

ISAS - International School for Advanced Studies



Coarse-grained models for Protein Folding and Function

Thesis submitted for the degree of Doctor Philosophiæ

Candidate
Luca Marsella

Supervisors
Amos Maritan
Antonio Trovato
Cristian Micheletti

October 2005

Contents

Introduction	5
1 Proteins and Disorder	9
1.1 Protein structure	11
1.2 Solving protein structures	17
1.3 Folding and design	18
1.4 Intrinsically unstructured proteins	22
1.5 Folding upon binding	24
1.6 Properties of the disordered state	26
2 Geometric approach to protein folding	31
2.1 Tube model of a thick polymer	32
2.2 Refined tube model for proteins	35
2.3 Presculpted energy landscape	37
2.4 Summary	42
3 Binding and Folding	43
3.1 Methods	44
3.2 Tuning model's parameters	48
3.3 Substrates for binding and folding	52
3.4 Unbiased interactions with the substrate	55
3.5 Enhancing specificity of effective contacts	60
3.6 One-to-one interactions	63
4 Gaussian models for protein function	71
4.1 Structural characterization	72
4.2 Theory	74
4.3 Tuning model parameters	77
4.4 Temperature factors and heme modeling	78
4.5 Results and Discussion	81
4.6 Elasticity and time scale	87

Conclusions and perspectives	91
Bibliography	103
Acknowledgements	105

Introduction

In recent years the development of experimental techniques suitable for the investigation of the behaviour of biomolecules *in vitro*, techniques such as Nuclear Magnetic Resonance, Circular Dichroism, Infrared Spectroscopy and Small Angle Scattering among the others, helped to understand the properties of biomolecules of vital importance for the higher evolved organisms.

Processes such the activation of transcription of the genetic code, and other regulatory mechanisms performed within the cell, need very specific and accurate description of the three-dimensional structures of the involved protein complexes, in order to assess the basic physicochemical mechanisms implicated in them.

Although important advances have been made in the field of structure prediction from the sequence, mainly using the tools of bioinformatics, still we miss a *first principle* theoretical understanding of the key mechanisms that form the basis for the folding process of even the simplest proteins, like the small single-domain globular ones, because of the sheer complexity of the problem – the huge number of degrees of freedom associated with the protein atoms and the surrounding water molecules, as well as the history dependence implicit in an evolutionary process.

On the other hand a great variety of effective potentials for folding exist, which exploit a huge number of empirical parameters, especially when they prove to be successful: it may be difficult to extract the common physical principles underlying the different approaches. This problem can be more easily addressed by means of simple coarse-grained models, but the question is then whether the modeling is realistic enough in order to tackle the issue at hand. Throughout this work we will pursue a qualitative description of several properties of small globular proteins. Our aim is to show that a unified framework, based on a coarse-grained modeling of the common physicochemical properties of globular proteins can be used to understand the mechanism of binding and folding adopted throughout the evolution by the intrinsically unstructured proteins (IUPs).

IUPs are a family of proteins recently discovered, deeply related to the

regulatory mechanisms of the living cells [1, 2, 3, 4, 5, 6, 7]: they are biological molecules which under physiological conditions do not exhibit extensive structural order in solution, but often display local and limited residual structure [1].

A key feature of IUPs is their high conformational flexibility, that allows them to interact with different molecular partners and adopt relatively rigid conformations in the presence of natural ligands, thus undergoing a loss of conformational entropy upon binding. Furthermore, the conformation in the bound state is determined not so much by the amino acid sequence but by the structure of the interacting partner.

On the other hand, it is well known that the sequence of amino acids comprising a protein encodes its native state structure. It has been recently shown [8, 9] that considerations of symmetry and geometry determine the limited menu of folded conformations that a protein can choose from for its native state structure. Such studies provide compelling support for the idea that protein native state structures reside in a physical state of matter in which the free energy landscape is sculpted by considerations of geometry and symmetry.

According to this framework, protein structures belong to a novel phase of matter associated with the marginally compact phase of short tubes with a thickness specially self-tuned to be comparable to the range of attractive interactions promoting the compaction.

This phase is a finite size effect and exists only for relatively short tubes: the proximity to a phase transition provides a simple explanation for the flexibility of native state structures. The marginally compact phase is stabilized by the interplay of the hydrophobic effect and hydrogen bond formation [8, 9]. The structures that one finds in it are modular in construction being made up of two principal building blocks, helices and planar sheets: the degeneracy is greatly reduced so that the number of the resulting energy minima is relatively small.

We will show that the interactions between the IUP and its partner play a role analogous to a design process, in which the geometric properties of the pattern of interactions result in the IUP adopting one of the best-fit structures from the menu predetermined at the homopolymeric level by geometrical considerations [8, 9]. In order to emphasize this point, we did not introduce any sequence heterogeneity in our modeling of IUPs except for the interaction with the target geometry, which may be tuned in order to fit a target fold.

