimize their function such as enzymatic catalysis or allosteric regulation [88]. Folding speed and flexibility are known to correlate, as the formation of the compact state with no or only minor native contacts is much quicker than the arrangement of the native – often long-range – contacts [89]. Fewer native contacts in turn result in lower stability and may increase conformational flexibility as required for some biological functions [90]. Our analysis of protein folding

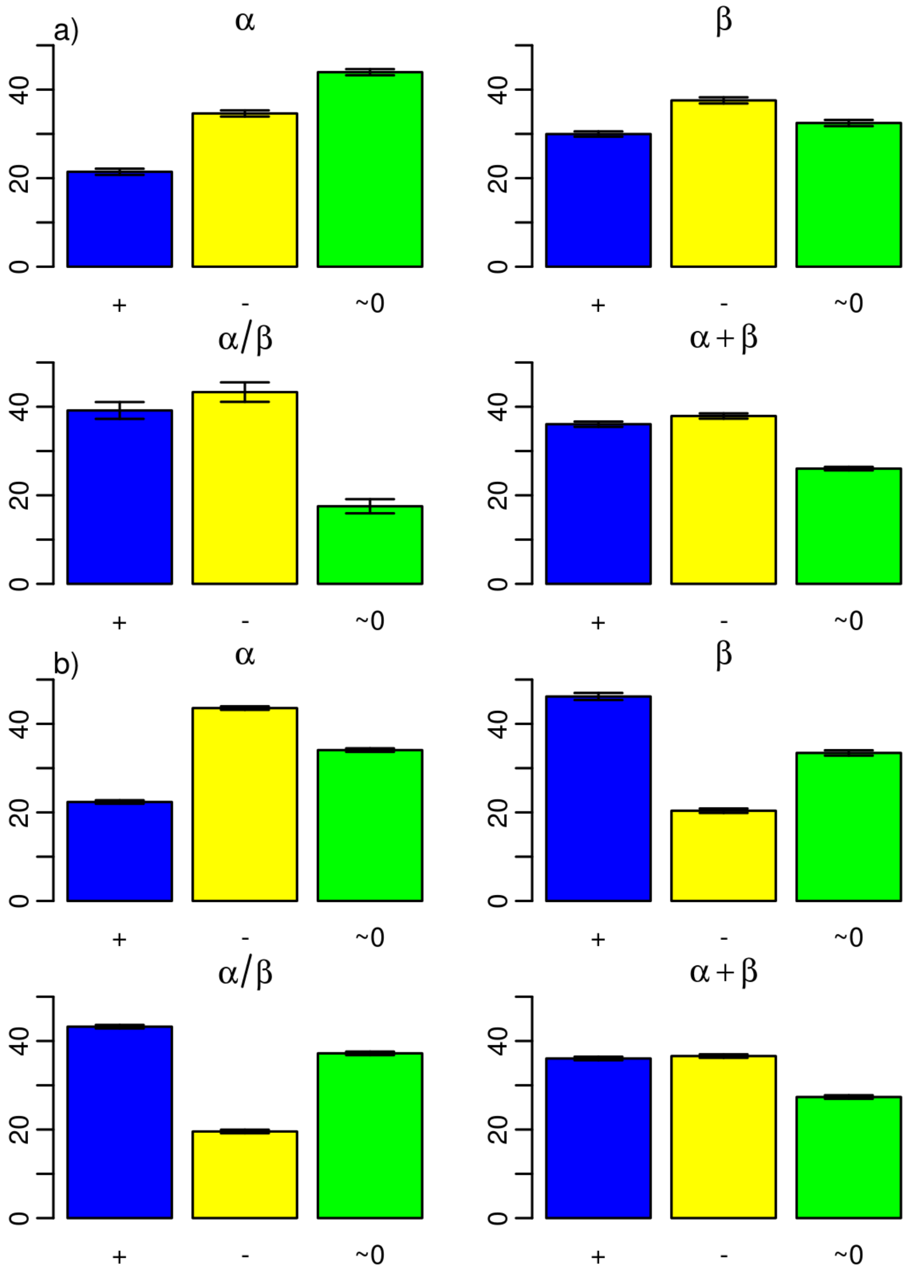


Figure 5.7: Percentage of all domains with a positive (blue), negative (yellow), and insignificant (green) ΔSMCO. a) for 3.8-∼1.5 Gya, and b) ∼1.5-0 Gya. Each barplot considers one of the four fold classes according to their secondary structure: all-α, all-β, α/β, and α+β, as indicated. The barplots were obtained from domain length distributions analogous to those shown in Figure 5.6.

speed on an evolutionary time line can be similarly carried out for measures of flexibility to test this scenario.

Evolutionary constraints on folding are apparently not uniformly imposed onto the full repertoire of protein structures and during the entire protein history. Instead, our analysis revealed a bimodal evolutionary pattern, with folding speed increasing and decreasing before and after ∼1.5 Gya, respectively. The speed-up of folding was most pronounced for all-α folds. The evolutionary inflexion point coincides with the previously identified protein "big bang", which features a sudden increase in the number of domain architectures and rearrangements in multi-domain proteins triggered by increased rates of domain fusion and fission. We speculate that the slow down of folding that ensues could be due to cooperative interactions during folding of domains in the emerging multi-domain proteins [91]. Alternatively, the observed slow-down after the "big bang" could be related to the appearance of protein architectures that are known to help proteins to fold, such as chaperones [92, 93] Moreover, protein architectures specific to eukaryotes appeared at ∼1.5 Gya [73]. The Eukaryotic domain of life has the most elaborate protein synthesis and housekeeping machinery, including enzymes for post-translational modification. This machinery might have mitigated the constraints for fast folding, thereby increasing evolutionary rates of change [87], while preventing misfolding and aggregation prior to attaining the native fold [94].

Finally, we revealed striking evolutionary diversity in protein folding when comparing all-α and all-β fold classes from ∼1.5 Gya. Their average folding times diverged after the "big bang", with the all-α class further decreasing and the all-β class instead increasing their folding times. This result can support the idea of an optimization of folding that increased the difference in folding time between all-β and all-α through evolution. As previously shown [56], all-β folds have on average higher SMCO and fold slower than their all-α-counterparts. This simply results from their different topology and is also the result of our analysis (Figure 5.8). We here show that earlier in evolution, however, folding times have been more similar and only diverged from each other as late as after 1.8 Gya. But why would all-β folds have been relieved from the evolutionary constraint of fast folding? Since the "big bang" is responsible for the discovery and optimization of many new functions, including an elaborate protein synthesis and folding machinery, we speculate that the divergence of averge folding times of all-α and all-β folds probably reflects an optimization of function. This optimization happens to be on the expense of foldability for only the all-β class, the reasons of which remain unknown. One possible scenario would be that all-α have the tendency to carry out functions that require high flexibility, a property that correlates with few long-range contacts, i.e. high foldability.

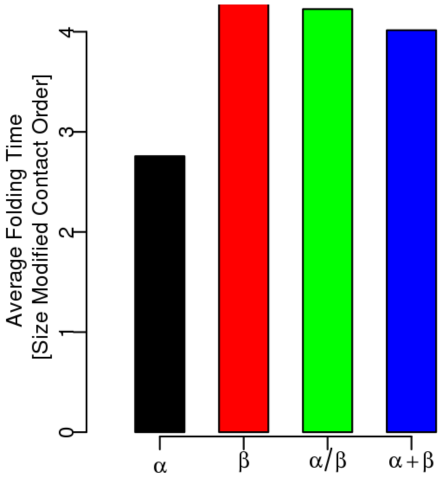


Figure 5.8: Average SMCO for the four fold classes according to their secondary structure: all-α, all-β, α/β and α+β. all-β proteins fold significantly more slowly than all-α proteins. The Wilcoxon rank-sum test return a p-value ± 2.2e-16 for every pair of datasets. The higher average SMCO for all-β as compared to all-α proteins confirms earlier findings.

An important experimental study by Baker and colleagues [95] tested the idea that rapid folding of biological sequences to their native states does not require extensive evolutionary optimization. Using a phage display selection strategy, the barrel fold of the SH3 domain protein was reproduced with a reduced alphabet of only five amino-acids without any loss in folding rate. Despite extensive changes to protein sequence, experimental manipulation preserved contact order. While these results should not be generalized to the thousands of other fold topologies that exist in nature, they are revealing. They suggest that stabilizing interactions and sequence complexity can be sufficiently small and still enable evolutionary folding optimization. In other words, optimal folding structures can find their way through the free energy landscape without extensive explorations of sequence space. This property of robustness could be a recent evolutionary development, since the SH3 domain F appears very late in our timeline of protein history. Alternatively, it could represent a general structural property. The fact that we now see clear and consistent foldability patterns along the entire timeline supports the existence of limits to evolutionary optimization of folding that are being actively overcome in protein evolution. We conjecture that these limits were initially imposed by the topologies of the early folds, and that structural rearrangements (resulting from insertions, tandem duplication, circular permutations, etc [96–99]) offered later on opportunities for fast and robust folding as evolving structures negotiated trade-offs between function and stability.

We end by noting that we cannot exclude overlooking effects on folding times from cooperative folding. These could influence trends of folding times. The SMCO is known to show high correlations with folding times only for single-domain proteins [56]. Developing schemes for estimating folding times from structures comprising more than one domain is a challenge [91] but would enable a more general view onto protein foldability as a constraint throughout evolution. Moreover, our analysis is based on the sequence and structural data that is available. Results might therefore be biased by the choice of proteins and their accessibility. However, the structure of most protein folds and families have been acquired and will not exceed those that are expected [100]. Moreover, our approach allow us to steadily test if the predicted evolutionary trends of foldability are maintained upon inclusion of new sequences and protein folds into the analysis. Interestingly, multiple studies have found folding rates to correlate with stability rather than contact order [101]. Analyzing phylogenomic trends of stability might in this light be an important study to further elucidate evolutionary contraints on protein structure.

Chapter 6

Evolution of protein mechanical stability

6.1 Introduction

Compression, tension, and friction are the most common mechanisms of how mechanical forces are applied and transduced during biological functions, ranging from cell-cell adhesion and muscle contraction to protein degradation and translocation [102–105]. Protein structures in control of such biological processes evolved via the optimization and discovery of new topologies, in order to fulfill their function under such mechanical constraints [106–108]. Thus, mechanical properties of proteins might play the role of a constraint or driving force during protein evolution. Identifying driving forces that recruit new topologies will help to understand how the current protein universe was shaped [109–113]. Physical and chemical factors [114, 115] molding the protein structure catalog need to be elucidated. Deciphering what are the crucial factors contributing to the evolutionary history of the protein structures is a relevant question toward a better understanding of the mechanism of evolution. One such factor could be mechanical stability, i.e. the ability to withstand forces, which can be measured by pulling a protein by both ends. The force applied for unfolding a protein can be monitored during protein extension. As a result, a force-extension curve describing the unfolding pathway of the molecule can be examined

(Figure 4.3). Peaks in the force-extension curve represent forces required to rupture critical building block such as a force clamp. Force clamps are an assembly of hydrogen bonds formed by a core of residues responsible for the mechanical stability of the protein. In other words, they are an area of proteins which possesses a high mechanical propensity. Unfolding pathways are therefore highly dependent on the shape of and interactions within the protein. Examining the topology adopted by a protein can give valuable information on its mechanical stability. A common example of such topology is the shear topology which possesses a high mechanical stability [116, 117]. It features two force-bearing β-strands arranged in parallel and is particularly adapted to withstand a stretching force. This topology is used in several proteins having different functions requiring mechanical properties. Therefore, evolutionary processes driving selection and optimization of protein topology is likely to have occurred and probably still occurs. Among those external forces, the ability to oppose tension is of great biological interest. Nature might have selected protein structures to resist such forces using evolutionary mechanisms such as mutations or recombinations (Figure 6.1).

One example of a protein with mechanical function is the giant muscle protein titin. Titin is a structural protein composed of 244 individual protein domains that takes advantage of the shear topology to confer resilience and elasticity to muscle fibers [118]. Since muscle fibers and also titin therein are only present in multicellular organisms and might have originated from contractile cells in sponge-grade organisms [119](with a common ancestor probably 700 million years ago), they are certainly the result of an optimization of protein folds. The association of small but highly mechanically stable domains constitute titin and is a key factor in defining its properties.

How evolutionary pressure, such as mechanical force, influenced the demand for new structural topologies and functions is still an open question. Using available genomic data and protein structure models acquired via ex-

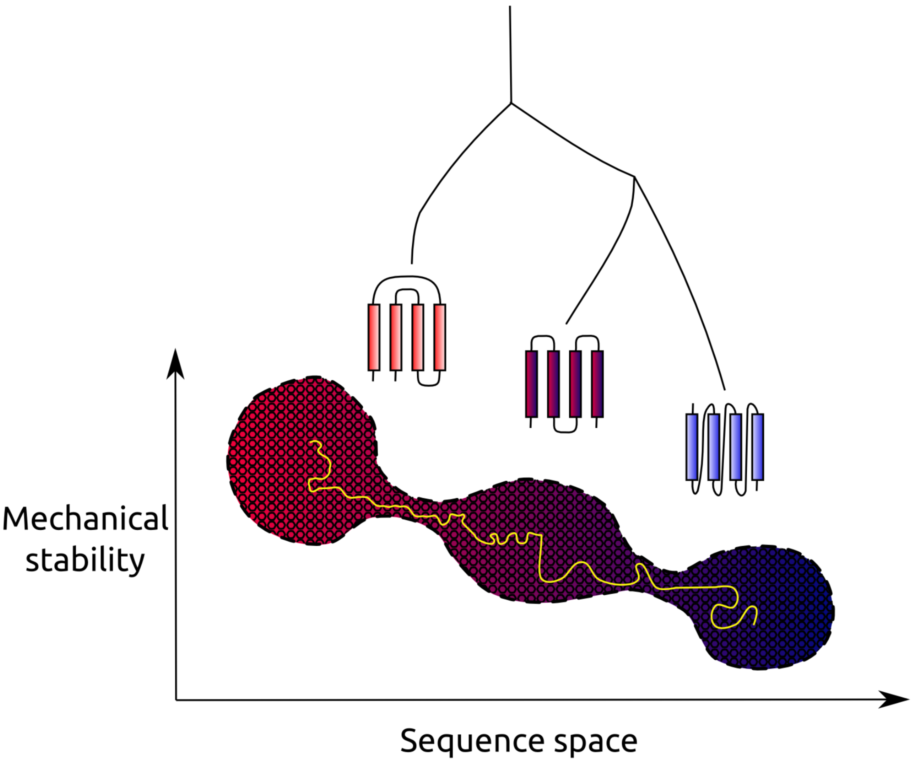


Figure 6.1: Scheme representing a possible path of protein structure evolution.

periments, combined with techniques to study mechanical properties of proteins in a high-throughput way can now open new avenues of inquiry. Over the past decade, new experimental techniques such as magnetic and laser tweezers or Atomic Force Microscopy (AFM) have enabled the characterization of mechanical properties of proteins at the single molecule level [33,120] (Section 4.3.2). However, AFM experiments of protein unfolding are too demanding to systematically evaluate the mechanical stability of the available structural models of protein domains, currently ∼100,000 in total. Such a survey, instead, has been conducted by computational means on proteins [121]. We here followed a similar computational approach to evaluate mechanostability. Our study differs from the previous survey in three aspects: first, we used the definition of protein domains by SCOP (Section 3.6.1), and second, we also considered domains with a size of up to 400 a.a. Thirdly, and most importantly we finally mapped the aquired data on an phylogenomic tree of SCOP domains. A domain based analysis, as provided by our studies, allows a better understanding of the connection between topology and mechanical stability by removing possible combination of different topologies in proteins.

We aimed at estimating the rupture force of all ∼100,000 SCOP domains as a measure for mechanical stability, using computer simulations mimicking force spectroscopy experiment. It remains virtually impossible to reach the biologically relevant millisecond to second timescales using theoretical pulling rates, even for a very small system of a few tens of residues, using conventional all-atom Molecular Dynamics (MD) methods in explicit and implicit solvent (water) implemented on the most powerful distributed computer clusters. Brownian Dynamics (BD) simulations using a Go-model [122] as introduced in Section 4.3.1 is a mesoscopic method, in which explicit solvent molecules are replaced by a stochastic force. This technique has the advantage to allow simulations on much larger time scales than MD simulations [83], and was the method of choice here. Its high computational efficiency allows to evaluate the dynamic response of ∼100.000 SCOP domains to a tensional force. Loading rates can be chosen such that timescales are close to those of the experimental studies.

Here, we reveal evolutionary patterns of mechanical stability by mapping the peak force of the force-extension curve F_{max} onto time-lines derived from phylogenomic trees of domain structures [18]. Remarkably, we find a selection pressure to decrease the overall mechanical stability, and on the other hand to increase the ratio of mechanical stability to protein length during the overall protein history. Our results suggest a reduction of the material without a loss of mechanical stability, leading to a forced optimization of mechanical clamps, or reflecting a need for more compact protein domains to become compatible with a multi-domain protein world [19], allowing the rise of multi-cellular organisms and Eukaryota lineage. Our studies yield valuable new information on the mechanical properties of domains. We identify specific force-resistant topologies and their emergence during evolution, as an answer to the need for abilities to transmit and withstand forces in a biological context. In addition, studies on supposedly non-mechanical proteins yielded interesting results about their behavior under forces [123, 124], suggesting mechanical properties to be a critical aspect of protein structures in general.

6.2 Results

6.2.1 Change in mechano-stability during evolution

To trace mechanical stability in evolution, we determined the mechanical unfolding forces of protein domain structures at the Family (F) level of structural organization. Figure 6.2a shows the mechanical stability of each F, as measured by its average maximum peak force (F_{max}), as a function of evolutionary time. Using polynomial regression, we observed a significant decrease (p-value = 2.0e-16) in F_{max} in proteins appearing between ∼3.8 and ∼1.5 billion years ago (Gya). Trends were maintained when studying domain evolution at more or less conserved levels of structural abstraction of SCOP hierarchy. We found a significant decrease of mechanical stability at the level of domains with less than 40 % or 95 % sequence identity (p-value = 2.0e-16). Consistent results were also obtained at the F level using linear regression. Within a smaller dataset of only 13 proteins, for which mechanical stabilities have been measured, we find that the experimental F_{max} does not exhibit any significant tendency early in evolution. Experimental data for at least 50-60 proteins would be needed to validate our computional results on the large data set.