It is important to note that for a IUP, the sequence heterogeneity, and more specifically the presence of polar residues, leads to no intrinsic ordering. The case considered here is much simpler – we have just one kind of amino

acid, the ground state in the absence of any target geometry is ordered, namely a single helix. This results in a slightly more complex free energy landscape with several other competitive local minima. The target fold we wish to observe upon binding to a specific partner is chosen from among those minima.

Monte Carlo simulations have been performed within a cubic box of side 50.0 Å on a homopolymeric chain of 24 amino acids. The partner of the IUP is represented in a simple coarse-grained framework through three carefully chosen contact points lying on the inner part of the box bottom face which each has a special affinity to a specific amino acid of the IUP, mimicking a molecular recognition mechanism in the crudest way.

Also, the bottom wall serves to capture the steric hindrance of the ligand/substrate partner and the associated loss of conformational entropy induced upon the IUP.

We present the results of simulations run for different instances of binding patterns, chosen to get the IUP folded onto a three stranded β -sheet, a zinc-finger-like conformation, a two-helix bundle and a $\beta-\alpha-\beta$ kind of secondary structure respectively, through different levels of contact bias between the polymer and the substrate.

After this brief outline on the topics of disordered proteins, we are left with the awareness that the relationship of the three-dimensional structure of a protein to the function needs then a reassessment [10]. The behaviour of disordered proteins proves that there is no need, for a biomolecule, to have a permanently folded structure to perform a specific task. On the contrary, the structure adopted may vary from case to case, depending on the molecular binding partner.

Nonetheless, it is evident that IUPs get structured before they accomplish their job: knowing the structure will still be a key point for the understanding of the related function, once we know the particular partners involved. Keeping in mind this revised version of the “structure-function paradigm”, we will present a simple coarse-grained model that tries to infer near equilibrium functional motions of proteins, from the knowledge of their native state.

Effective beta carbon atoms are taken into account besides C^α s for all residues but glycines in the coarse-graining procedure, without leading to an increase in the degrees of freedom (β Gaussian Model). Normalized covariance matrix and deformation along slowest modes with collective character are analyzed, pointing out anti-correlations between functionally relevant sites for the proteins under study.

In particular we underline the functional motions of an extended tunnel-cavity system running inside the protein matrix, which provide a pathway

for small ligands binding with the iron in the heme group.

We give a rough estimate of the order of magnitude of the relaxation times of the slowest two overdamped modes and compare results with previous studies on proteins.

The plan of this thesis is as follows: in the next chapter (chapter 1) we review the main properties of globular proteins, in particular focusing on the state of the art of protein folding and design. Then we describe in detail the simple model for folding adopted throughout the present work (chapter 2). Chapter 3 shows the further modeling introduced to handle the specific subject of disordered proteins, with full explanation of all parameters used and with some possible interpretation of the results obtained.

In the last chapter of this work (chapter 4) we present a study on the near equilibrium dynamics of two small proteins in the family of truncated hemoglobins, developed under the framework of a Gaussian network approach.

Chapter 1

Proteins and Disorder

Proteins are heteropolymer chain molecules, built by the assembly of the twenty amino acids occurring in nature through the chemically stable peptide bond: the number of amino acids that build a protein can vary from few tens to several hundreds. Proteins belong to the group of biopolymers, which includes also nucleic acids (DNA, RNA) and polysaccharides: while the latter have evolved in order to perform a particular task – mainly information and energy storage respectively – proteins can potentially cover an unlimited number of different functions in the living world.

In fact they control and affect most biological functions in living organisms: apart from catalyzing almost all biochemical reactions, they are responsible for the transport and store of a variety of elements ranging from macromolecules to electrons, for the transmission of information between specific cells and organs, for the passage of molecules across cell membranes, for the regulation of the activity of the immune system in complex organism and for the genetic expression.

Protein activity depends on the complex relationship between the sequence of amino acids forming the polypeptide chain and the associated three-dimensional structure, that is stable against slight variations of environmental conditions.

When the protein is synthesized it is not yet biologically active, since it has to fold itself into a unique and specific three-dimensional structure. The stability of the polymeric chain in this functional conformation, known as the native state of the protein, is guaranteed by the interplay of several concurring factors: hydrogen bonds and disulphide bridges, hydrophobic and steric effects, electrostatics.

Under normal physiological conditions (i.e. aqueous solvent, near neutral pH, temperature between 290 and 310 Kelvin) a small protein folds spontaneously in its native state, as it was showed by Anfinsen [11] in the late '50s.

In addition Anfinsen's experiment proved that the sequence entirely determined the stable spatial structure of the protein, which do not rely on any special biological machinery, except for large proteins and their assemblies, since they need the help of special proteins, called chaperonins, in order to fold. This holds for globular proteins, while disordered ones behave quite differently, as we will see in more detail the following sections.

The pioneering work of Anfinsen naturally led to the formulation of the following questions: how is a specific three-dimensional structure encoded in the amino acidic sequence and which are the sequences compatible with a given native state.

These two simple but extremely important issues are known respectively as the protein folding problem and the inverse protein folding problem, also known as protein design: they are just different formulations of the same issue, that is the relationship between amino acid sequences and native states. Although several advances have been made in these fields in the last four decades of study, the physical principles underlying the folding process are currently matter of discussion.