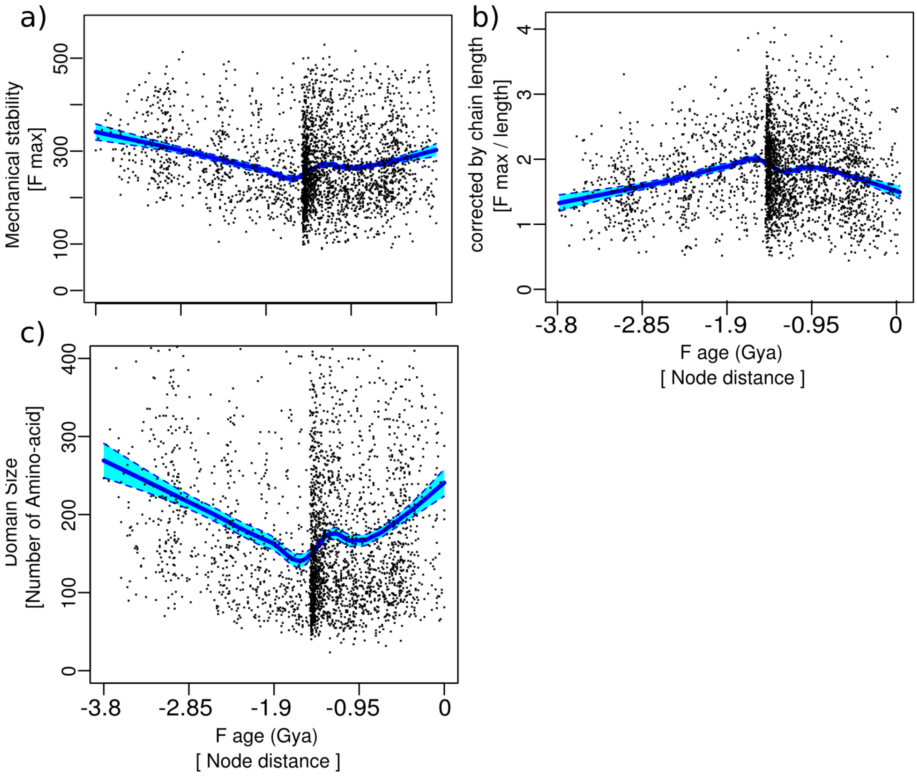


Figure 6.2: Change in length and mechanical stability during evolution: a) Mechanical stability (F_{max}) b) Mechanical stability corrected by chain length (F_{max}/l) versus approximative F domain age in billion of years (Gya). Each data point represents a mechanical stability average of domains belonging to the same F. c) Average amino-acid chain length for domains belonging to the same F versus F domain age in Gya. The solid line shows a LOESS polynomial regression, and the blue shade the 95% confidence interval. The appearance of many new F at 1.5 Gya corresponds to the so-called big-bang of protein stuctures.

6.2.2 Length, SMCO, and evolution of mechano-stability

At the same time during which we observe a decay in rupture forces, protein domains decrease in size suggesting that decrease in mechanical stability is only due to trend in evolution for domain size reduction. Aiming at testing how length influences the variation of mechanical stability, we divided mechanical stability by chain length. Figure 6.2b shows the mechanical stability of each F, as measured by its F_{max} corrected by size, as a function of evolutionary time. We observed a decrease of the ratio of mechanical stability and chain length before 1.6 Gya, suggesting an optimization of amino-acids usage to retain a relative mechanical stability of proteins while reducing the amount of material required. In other words, evolution might have favored the production of compact protein structures but on an only minor expense of mechanical stability. This trend was followed by a minor but significant increase in F_{max} in proteins appearing between 1.5 Gya to present day proteins. Again, trends were maintained when studying domain evolution at the level of domains with less than 40 %, 95 % sequence identity (p-value = 2.0e-16). The decrease of absolute mechanical stability during the first part of evolution suggests that the ability to withstand forces decreased in this interval. Interestingly, when looking at the tightness of proteins, which represents how compact the network of interactions in protein structures is, we observed a decrease suggesting that protein structures lost mechanical stability while becoming more compact. On the contrary, when we divide tightness and mechanical stability by chain length, we observe an increase in both ratios of mechanical stability and tightness divided by chain length. The length of the amino acid chain has already been reported to influence the mechano-stability of a protein, with longer chains creating a higher resistance to force [125]. We therefore asked whether the decrease in F_{max} and increase in F_{max}/L (Figure 6.2a,b) we observed from ∼3.8 to ∼1.5 Gya can be explained by a decrease in the chain length of proteins. Figure 6.2c shows how domain size has varied in evolution. Mechano-stability measured by F_{max} and domain size follow a very similar bimodal trend, with a clear decrease occurring prior to ~1.5 Gya and a slight increase after the "big bang". In accordance, we found the domain size and F_{max} to be correlated (Pearson correlation coefficient: 0.74; Figure 6.3). Taken together, those results suggest that nature optimized the number of amino-acids used for protein structures, while partly maintaining tight interactions and mechanical stability. This trend might reflect the pressure of an environment with limited resources. During the second part of evolutionary history (after ~1.5 Gya) we observe the opposite tendencies. It has been shown that this time marks the increase of folds belonging to the Eukaryota lineage. New folds are among others related to the extracellular matrix [126,127], as required by multicellular organisms, for cell-cell interactions and junctions. Adhesion, motility, and matrix proteins evolved to develop and spawn anisotropic and mechanically resilient scaffolds between and within cells.

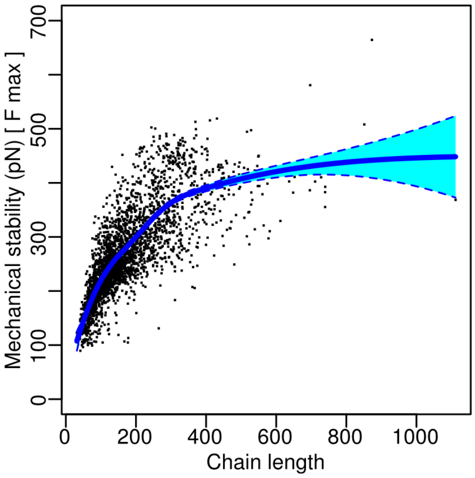


Figure 6.3: Chain length versus mechanical stability, F_{max}. A non-linear dependency is observed, which can be approximated by a linear dependency for small (200 aminoacids) proteins. The solid line shows a LOESS polynomial regression, and the blue shade the 95 % confidence interval.

Next, we eliminated the effect of domain size on evolutionary trends in mechano-stability by dissecting our dataset according to the amino acid chain length, in order to analyze other factors. This analysis was done with all ~92,000 domains to ensure enough data points for each length.

The distributions of chain length are shown in Figure 6.4a and b for the

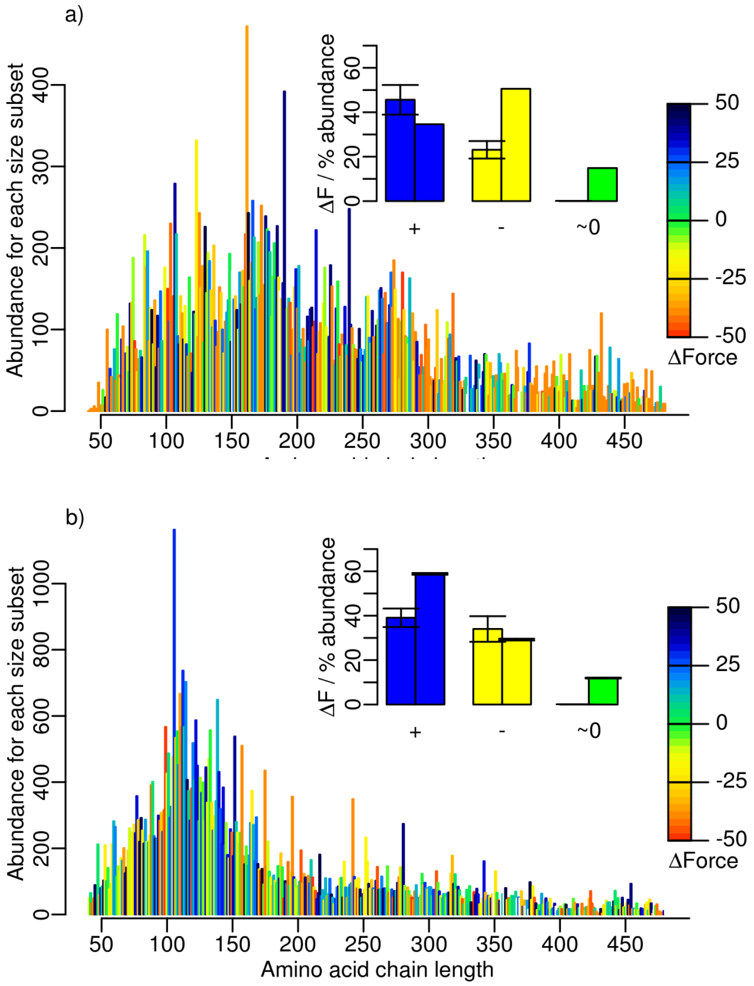


Figure 6.4: Change in foldability during evolution for subsets of chain size: Distribution of domain length for domains appearing a) 3.8-∼1.5 Gya and b) ∼1.5-0 Gya. Abundancies were colored according to the average ΔF_{max}, the difference between the end points of the polynomial regression of F_{max} in this dataset, for the specified initial (a) and later (b) time period. Yellow to red indicates a decrease, and blue an increase in F_{max}. The barplots (inset) show the Δ averaged of subset (left) and the percentage of domains (rigth) with positive (blue), negative (yellow), and insignificant (green) ΔF_{max}.

two time periods before and after the "big bang" (∼1.5 Gya). The length distribution for proteins appearing before the "big bang" exhibited a peak at around ∼150 amino acids, and shifted later (∼1.5 Gya to the present) to shorter chains with a peak at around 100 amino-acids, showing a decrease of domain size. We note that the resulting average chain length of three-dimensional structures in SCOP, which have been obtained from X-ray or NMR measurements, is smaller than the average length of sequences in genomes [83], due to the increasing experimental difficulties when working with large proteins. These tendencies are consistent with studies revealing that conserved protein domains have a longer length [128] and also agree with the theory that 200 amino-chains represent a barrier for the physical force helping folding [129, 130]. The need for smaller force with equivalent length was then required as observed in our results. Then we analyzed evolutionary tendencies for every domain length subset by measuring the variation in the end points of a polynomial regression. The color mapping in Figure 6.4a indicates an increase (blue), a decrease (yellow-red), or a non-significant change (green) of mechanical stability. During early protein evolution (3.8-1.5 Gya), we found that 52 % ± 0.3 % of all domains in each size subset decrease their mechano-stability during evolution. Conversely, 30 ± 0.4 % of domains showed an increase in mechano-stability, i.e. a significant increase in F_{max}. These results confirm the tendencies observed for the full data set (Figure 6.2a), and hold for different thresholds of identity, namely 95 % and 40 %. As expected, due to the smaller data set, partitioning domains defined at F and SF levels according to size yielded results that were not statistically significant. In summary, even after dissecting the effect of the chain length on changes in F_{max}, the tendency of proteins to lose mechano-stability during the evolution is confirmed.

6.2.3 Evolution of β-architectures

With the aim of understanding how topology may have affected the evolution of mechano-stability, we analyzed three specific β topologies: barrel, sandwich and pseudo-barrel. We choose β-topologies because of their outstanding mechanical stability, e.g. observed for the β-barrel GFP [131] and for the β-sandwich immunoglobulin [132]. Figure 6.5 shows the evolution of these β-topologies. We observed that the apparition of the barrel topology occurs prior to sandwich topology apparition.

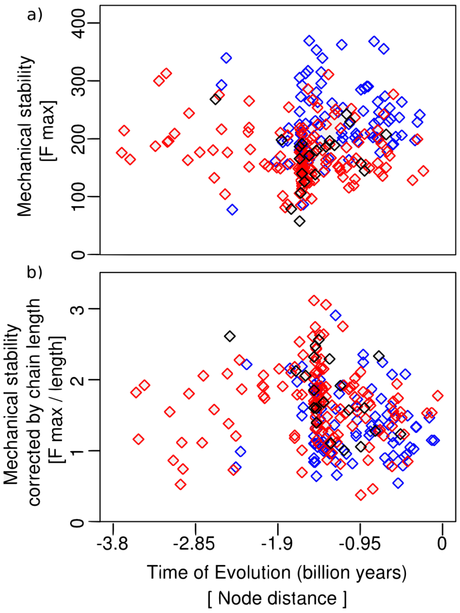


Figure 6.5: Change in mechanical stability for β-class proteins a) Mechanical stability (F_{max}) and b) Mechanical stability divided by length (F_{max}/l). Red diamonds are barrels, blue diamonds are sandwiches, and black diamonds represent pseudo-barrels.

A barrel topology is often related to channel functions, such as the exchange between outside and inside of the cell [133], while a sandwich topology can carry out functions such as motion, cell recognition or mechanical scaffolding [134–139] that are known to have appeared only later when Eukaryota kingdom arose [140,141]. Taken together, those results suggest that nature selected topologies for required functions, thereby potentially selecting β-barrels such that two sheets flattened to form first a pseudo-barrel and later a sandwich.

Aiming at tracking such transformations, we measured the structural similarity between β-topologies using SPalign [142]. Figure 6.6 represents a similarity graph of all-β domains with relative positions according to their evolutionary age (nd) and mechanical stability. Several pathways possibly representing transitions of protein structures from between topologies can be observed. Together with the successive apparition of those topologies in evolutionary time, our results suggest a continuous transition in protein topological space driven by functions such as high mechanical resistance.

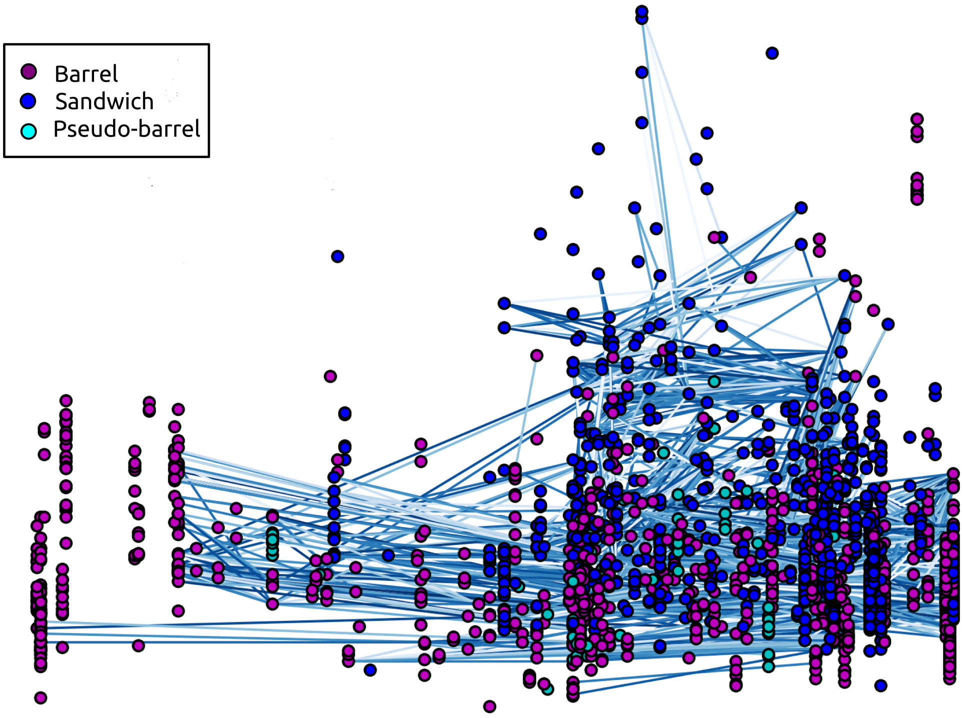


Figure 6.6: Graph representation of similarity between all-β domains: x-axis: evolutionary time (nd), y-axis: mechanical stability (F_{max})

6.3 Discussion

Protein mechano-stability is generally associated with functions related to mobility, force transmission and structural integrity. Those functions are required in particular in multi-cellular organisms to allow communication, adhesion, and mobility of the whole organism. Transition of protein from an intracellular to an extracellular environment may cause a higher need for mechanical stability due to the increase of forces acting on the protein. The later expansion of a uni-cellular to a multi-cellular organism may account for the increase in mechanical stability seen in our study. The turnover occurs at the same time as the protein "big bang", a period of time that exhibits an abrupt increase in the number of domain architectures and shuffling within multi-domain proteins set off by increased rates of domain fusion and fission. In our analysis, we considered only the highest peak of force required to pull a protein as a reference for mechanical stability.

In this aspect, protein structure nearly maintained their mechanical stability over the first period in spite of losing amino acids. This suggests that a simplification of protein structures occurred during this period of time. Despite the increase of mechano-stability after the "big bang" our results show that the evolutionary pressure for mechanically stable protein is not globally applied onto every protein structure, but that mechanical stability will increase in specific protein families, depending on requirements for protein function and other evolutionary pressures. Classification of force profiles could help us learn more about specific protein families and their evolution. More precisely, we would like to have a better understanding of the correlation of force peaks with set of residues in a given structure possibly by using WLC theory [143]. This could improve the classification of different fore clamps, and could yield a more detailed picture of the evolutionary selection of stable proteins.

Single protein folds do not correspond to individual functions, that is to say, the same fold could have many functions or a similar function in two instances may require different protein folds. Our work examines the relationship between structure and function and may be used to uncover folds that might be applicable to nanotechnology [144]. However, mechanical strength is a function not necessarily important to an equal extend throughout the protein repertoire. As a consequence, we find mechanical stability to follow more complex trends as compared to folding, in particular when separating protein size effects in contrast to a more locally applied pressure for mechanical performance. This may be a result of a more global evolutionary pressure applied to protein folding time in contrast to a more locally applied pressure for mechanical performance

In this study, we only considered pulling along the termini axis. As shown previously in different studies, [145] the variation of the pulling direction may impact the proteins response to force. Therefore this study describes the evolution of protein structure for resistance to tensional force propagating to the protein through the termini. Future studies using different pulling velocities can further complete our understanding of how mechanical properties evolved.

Solenoid proteins represent one of the possible ancestors of intrinsically disordered proteins (IDP) and are found late in evolution (0.95 nd) suggesting the late appearance of IDPs. Their interesting mechanical properties such as high elasticity and extensibility, cannot be covered by this study but would be an interesting aspect to examine in future studies [146]. Apart from IDPs, we note that the current protein structure (PDB) database is predicted to include already the majority of protein structures, and that the rate of discovery of novel structures has declined over the last 2 years [100]. Further studies on augmented protein structure data sets, on structures beyond single domains, or also on other interesting protein features such as hydrophobicity [147, 148], would further elucidate the possible evolutionary constraints on protein structure.

Chapter 7

Multivariate analysis of physical constraints on protein structures

Chapters 5 and 6 analyzed evolutionary trends of physical constraints, namely foldability and mechanical stability. We next asked how those potential constraints depend on each other and also vary with other attributes such as protein flexibility, function, localization, or secondary structure composition. These results give insight into the property space sampled by today's protein structural repertoire.

7.1 Mapping between localization, function, and domains

We mapped protein domains and associated physical measures to functions and localizations using data from Gene Ontology (Go) [1] [149]. Go associates every gene or protein with controlled vocabulary terms. Vocabulary terms are split into three main branches: cellular component, molecular function, and biological process. For every PDB or chain, several Go IDs are available for each ontology. They correspond to the different levels of definition for a given ontology (e.g. for the cellular component, Figure 7.1). Using the Go, we assigned the localization and function to domains by mapping between pdb structures to their Go [2]. The mapping covered about half of our dataset ($\approx$ 50,000 domains).