The understanding of such principles would be of great importance for medicine, since it would greatly enhance the design of novel proteins with the desired biological function, the design of new and more effective drugs, the prediction of the function of a protein from the knowledge of the bare sequence, which is by far easier to get rather than solving the whole three-dimensional structure. The database of sequences [12] is already huge, hundred times larger than the database of solved three-dimensional structures [13]: the number of known sequences approximately doubles every year, since they are obtained by biochemical methods - either directly by the protein itself or indirectly by the corresponding gene on DNA.

On the other hand determining the structure requires a long process: most of them are in fact obtained through x-ray crystallography, while the smaller ones via nuclear magnetic resonance (NMR). In general the latter gives structures at a lower resolution than the first one, which can provide native structures at a resolution lower than 2.0 Å.

In order to study a structure with x-ray the crystallization of the protein is needed, which is usually difficult to achieve. Moreover, proteins packed in a crystal may show a slightly different structure than the one that might be observed in solution, due to crystal packing.

In principle one could explore proteins behaviour by numerically integrating the equation of motion for each degree of freedom, using interaction energies obtained experimentally and including all the details of the interaction between the protein and the surrounding solvent: big clusters can currently simulate several tens of nanoseconds of real time classical dynam-

ics with time steps appropriate to describe harmonic motions of covalently bonded atoms.

This time scale is still too small in comparison to the typical time required for a protein to fold, which starts from the range of milliseconds up to several seconds for bigger proteins. Furthermore, the use of force fields by classical molecular dynamics is an approximation to the real quantum behaviour of biomolecules.

Another difficulty, which affects mostly coarse-grained models of biomolecular processes, is related to the existence of a huge number of local energy minima, even in the neighborhood of the native state. The latter is believed to be the absolute minimum of the free energy landscape, still the great number of locally stable states prevents the simulation to sample rapidly the conformational space.

These limitations encountered when facing the protein folding problem, usually lead to replace atomistic models with coarse-grained ones, where amino acids are represented in simplified ways, averaging over suitably chosen degrees of freedom.

In the remainder of the present chapter we wish to present an overview of the experimental results and the theoretical interpretation which lie at the basis of the work presented within this thesis.

1.1 Protein structure

Proteins are macromolecules composed of up to several thousands atoms without apparent symmetries or regularities. Describing such large objects at the atomic level is a quite discouraging task: since 1958, when the first protein structure has been determined by x-ray crystallography, a number of recurrent structures and motifs have been discovered. In some cases the description of protein properties by these motifs is helpful and simplifies concepts. Nonetheless, depending on the type of study to be performed, a resolution at the atomic level may be necessary.

At the lowest level of this hierarchy, there are the 20 amino acids, whose covalent structure is the base for the structure of proteins. Amino acids are bound together to form a linear chain, through the peptide bond, which constitutes the backbone of the structure. Though the polymeric chain is flexible and can adopt, in principle, many different conformations, the interactions among the different regions of the chain are such that only one conformation – called native – will be adopted by the protein under physiological conditions (temperature, pressure, pH). The order of amino acids placed along the chain is of fundamental importance, since changing it may dramatically

change the interactions, destabilizing the native conformation.

The sequence – the order according to which the amino acids are placed along the protein backbone – is the first level of complexity. It can be in fact represented by a one-dimensional string, where each letter is associated to one of the twenty types of amino acids (see table 1.1).

The primary structure apparently does not contain much information, but one has to associate the structure of every amino acid to each letter in the sequence. By doing so a polymeric chain is obtained, which can assume in principle many different conformations, compatible with steric constraints. One needs to know amino acids structures and how they bind together to form the peptide chain in order to understand which conformations are allowed and which are not.

The α carbon atom is bonded to the aminic and carboxylic groups (NH_2 and COOH , respectively), the chemical group R – usually called side chain – and a hydrogen atom (fig. 1.1).

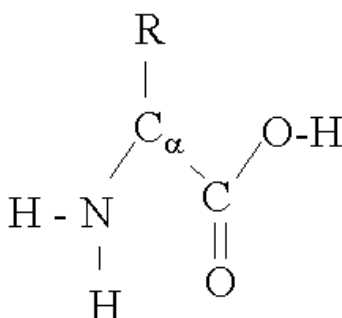


Figure 1.1: Schematic chemical structure of an amino-acid.

Apart from proline, whose carbon atom in the carboxylic group is bonded to the side chain itself, the other amino acids differ only in the nature of the side chain's group. The number of atoms forming the side chain can vary from one – in glycine is just one hydrogen atom – to a maximum of eighteen for tryptophan and arginine.

Side chains are formed by different combination of carbon and hydrogen atoms, as well as oxygen and nitrogen for some amino acids. In cysteine and methionine a sulfur atom is present: it is responsible for the stabilization of three-dimensional structures through a disulfide bridge.