[1] http://www.geneontology.org/

[2] https://lists.sdsc.edu/pipermail/pdb-l/2011.../005385.html.

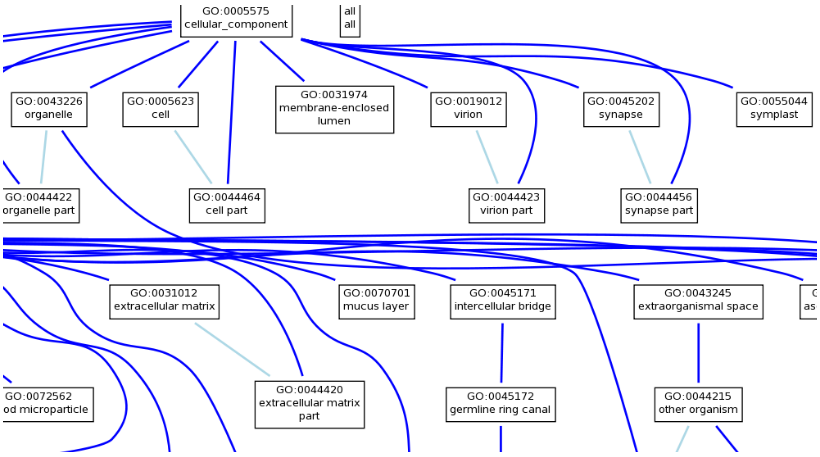


Figure 7.1: Portion of the cellular component tree structure of Go.

7.1.1 Mapping between Taxa and protein domains

A mapping between taxid and taxa name was obtained from querying NCBI taxonomy files [3] [150]. From the NCBI data, tree structures for every species were collected. A tree structure contains several layers of every species, such as: domain, kingdom, phylum, class, order, family, genus, and species.

The Star Data Model [151] (Figure 7.2) is composed of one central fact table called Domain. The fact table contains all the physical quantities, or any other values related to a given domain. Surrounding tables are called dimension tables. Each dimension corresponds to an analysis axis, in other words to criteria that are relevant for data analysis. The dimensions here are the SCOP classifications comprising four layers. (namely the general composition, the fold class, the superfamily, and the family), the localization, the function, the taxon (evolution).

Then, queries can be built such as a comparison of average foldability (Section 4.2.1) against localization for each domain of life.

[3] http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

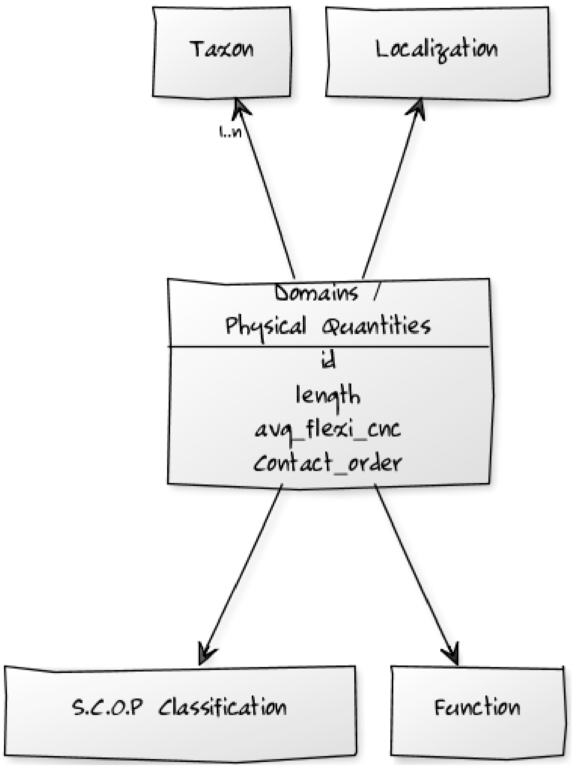


Figure 7.2: Star scheme model with fact table domains (id, length, etc) surrounded by dimensions, which here are taxon, localization, SCOP classification, and function.

7.2 Multivariate analysis

We first analyzed the covariation of foldability as measured by the Size-Modified Contact Order (SMCO, Section 4.2.1 and Chapter 5) with mechanical stability measured from unfolding forces (F_{max}, Section 4.3, Chapter 6) Interestingly, S.C.O.P domains with mixed α/β structures show a slight tendency for an increase of foldability that involves a decrease of mechanical strength (Figure 7.3). Moreover, we can observe a lower contact order for high values of rupture force for the purely β-class structures. This tendency could be explained by the fact that a low Contact Order translates into a high number of short range or local contacts (Section 4.2.1), which in contrast to non-local contacts can very efficiently increase the global stability of protein structures. On the contrary, from a global perspective, purely β-sheet structures show both overall higher mechanical strength and contact order, as compared to all-α proteins. This underlines that other factors influence the mechanical strength of a domain, such as architecture or fold (Section 4.2.1).

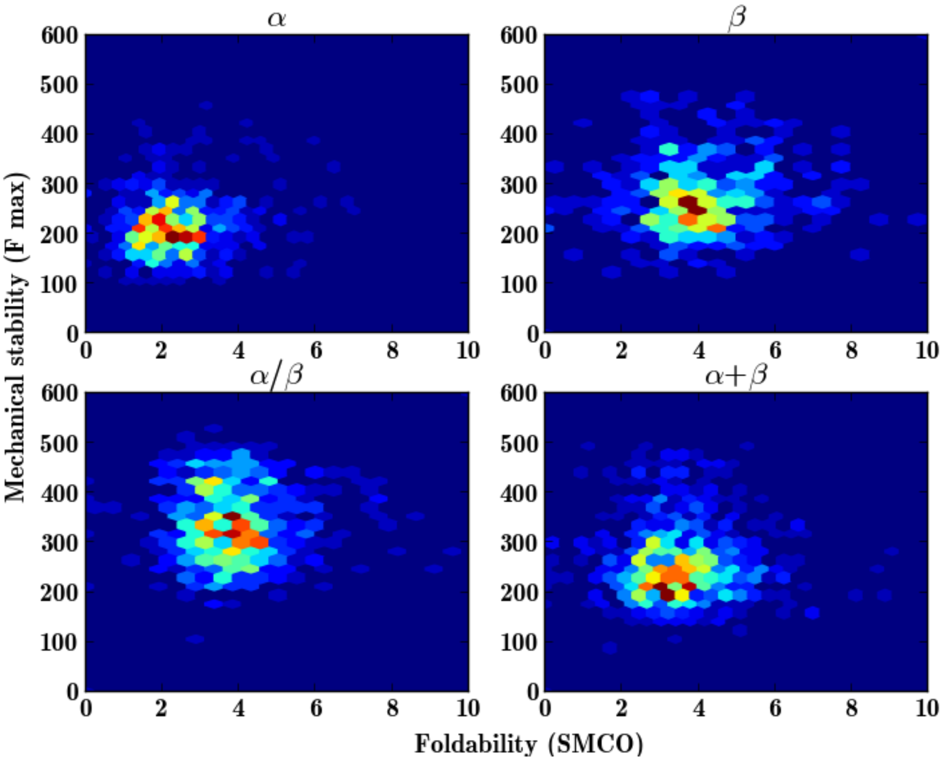


Figure 7.3: Domain distributions according to their relative Size Modified Contact Order and rupture force (F_{max}). The color code corresponds to the density of domains at a given coordinates, with red for high to blue for low probabilities.

We next compared the mechanical strength to the flexibility of domains, measured by the structural deviations (RMSD) within the conformational ensemble of a domain (Section 4.1). Unsurprisingly, an increase of flexibility generally involves a decrease of mechanical strength (Figure 7.4). We could not detect pronounced differences in flexibility for different secondary structure classes. Interestingly, domains distribute into some distinct areas, which could contain domains with similar properties or evolutionary connections, an observation, which requires further investigation.

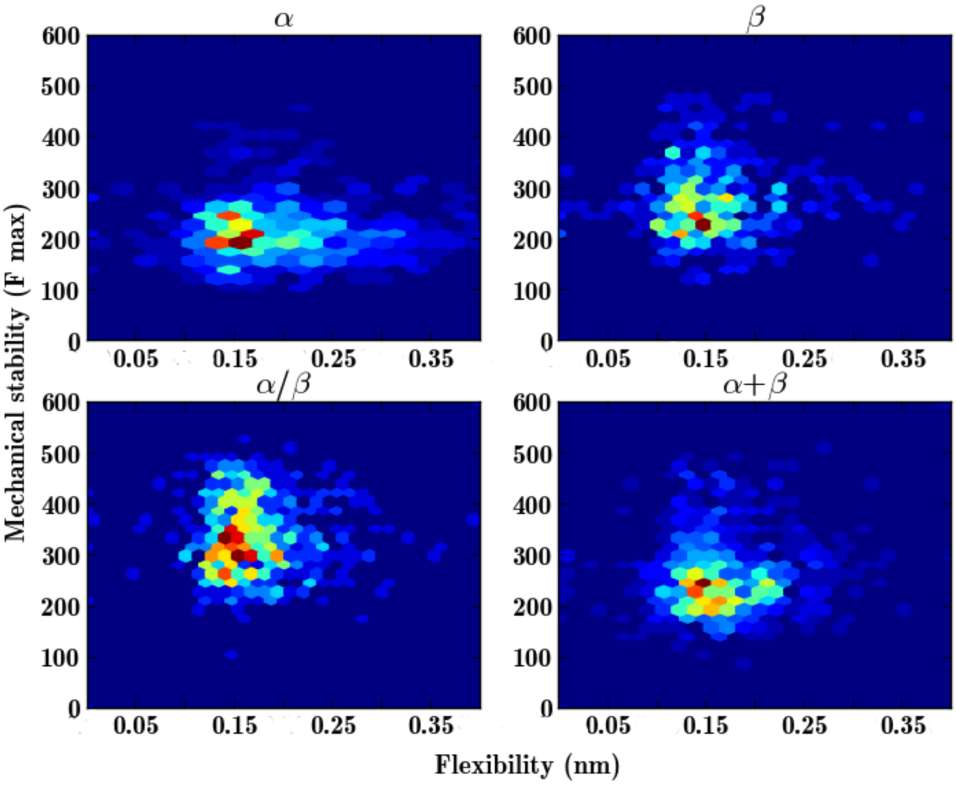


Figure 7.4: Domain distributions according to their flexibility (RMSD) and rupture force (F_{max}). Increase of flexibility involves a decrease of rupture forces for all structural classes.