While the residue part of an amino acid characterizes the chemical properties of the molecule, the aminic and the carboxylic groups have an important role to connect amino acids in a polymeric chain.

Residue			Frequency
alanine	ALA	A	8.3
arginine	ARG	R	5.7
asparagine	ASN	N	4.4
aspartic acid	ASP	D	5.3
cysteine	CYS	C	1.7
glutamine	GLN	Q	4.0
glutamic acid	GLU	E	6.2
glycine	GLY	G	7.2
histidine	HIS	H	2.2
isoleucine	ILE	I	5.2
leucine	LEU	L	9.0
lysine	LYS	K	5.7
methionine	MET	M	2.4
phenylalanine	PHE	F	3.9
proline	PRO	P	5.1
serine	SER	S	6.9
threonine	THR	T	5.8
tryptophan	TRP	W	1.3
tyrosine	TYR	Y	3.2
valine	VAL	V	6.6

Table 1.1: List of the twenty amino acids with their frequency in proteins, taken from [14]: amino acids can be identified by a three-letter or a one-letter code, shown in the second and third column respectively.

When two amino acids are hydrolyzed, the aminic group and the carboxylic group of different amino acids form a covalent bond, shown in fig. 1.2. After an amino acid has lost a water molecule it is called residue.

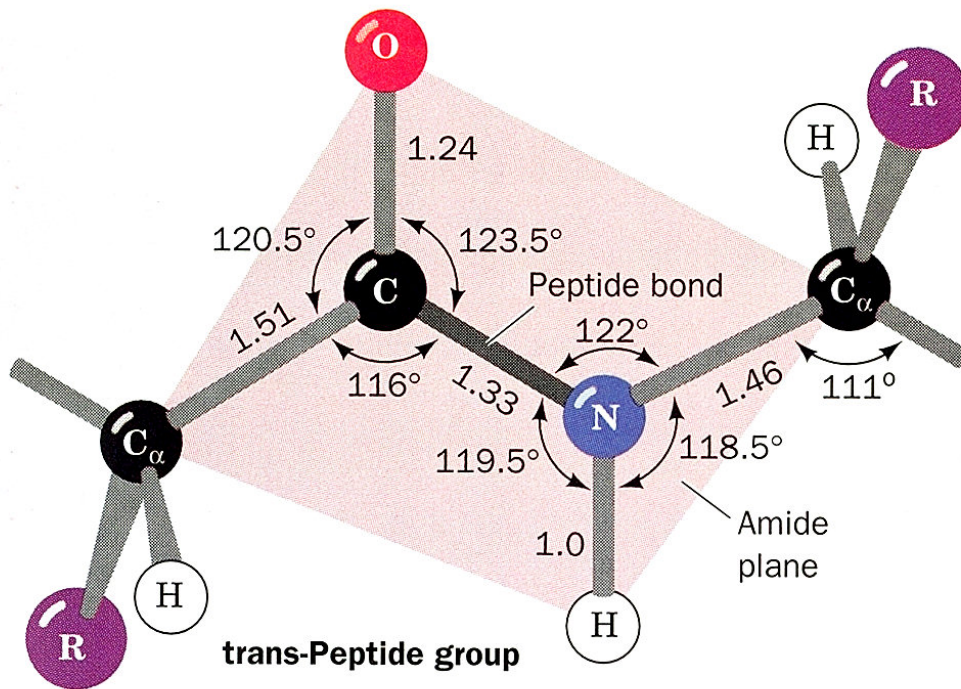


Figure 1.2: Schematic chemical structure of a peptide bond between two residues.

The bond between the carbon and the nitrogen is called peptide bond and, since it is a partial double-bond, rotations along this axis are forbidden (except rotations of 180°).

On the other hand, rotations are allowed along the single bonds between C^α and N and between the two carbon atoms, as far as steric clashes do not occur: rotations along these axes are represented by two torsional angles called ϕ and ψ , respectively (fig. 1.3). Since bonds between nearest neighbouring atoms are not aligned, these rotations cause a conformational change in the polypeptide chain.

The dihedral angles ϕ and ψ can assume all the values within the range $[-\pi, \pi]$: some values are in fact more likely than others. In particular, some values are never allowed due to steric reasons, since they would correspond to an overlap of atoms of the side-chain with atoms of the backbone. The

permitted values of ϕ and ψ were first determined by Ramachandran and collaborators [15], using hard-sphere models of the atoms and fixed geometries of the bonds. The permitted values of ϕ and ψ are usually indicated on a two-dimensional map of the $[\phi\psi]$ plane, known as a Ramachandran plot 1.4(a). Since the size of the residue strongly depends on the amino acid type, Ramachandran plots are amino acid specific. In particular, glycine, which has the smallest residue, has a Ramachandran plot with several allowed regions, indicating a flexibility uncommon to other amino acids.

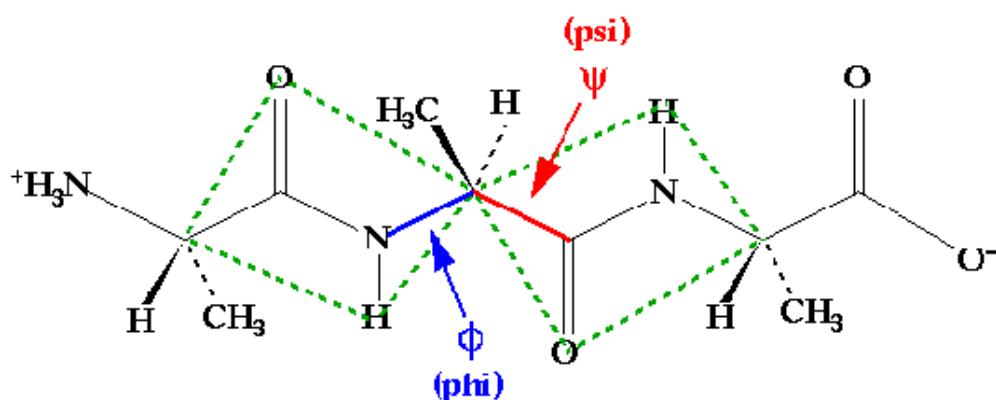


Figure 1.3: Protein flexibility is due to the presence of single bonds along the main chain between the nitrogen and α carbon atoms (ϕ) and between the two carbon atoms (ψ).

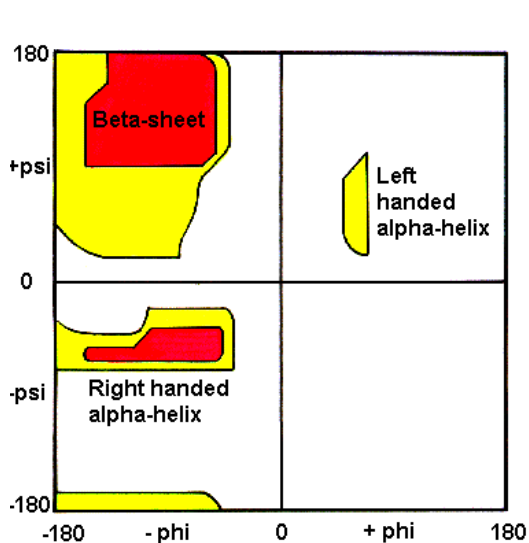
The allowed regions of the Ramachandran plots are not equally likely in real proteins. A statistical analysis of protein structures shows that some regions of the $[\phi\psi]$ plane are more populated than others. The most populated region corresponds to angles around $(-60^\circ, 50^\circ)$. Several consecutive amino acids with such values of the dihedral angles take part to a helical structure that is called α -helix and it is a motif quite recurrent in proteins. Each turn in the helix is formed on average by 3.6 amino acids, the i -th amino acid being in spatial contact with the $(i+3)$ -th and with $(i+4)$ -th one.

Another recurring secondary structure is the extended conformation, or β -strand, that is associated to the angles $(-130^\circ, 120^\circ)$. Extended conformations are frequently found associated together to form β -sheets. It is possible to distinguish between parallel and anti-parallel sheets. In the first case if the i -th and j -th amino acids are in contact then $(i+1)$ -th and $(j+1)$ -th will be still in contact. In the second case it will be true for the $(i+1)$ -th and $(j-1)$ -th ones.

Secondary structures are assembled together to form more complex structures by turns and loops. In the first case, an amino acid, usually glycine since it is the smallest residue, makes a tight turn which changes completely the direction of the backbone. In loops the change of direction is more gradual, being distributed over several amino acids. Finally, random coils are protein regions without a definite shape.

Such an assembly of secondary motifs is called tertiary structure. This is the functional three-dimensional configuration of the protein, whose stability and compactness is due to different types of interactions between amino acids far apart on the chain, briefly listed at the beginning of this chapter.

Secondary structures contain information on the conformation of the protein and can be represented by a one-dimensional string, using the following convention: H for a α -helix, S for a bend or loop, T for a turn and R for a random coil [16, 17]. Secondary structures are generally highlighted by the cartoon scheme of visualization (fig. 1.4(b)).



(a) Ramachandran plot of a tripeptide, showing sterically forbidden areas for all amino acids except glycine (white), and allowed ones (colour). The use of van der Waals radii smaller by 0.1 Å allows more conformations and lets a new distinct area available.



(b) Cartoon representation of an ATP binding domain (pdb id: 1B0U) with secondary structures highlighted: helices, strands (arrows) and random coils and loops (thin tubes).

1.2 Solving protein structures

One of the most important challenges in understanding biological reactions occurring in higher organisms is the determination of the structure of the molecules participating to the reaction. This is especially true for proteins, whose structure has been selected by evolution for a specific biological task. In some cases, it is just the geometrical shape that contains important information on the function, especially when a cavity in the structure is complementary to the geometrical shape of another macromolecule ligand (docking). In other cases the geometrical shape can give only generic indications where the binding site is located, and only detailed electrostatic calculations can solve the docking problem.