Finally, we analyzed how flexibility and foldablity, i.e RMSD and SMCO, covary (Figure 7.5). SCOP domains with only α-helical structures show a higher flexibility combined with a lower contact order than mixed or purely β-sheet structures. A remarkable trade-off is observed between flexiblity and foldability. Thus, rigid proteins are designed from many long-range contacts, while flexible proteins are held together by rather local contacts.

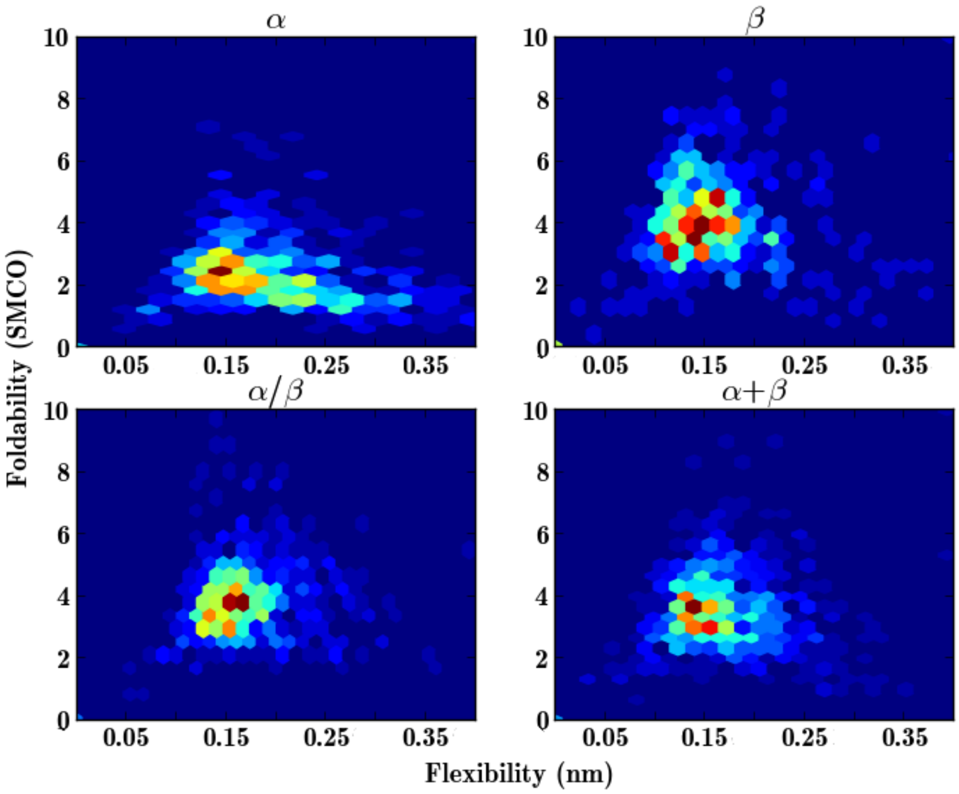
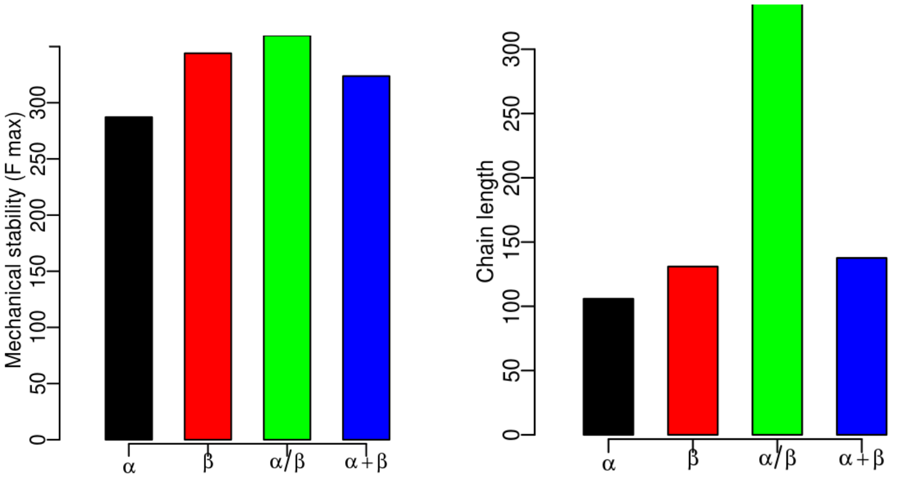


Figure 7.5: Domain distributions according to their flexibility (RMSD) and foldability (SMCO). Increase of flexibilty involves a decrease of folding time for all structural classes. Also, β-sheets fold more slowly than helical proteins.

Those results bring to light that resistance to tensile strength implies a longer folding time, and flexibility a shorter folding time, when considering global differences between protein classes.

7.3 Mechanical strength, fold classes, and localization of proteins

In Chapter 6, we observed a decrease in mechanical stability during evolution until the big bang ~1.5 Gya. We argued that this tendency was mainly due to the loss in protein size, as we observe an overall increase of F_{max} per amino-acid. We next asked how different structural composition classes of SCOP, namely all-α, all-β, mixed α/β and segregated α+β, differ in these aspects. We find mixed α/β structures to show the highest average mechanical force (341 pN) followed by β-sheet structures (283 pN), and α+β structures (244 pN) (Figure 7.6a). These values have to be compared to the average length of each class (Figure 7.6b).



(a) Average mechanical strength.

(b) Average length (in amino-acids).

Figure 7.6: Average mechanical strength (F_{max}) and length (in amino-acids) according to S.C.O.P general composition classes.

The ratio force/length (F_{max}/N, Figure 7.7) shows a different order as compared to Figure 7.6a. The β-sheet structures possess the highest ratio, followed by mixed α/β structures. Thus, in agreement with experimental observations, β-sheet structures outperform others in terms of mechanical resistance, and have been possibly designed for this function at least partly.

Finally, we asked the question if the protein localization has an effect on the average mechanical stability of proteins, which would suggest an adaptation of proteins to the mechanical stress present in their respective

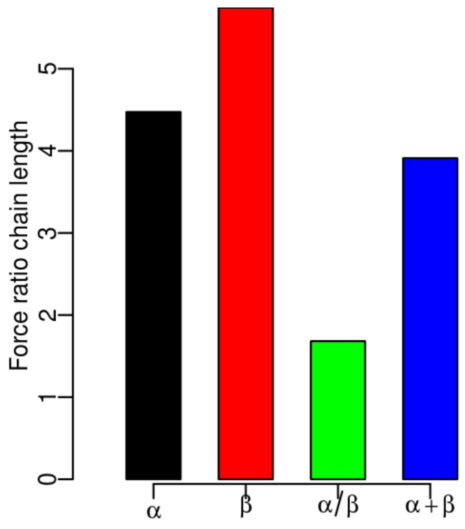


Figure 7.7: Relative mechanical strength (F_{max}/N) obtained from normalizing F_{max} by protein length N, for SCOP general composition classes.

environment. Figure 7.8 shows that intra-cellular domains clearly possess a lower mechanical stability as compared to extracellular domains, but the distribution is much more diffuse.

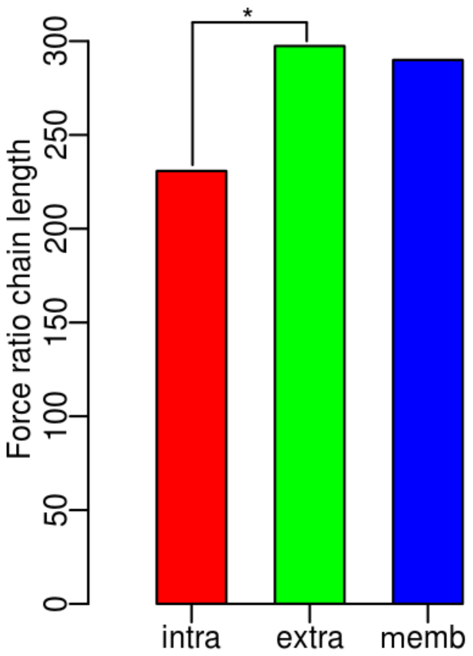


Figure 7.8: Relative mechanical strength (F_{max}/N) for three different localizations: intra-cellular, extra-cellular, and plasma membrane.

The average F_{max} for localizations in the Endoplasmic Reticulum, the lysozyme, microtubuli, the Golgi, sarcolemma, and mitochondria is very similar. One exception is the lysosome, hosting proteins with significantly higher mechanical strength than the other compartiments, the reason of which remains to be elucitated.

Bibliography

[1] Oparin AI (1961) The origin of life. Nord Med 65: 693–697.

[2] Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134: 191–203.

[3] Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, et al. (1999) Structural genomics: beyond the human genome project. Nat Genet 23: 151–157.

[4] Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311: 347–351.