While the structure is important to understand the function of the protein, its experimental determination is difficult and expensive. By contrast, it is very easy to determine the sequence of amino acids by experimental measurements (sequencing) or by translating the associated gene. The number of sequences that have been determined up to now is almost hundred times larger than the number of structures, and the number of sequences that will be acquired per day grows rapidly. It follows that one of the most important research field in bioinformatics and biophysics is the prediction of the structure of already known sequences. In principle this problem can be solved by following the dynamics of the protein embedded in the solvent (which has a fundamental role in driving the folding of the protein) on a computer and finding the lowest free energy conformation. However, the complexity of the atomic structure of a protein and the time scale on which the folding occurs, make this approach unfeasible.

A possible way out to overcome this kind of problems might be to use simplified descriptions of proteins in which amino acids interactions and steric constraints are described in an effective way. These models have received a lot of interest in the community of physicists. However, because of their lack of atomic details and their approximate description of interactions, simplified models are still far to be successfully applied in structure prediction.

Among the methods more reliable for structure prediction, there are those based on homology modeling. Homology modeling deals with the problem to detect an homology, i.e. an evolutionary relationship, between the protein under study and proteins of known structure. Usually homology is detected by aligning and comparing the sequence of unknown structure (or target sequence) with the sequences of proteins of known structure.

Such structures can be used as templates to make a model, to be carefully refined. A similar procedure has many advantages: first, it can be automated, thus allowing many researchers to access model structures for

proteins, whose structure has not yet experimentally determined. Then an experimental determination could be not necessary, especially when a high homology has been detected. Finally, it can be used on a large number of known protein sequences: it has been estimated that it is currently possible to model with useful accuracy significant part of approximately one third of all known protein sequences. Furthermore, the number of proteins of known structures is destined to increase.

The basic idea of homology modeling is that similar sequences are likely to have similar structures: similarity above 25% can be enough to produce a good model of the unknown structure. Usually, structures are much more conserved than sequences by the evolution: this implies that two related sequences can share similar structures [18]. However, deciding on the base of sequence analysis if two structures are similar, is still a difficult task.

For these reasons and for the importance that the structure has for molecular biologists, structure data have been collected in a unique big database, called Protein Data Bank (PDB [13]). Since 1975, when PDB has been built up, a lot of structures of macromolecules – mainly proteins – have been collected: at the moment there are more than thirty thousands structures and structure models: most of the protein structure data were obtained by X-ray crystallography (27693) and by NMR (4741), while only few are obtained by theoretical modeling. Structures deposited on PDB constitute an important source for molecular biologists and for people working in the expanding fields of bioinformatics and biophysics.

1.3 Folding and design

In the higher organisms proteins are synthesized in the cytoplasm through a complex mechanism of biosynthesis. Once the sequence is synthesized the protein is not yet active. To become biological active it has to fold into a specific three-dimensional conformation, i.e. the native state. In principle, there are a lot of different conformations that the sequence can adopt.

Assuming three different coarse-grained conformations per amino acid, the number of possible distinct conformations, for a protein with 100 amino acid, which is a relatively small one, should be $3^{100} \approx 10^{48}$.

Some of these conformations are not accessible, due to steric reasons. Nevertheless, even taking into account this observation, the number of physical conformations is enormous and the protein should need a folding time larger than the age of the universe, in order to find the native state by a random exploration. This is known as the Levinthal paradox, from the name of the first one that arose it [19].

How can protein find their native conformation among a huge number of conformations? A first attempt to answer this question came from Anfinsen and coworkers [11]. Before their studies, the nature of the sequence-structure relationship was completely unknown and it was still an open question if the structure was written in the sequence as a physicochemical message or if there was a biological machinery similar to enzymes regulating biosynthesis.

Anfinsen's studies on the re-folding of ribonuclease showed clearly that protein sequences under physiological conditions can automatically find their native state by minimizing the free energy. In other words, proteins with their solvent constitute a physical system that is thermodynamically stable only in their native conformation.

This hypothesis excludes the possibility that proteins adopt their native conformation due to a complex biological machinery, since the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, ...) is the one in which the Gibbs free energy of the whole system is the lowest: that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment [11].

Anfinsen's discovery had an enormous impact on molecular biology: if the native state of a protein is the global minimum of the free energy, then it must be possible to predict its structure just by simulating its dynamics using the standard laws of physics.

This is still one of the fundamental unsolved problems in biophysics. The complexity of the problem is mainly due to the size of proteins, since large proteins are made up by several hundreds of amino acids, i.e. by thousands of atoms, and to the difficulty to treat the solvent accurately.

The integration of Schrödinger's equation is possible numerically, by using supercomputers, only for time intervals of few picoseconds. Within the framework of classical molecular dynamics instead, one can follow the dynamics of biomolecules, guided by approximated force fields, for much longer times.

A striking result in this field has been obtained in 1998 by Duan and Kollman, who were able to follow the folding of a small protein for 1 μ s [20]. The time necessary for the simulation was 4 months on a 256 processors supercomputer corresponding to a total CPU time of 80 years. At the end of the simulation they observed the presence of an intermediate state in the folding pathways.