[5] Caetano-Anollés G, Caetano-Anollés D (2003) An evolutionarily structured universe of protein architecture. Genome Res 13: 1563–1571.

[6] Doolittle RF (1995) Of archae and eo: what's in a name? Proc Natl Acad Sci U S A 92: 2421–2423.

[7] Doolittle RF (1995) The multiplicity of domains in proteins. Annu Rev Biochem 64: 287–314.

[8] Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE (2009) The origin, evolution and structure of the protein world. Biochem J 417: 621–637.

[9] Chothia C, AG M, L LC, A A, D H, et al. (1995) J Mol Biol 247: 536-540.

[10] Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, et al. (2013) New functional families (funfams) in cath to improve the mapping of

conserved functional sites to 3d structures. Nucleic Acids Res 41: D490–D498.

[11] Hou J, Jun SR, Zhang C, Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci U S A 102: 3651–3656.

[12] Holm L, Sander C (1996) Mapping the protein universe. Science 273: 595–603.

[13] Osadchy M, Kolodny R (2011) Maps of protein structure space reveal a fundamental relationship between protein structure and function. Proc Natl Acad Sci U S A 108: 12301–12306.

[14] Zotenko E, Dogan RI, Wilbur WJ, O'Leary DP, Przytycka TM (2007) Structural footprinting in protein structure comparison: the impact of structural fragments. BMC Struct Biol 7: 53.

[15] Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31: 45–71.

[16] Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134: 167–185.

[17] Caetano-Anollés G, Kim KM, Caetano-Anollés D (2012) Erratum to: The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. J Mol Evol .

[18] Wang Y, Wang S, Gao YS, Chen Z, Zhou HM, et al. (2011) Dissimilarity in the folding of human cytosolic creatine kinase isoenzymes. PLoS One 6: e24681.

[19] Wang M, Caetano-Anollés G (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure 17: 66–78.

[20] Kimura M (1979) Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Acad Sci U S A 76: 3440–3444.

[21] Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.

[22] Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A 101: 9205–9210.

[23] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, et al. (2005) Evolutionary information for specifying a protein fold. Nature 437: 512–518.

[24] Russell RB (1994) Domain insertion. Protein Eng 7: 1407–1410.

[25] Graur WH Dan; Li (2000) Fundamentals of Molecular Evolution: Second Edition. Sunderland, Massachusetts: Sinauer Associates.

[26] Ohno S (1967) Sex Chromosomes and Sex-linked Genes. springer.

[27] Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420: 218–223.

[28] Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 313: 673–681.

[29] Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 313: 903–919.

[30] Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. Horizons in Biochemistry. Academic Press, New York, 189–225 pp.

[31] Ayala FJ (1999) Molecular clock mirages. Bioessays 21: 71–75.

[32] Katju V (2012) In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. Int J Evol Biol 2012: 341932.

[33] Galera-Prat A, Gómez-Sicilia A, Oberhauser AF, Cieplak M, Carrión-Vázquez M (2010) Understanding biology by stretching proteins: recent progress. Curr Opin Struct Biol 20: 63–69.

[34] Söding J (2005) Protein homology detection by hmm-hmm comparison. Bioinformatics 21: 951–960.

[35] Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology, 406-416 pp.

[36] Gerstein M, Hegyi H (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 22: 277–304.

[37] Gerstein M (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins 33: 518–534.

[38] Pastore A, Lesk AM (1990) Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. Proteins 8: 133–155.

[39] Schnuchel A, Wiltscheck R, Czisch M, Herrler M, Willimsky G, et al. (1993) Structure in solution of the major cold-shock protein from bacillus subtilis. Nature 364: 169–171.

[40] Yule GU (1925) A mathematical theory of evolution. Philosophical Transactions of the Royal Society of London 213: 402-410.

[41] Letunic I, Bork P (2011) Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39: W475–W478.

[42] Eisenberg D (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A 100: 11207–11210.

[43] Pauling L, Corey RB (1951) Configuration of polypeptide chains. Nature 168: 550–551.

[44] Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552–558.

[45] Phillips DC (1966) The three-dimensional structure of an enzyme molecule. Sci Am 215: 78–90.

[46] Drenth J, Jansonius JN, Koekoek R, Swen HM, Wolthers BG (1968) Structure of papain. Nature 218: 929–932.

[47] Chothia C (1992) Proteins. one thousand families for the molecular biologist. Nature 357: 543–544.

[48] Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.

[49] Levitt M (2007) Growth of novel protein structural data. Proc Natl Acad Sci U S A 104: 3183–3188.

[50] de Leeuw M, Reuveni S, Klafter J, Granek R (2009) Coexistence of flexibility and stability of proteins: An equation of state. PLoS ONE 4: e7296.

[51] Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design 2: 173 - 181.

[52] de Groot B, van Aalten D, Amadei R, Scheek A, Vriend G, et al. (1997) Prediction of protein conformational freedom from distance onstraints. Proteins: Structure, Function, and Bioinformatics 29: 240 - 251.

[53] Tolman JR (1997) J R Tolman 4: 292-297.

[54] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, et al. (1977) The protein data bank. a computer-based archival file for macromolecular structures. Eur J Biochem 80: 319–324.

[55] Yang LW, Eyal E, Chennubhotla C, Jee J, Gronenborn AM, et al. (2007) Insights into equilibrium dynamics of proteins from comparison of nmr and x-ray data with computational predictions. Structure 15: 741–749.

[56] Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277: 985–994.

[57] Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, et al. (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12: 2057–2062.

[58] Abe H, Go N (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins. ii. application to two-dimensional lattice proteins. Biopolymers 20: 1013–1031.

[59] Thirumalai D, Klimov DK (1999) Emergence of stable and fast folding protein structures. Technical Report cond-mat/9910248.

[60] Thirumalai D (1995) From minimal models to real proteins: Time scales for protein folding kinetics. J Phys I France 5: 1457-1467.

[61] Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, et al. (2008) Data growth and its impact on the scop database: new developments. Nucleic Acids Res 36: D419–D425.

[62] Qiu L, Pabit SA, Roitberg AE, Hagen SJ (2002) Smaller and faster: the 20-residue trp-cage protein folds in 4 micros. J Am Chem Soc 124: 12952–12953.

[63] Goldberg ME, Semisotnov GV, Friguet B, Kuwajima K, Ptitsyn OB, et al. (1990) An early immunoreactive folding intermediate of the tryptophan synthase $\beta 2$ subunit is a 'molten globule'. FEBS Letters 263: 51 - 56.

[64] Matagne A, Chung EW, Ball LJ, Radford SE, Robinson CV, et al. (1998) The origin of the alpha-domain intermediate in the folding of hen lysozyme. J Mol Biol 277: 997–1005.

[65] Onuchic JN, Wolynes PG (2004) Theory of protein folding. Curr Opin Struct Biol 14: 70–75.

[66] Levinthal C (1969) How to fold graciously. In: Debrunnder JTP, Munck E, editors, Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. University of Illinois Press, pp. 22–24.

[67] Nölting B, Schälike W, Hampel P, Grundig F, Gantert S, et al. (2003) Structural determinants of the rate of protein folding. J Theor Biol 223: 299–307.

[68] Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. Proteins 35: 408–414.

[69] Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A, et al. (2010) Exploring the universe of protein structures beyond the protein data bank. PLoS Comput Biol 6: e1000957.

[70] Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–196.

[71] Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. Curr Opin Struct Biol 14: 202–207.

[72] Ortiz AR, Skolnick J (2000) Sequence evolution and the mechanism of protein folding. Biophys J 79: 1787–1799.

[73] Caetano-Anollés G, Caetano-Anollés D (2005) Universal sharing patterns in proteomes and evolution of protein fold architecture and life. J Mol Evol 60: 484–498.

[74] Bowman GR, Voelz VA, Pande VS (2011) Taming the complexity of protein folding. Current Opinion in Structural Biology 21: 4 - 11.

[75] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334: 517–520.

[76] Bogatyreva NS, Osypov AA, Ivankov DN (2009) Kineticdb: a database of protein folding kinetics. Nucleic Acids Res 37: D342–D346.

[77] Ouyang Z, Liang J (2008) Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 17: 1256–1263.

[78] Kubelka J, Hofrichter J, Eaton WA (2004) The protein folding 'speed limit'. Curr Opin Struct Biol 14: 76–88.

[79] Cleveland WS, Devlin SJ (1988) Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83: pp. 596-610.

[80] Vendruscolo M, Dokholyan NV, Paci E, Karplus M (2002) Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlin Soft Matter Phys 65: 061910.

[81] Sancho DD, Doshi U, Muñoz V (2009) Protein folding rates and stability: how much is there beyond size? J Am Chem Soc 131: 2074–2075.

[82] Portman JJ (2010) Cooperativity and protein folding rates. Curr Opin Struct Biol 20: 11–15.

[83] Cieplak M, Xuan Hoang T (2000) Scaling of folding properties in go models of proteins. Journal of Biological Physics 26: 273-294.

[84] Felice FGD, Vieira MNN, Meirelles MNL, Morozova-Roche LA, Dobson CM, et al. (2004) Formation of amyloid aggregates from human lysozyme and its disease-associated variants using hydrostatic pressure. FASEB J 18: 1099–1101.

[85] Tanzi RE, Bertram L (2005) Twenty years of the alzheimer's disease amyloid hypothesis: a genetic perspective. Cell 120: 545–555.

[86] Ross CA, Poirier MA (2004) Protein aggregation and neurodegenerative disease. Nat Med 10 Suppl: S10–S17.

[87] Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8: 737–742.

[88] Ramanathan A, Agarwal PK (2011) Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol 9: e1001193.