The result of Duan and Kollman shows that we are still far from solving the protein folding problem by brute force, i.e. simulating the real dynamics of proteins until the global free energy minimum is found. Most of the pro-

teins are much longer than the one used in their simulation and they would fold in a CPU time that is 10^9 to 10^{12} times longer. Furthermore following biophysical processes by simulating them in full details does not necessarily mean to understand them.

In order to understand protein folding more deeply, a lot of simplified models have been proposed. Usually, simplified models have not, at the moment, the aim to find the native state of a given protein sequence: they focus rather on the dynamics and thermodynamics of such physical systems. This is the case, for example, of the Gō model [21], in which the knowledge of the native structure is the input of the model itself.

The main idea of this model is that an energetic bias towards the native state, without any realistic description of physical interactions, allows the study of protein dynamics. In the simplest form, the energy of a conformation is defined as the number of contacts in common with the native state, taken with the negative sign: two amino acids are said to be in contact if the distance of their C $^{\alpha}$ -s is less than a given cutoff, usually taken between 6 and 8 Å, and if they are not consecutive along the peptide chain.

Since the energy is well specified and the native state is by definition the state with the lowest free energy, the only problem is how to define the dynamics. The basic feature of this model is specificity, since the native conformation is by definition the ground state and an ergodic dynamics will reach always the native state. However, such model is interesting because it allows an analysis of the folding process of a well designed sequence and of the way dynamics is controlled by the topology of the native state.

Models for random heteropolymers can be used to study protein-like features. Usually, models amenable to analytical calculations are too simple to capture the sequence-structure relationship, typical of proteins.

On the other hand, models apt for numerical implementation, like the one described in next chapter, capture some of general features of proteins and allow a deeper study of protein folding and design. Proteins, in fact, can be considered special heteropolymers that have evolved for fast-folding into a unique and thermodynamically stable conformation [22]. At variance with the Gō model, such models are completely unbiased, since fast folding and other protein-like features are expected to emerge through a suitable sequence selection.

In fact, heteropolymers at low temperature behave differently from proteins. They show a glassy dynamics and the state in which they fold depend on the initial conditions [22]. Kinetic and energetic barriers prevent an easy access to the ground state and the search of the global minimum is more similar to that prospected by Levinthal [19]. This is not the case for sequences that show protein-like features: they fold through a two-state mechanisms

rapidly and reversibly in the native state [23].

It follows that a rigorous study of protein folding has to be preceded by a suitable optimization of the sequence: for such a sequence physical properties, for example the type of transition to the native state, can be compared with real proteins.

Protein design deals with the problem to find how many and which kind of sequences fold on a given target structure. In principle the problem can be solved enumerating all the possible sequences and attempting to solve the protein folding problem for each of them. Obviously, a similar strategy is not feasible. Furthermore, as we have seen previously, the structure prediction problem is far from a general solution even in an approximated way.

A first tentative solution to protein design was proposed, and checked with experiments, by Hecht and coworkers [24]. Their method was based on the assumption that the hydrophobic force is the main factor driving proteins into their native conformation. The hydrophobic force is the propensity of hydrophobic amino acids to cluster in buried regions, leaving polar ones exposed to a polar solvent – which is water, in living organisms. By specifying explicitly the positions of hydrophobic and polar amino acids within the sequence, they tried to design stable conformations without the explicit design of specific interresidue contacts. Through the use of binary patterns they were able to produce compact folds with high α -helical content, comparable to a four helix bundle for a large fraction of the designed sequences [24].

This strategy is different from modern computational approaches to protein design with automated sequence selection [25], where the a major ingredient is the optimization of side-chains packing [26].

In fact hydrophobic-polar patterning of sequences is not a sufficient condition for a successful design, even if it is a good filter to reduce the gigantic number of sequences. The burial of amino acids cannot be the only criterion of selection, since hydrophobic-polar patterning is not even a necessary conditions for protein design.

There are two reasons: first, hydrophobic interactions are not the only ones, since stabilization in the native state is increased by hydrogen bonds, polar effects between amino acids and van der Waals forces. Furthermore, side-chains packing plays an important role in discarding otherwise energetic favorable conformations, emphasizing the need to couple sequence design with backbone flexibility for general protein design, as shown by the group of Baker [26].

They developed a method for de novo designing stable folds, successfully applied to a 93-residue α/β structure, whose topology was not present in the Protein Data Bank [13]. As they noticed, computational design of novel protein structures is a more rigorous test of force fields than the redesign of

naturally occurring proteins [26].

The design procedure they adopted include a search of nearby conformational space, in addition to sequence space: this is accomplished by iterating between sequence optimization and structure prediction.

1.4 Intrinsically unstructured proteins

Disordered proteins – referred to as intrinsically unstructured proteins (IUPs) as well – are biological molecules devoid of extensive structural order, but often displaying signs of local and limited residual structure [1]: the word “natively unfolded” was introduced in 1994 to describe the behaviour of tau protein, which turned out to have the properties of a denatured molecule in solution, without any evidence of compact folding and only a minimal content of secondary structure elements [27].