[89] Hagen SJ, Hofrichter J, Szabo A, Eaton WA (1996) Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. Proc Natl Acad Sci U S A 93: 11615–11617.

[90] Jaenicke R (1991) Protein stability and molecular adaptation to extreme conditions. Eur J Biochem 202: 715–728.

[91] Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8: 319–330.

[92] Pauwels K, Molle IV, Tommassen J, Gelder PV (2007) Chaperoning anfinsen: the steric foldases. Mol Microbiol 64: 917–922.

[93] Bogumil D, Landan G, Ilhan J, Dagan T (2012) Chaperones divide yeast proteins into classes of expression level and evolutionary rate. Genome Biol Evol 4: 618–625.

[94] Vendruscolo M (2012) Proteome folding and aggregation. Curr Opin Struct Biol 22: 138–143.

[95] Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, et al. (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 4: 805–809.

[96] Li L, Shakhnovich EI (2001) Different circular permutations produced different folding nuclei in proteins: a computational study. J Mol Biol 306: 121–132.

[97] Jung J, Lee B (2001) Circularly permuted proteins in the protein structure database. Protein Sci 10: 1881–1886.

[98] Bliven S, Prlič A (2012) Circular permutation in proteins. PLoS Comput Biol 8: e1002445.

[99] Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, et al. (2006) Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure 14: 1489–1498.

[100] Wolf YI, Grishin NV, Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. J Mol Biol 299: 897–905.

[101] Muñoz V, Serrano L (1996) Local versus nonlocal interactions in protein folding and stability - an experimentalist's point of view. Folding and Design 1: R71 - R77.

[102] Leckband D (2000) Measuring the forces that control protein interactions. Annu Rev Biophys Biomol Struct 29: 1–26.

[103] bo Hu X, yun Cheng D, yi Zhang Y, li Fan L, xia Li X, et al. (2009) [effects of mechanical force on expressions of intercellular adhesion molecule-1 in cultured human alveolar type 2 cells]. Zhonghua Jie He He Hu Xi Za Zhi 32: 99–102.

[104] Liu M, Tanswell AK, Post M (1999) Mechanical force-induced signal transduction in lung cells. Am J Physiol 277: L667–L683.

[105] Yin J, Kuebler WM (2010) Mechanotransduction by trp channels: general concepts and specific role in the vasculature. Cell Biochem Biophys 56: 1–18.

[106] Cao B, Elber R (2010) Computational exploration of the network of sequence flow between protein structures. Proteins 78: 985–1003.

[107] Fu X, Yu LJ, Mao-Teng L, Wei L, Wu C, et al. (2008) Evolution of structure in gamma-class carbonic anhydrase and structurally related proteins. Mol Phylogenet Evol 47: 211–220.

[108] Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16: 399–408.

[109] Burley SK, Bonanno JB (2002) Structuring the universe of proteins. Annu Rev Genomics Hum Genet 3: 243–262.

[110] Dokholyan NV (2005) The architecture of the protein domain universe. Gene 347: 199–206.

[111] Godzik A (2011) Metagenomics and the protein universe. Curr Opin Struct Biol 21: 398–403.

[112] Grabowski M, Joachimiak A, Otwinowski Z, Minor W (2007) Structural genomics: keeping up with expanding knowledge of the protein universe. Curr Opin Struct Biol 17: 347–353.

[113] Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. Genome Biol 5: 107.

[114] Serohijos AWR, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep 2: 249–256.

[115] Mulkidjanian AY, Galperin MY (2007) Physico-chemical and evolutionary constraints for the formation and selection of first biopolymers: towards the consensus paradigm of the abiogenic origin of life. Chem Biodivers 4: 2003–2015.

[116] Silva IR, Reis LMD, Caliri A (2005) Topology-dependent protein folding rates analyzed by a stereochemical model. J Chem Phys 123: 154906.

[117] Jung J, Lee J, Moon HT (2005) Topological determinants of protein unfolding rates. Proteins 58: 389–395.

[118] Li H, Cao Y (2010) Protein mechanics: from single molecules to functional biomaterials. Acc Chem Res 43: 1331–1341.

[119] Steinmetz PRH, Kraus JEM, Larroux C, Hammel JU, Amon-Hassenzahl A, et al. (2012) Independent evolution of striated muscles in cnidarians and bilaterians. Nature 487: 231–234.

[120] Forman JR, Clarke J (2007) Mechanical unfolding of proteins: insights into biology, structure and folding. Curr Opin Struct Biol 17: 58–66.

[121] Sikora M, Sułkowska JI, Cieplak M (2009) Mechanical strength of 17,134 model proteins and cysteine slipknots. PLoS Comput Biol 5: e1000547.

[122] Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with paup*. Curr Protoc Bioinformatics Chapter 6: Unit 6.4.

[123] Cao Y, Lam C, Wang M, Li H (2006) Nonmechanical protein can have significant mechanical stability. Angew Chem Int Ed Engl 45: 642–645.

[124] Best RB, Li B, Steward A, Daggett V, Clarke J (2001) Can non-mechanical proteins withstand force? stretching barnase by atomic force microscopy and molecular dynamics simulation. Biophys J 81: 2344–2356.

[125] Prince A E Rouse (1953) A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. J Chem Phys 21: 1272.

[126] Niklas KJ, Newman SA (2013) The origins of multicellular organisms. Evol Dev 15: 41–52.

[127] Hynes RO, Zhao Q (2000) The evolution of cell adhesion. J Cell Biol 150: F89–F96.

[128] Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA (2002) The relationship of protein conservation and sequence length. BMC Evol Biol 2: 20.

[129] Meyerguz L, Grasso C, Kleinberg J, Elber R (2004) Computational analysis of sequence selection mechanisms. Structure 12: 547–557.

[130] Lin MM, Zewail AH (2012) Hydrophobic forces and the length limit of foldable protein domains. Proc Natl Acad Sci U S A 109: 9851–9856.

[131] Dietz H, Rief M (2004) Exploring the energy landscape of gfp by single-molecule mechanical experiments. Proc Natl Acad Sci U S A 101: 16192–16197.

[132] Rief M, Gautel M, Schemmel A, Gaub HE (1998) The mechanical stability of immunoglobulin and fibronectin iii domains in the muscle protein titin measured by atomic force microscopy. Biophys J 75: 3008–3014.

[133] Wimley WC (2003) The versatile beta-barrel membrane protein. Curr Opin Struct Biol 13: 404–411.

[134] Inoue T, Hagiyama M, Enoki E, Sakurai MA, Tan A, et al. (2013) Cell adhesion molecule 1 is a new osteoblastic cell adhesion molecule and a diagnostic marker for osteosarcoma. Life Sci 92: 91–99.

[135] Wong CW, Dye DE, Coombe DR (2012) The role of immunoglobulin superfamily cell adhesion molecules in cancer metastasis. Int J Cell Biol 2012: 340296.

[136] Labasque M, Devaux JJ, Lévêque C, Faivre-Sarrailh C (2011) Fibronectin type iii-like domains of neurofascin-186 protein mediate gliomedin binding and its clustering at the developing nodes of ranvier. J Biol Chem 286: 42426–42434.

[137] Fogel AI, Stagi M, de Arce KP, Biederer T (2011) Lateral assembly of the immunoglobulin protein syncam 1 controls its adhesive function and instructs synapse formation. EMBO J 30: 4728–4738.

[138] Golias C, Batistatou A, Bablekos G, Charalabopoulos A, Peschos D, et al. (2011) Physiology and pathophysiology of selectins, integrins, and igsf cell adhesion molecules focusing on inflammation. a paradigm model on infectious endocarditis. Cell Commun Adhes 18: 19–32.

[139] Babu K, Hu Z, Chien SC, Garriga G, Kaplan JM (2011) The immunoglobulin super family protein rig-3 prevents synaptic potentiation and regulates wnt signaling. Neuron 71: 103–116.

[140] Albani AE, Bengtson S, Canfield DE, Bekker A, Macchiarelli R, et al. (2010) Large colonial organisms with coordinated growth in oxygenated environments 2.1 gyr ago. Nature 466: 100–104.

[141] Bengtson S, Belivanova V, Rasmussen B, Whitehouse M (2009) The controversial "cambrian" fossils of the vindhyan are real but more than a billion years older. Proc Natl Acad Sci U S A 106: 7729–7734.

[142] Yang Y, Zhan J, Zhao H, Zhou Y (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins: Structure, Function, and Bioinformatics 80: 2080–2088.

[143] Doi M, Edwards SF (1988) The Theory of Polymer Dynamics. The International Series of Monographs on Physics.

[144] Bruning M, Barsukov I, Franke B, Barbieri S, Volk M, et al. (2012) The intracellular ig fold: a robust protein scaffold for the engineering of molecular recognition. Protein Eng Des Sel 25: 205–212.

[145] Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, et al. (2003) Pulling geometry defines the mechanical resistance of a beta-sheet protein. Nat Struct Biol 10: 731–737.

[146] Kappel C, Zachariae U, Dölker N, Grubmüller H (2010) An unusual hydrophobic core confers extreme flexibility to heat repeat proteins. Biophys J 99: 1596–1603.

[147] Bu T, Wang HCE, Li H (2012) Single molecule force spectroscopy reveals critical roles of hydrophobic core packing in determining the mechanical stability of protein gb1. Langmuir 28: 12319–12325.

[148] Sadler DP, Petrik E, Taniguchi Y, Pullen JR, Kawakami M, et al. (2009) Identification of a mechanical rheostat in the hydrophobic core of protein l. J Mol Biol 393: 237–248.

[149] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–29.

[150] Federhen S (2012) The ncbi taxonomy database. Nucleic Acids Res 40: D136–D143.

[151] Kimball R, Ross M (2002) The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. wiley.