Due to the resemblance of their structure to denatured states of globular proteins, IUPs have long been considered to exist in a random coil conformational state, although true random coils are not observed, since there is persistence of native-like topology even in denatured proteins [28].

Further analysis on this class of proteins showed that they can be classified in two main groups [3]: the first one comprehends flexible chains, yet compact and exhibiting properties typical of the molten globule, whereas the biopolymers traditionally described as “random coils” are found within the second group.

The latter are characterized by specific amino acid sequence “encoding disorder” [4, 5] with low overall hydrophobicity and high net charge, hydrodynamic properties typical of a random coil in poor solvent [29] and a low level of secondary structure.

The high conformational flexibility is essential for IUPs to accomplish their biological function, since it may allow them to interact with different molecular partners and adopt relatively rigid conformations in the presence of natural ligands [3]: many examples of coupled folding and binding events have been reported recently, providing new insights into mechanisms of molecular recognition [6].

The conformation they adopt is largely defined by their interacting partner rather than their amino acid sequence, differently from the case of globular proteins [7].

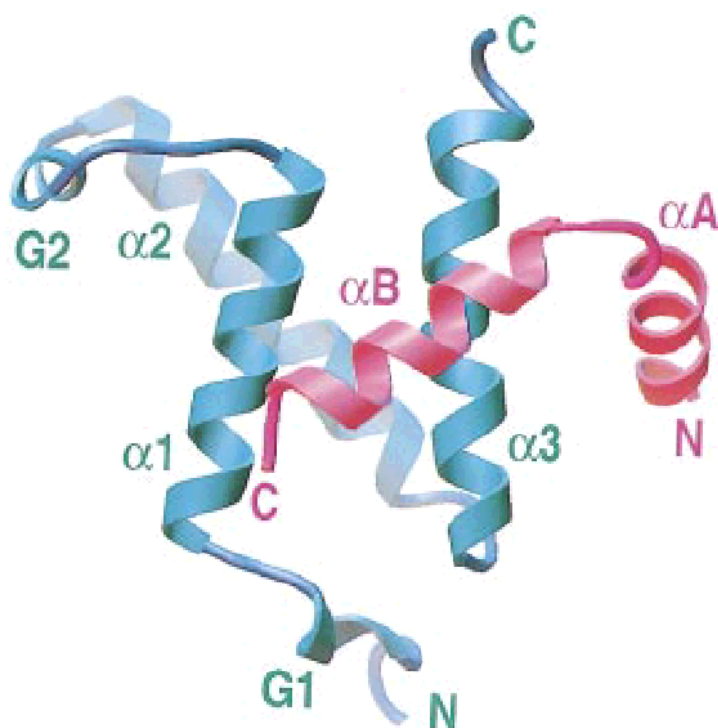


Figure 1.4: NMR structure of the complex formed between the KIX domain (cyan) of the coactivator CBP and the kinase inducible activation domain (pKID) of the transcription factor CREB (pink) [30].

1.5 Folding upon binding

There are now numerous examples of proteins that are unstructured or only partially structured under physiological conditions and yet are nevertheless functional [6]. Such proteins are especially prevalent in eukaryotes: in many cases, intrinsically disordered proteins adopt folded structures upon binding to their biological targets.

It has long been axiomatic that the function of a protein is directly related to its three-dimensional structure: recently however, it has been recognized that numerous proteins lack intrinsic globular structure or contain long disordered segments under physiological conditions and, furthermore, that this is their normal, functional state [10, 4, 5].

Such proteins are frequently involved in regulatory functions in the cell and the structural disorder may be relieved upon binding of the protein to its target molecule. The intrinsic lack of structure can confer functional advantages, including the ability to bind, perhaps in different conformations, to several different targets.

One speaks of folding upon binding or inducible binding, which is different from the so-called constitutive binding [31, 10]: the kinase inducible activation domain of CREB, which binds to the KIX domain of the coactivator CBP in the phosphorylated form (pKID, fig. 1.4), is a typical case of the first phenomenon, schematically shown in fig. 1.5.

The free pKID domain is intrinsically unstructured, but folds on binding to its target. The entropic penalty associated with the folding transition is counterbalanced by the large enthalpy of binding, partly due to the complementary intermolecular hydrogen bonds formed by the phosphoserine group of pKID.

In the unphosphorylated state, binding of KID is very weak since the smaller enthalpy of binding cannot compensate for the entropic cost of the folding transition. Thus, inducible binding is a consequence of the thermodynamic balance that arises from the coupling of folding and binding events (fig. 1.5).

By contrast, the transactivation domain of the c-Myb oncoprotein is folded into a helical structure in its free state and can bind constitutively to its target protein since both the ΔH and ΔS of association are favorable [31].

The sequential incorporation of unfolded monomers is a well-recognized mechanism of increasing the size of macromolecular assemblies: disordered segments appear to be very common in the proteins encoded by the various genomes, especially those of higher eukaryotes.

A recent survey of 31 genomes indicated that disordered segments longer

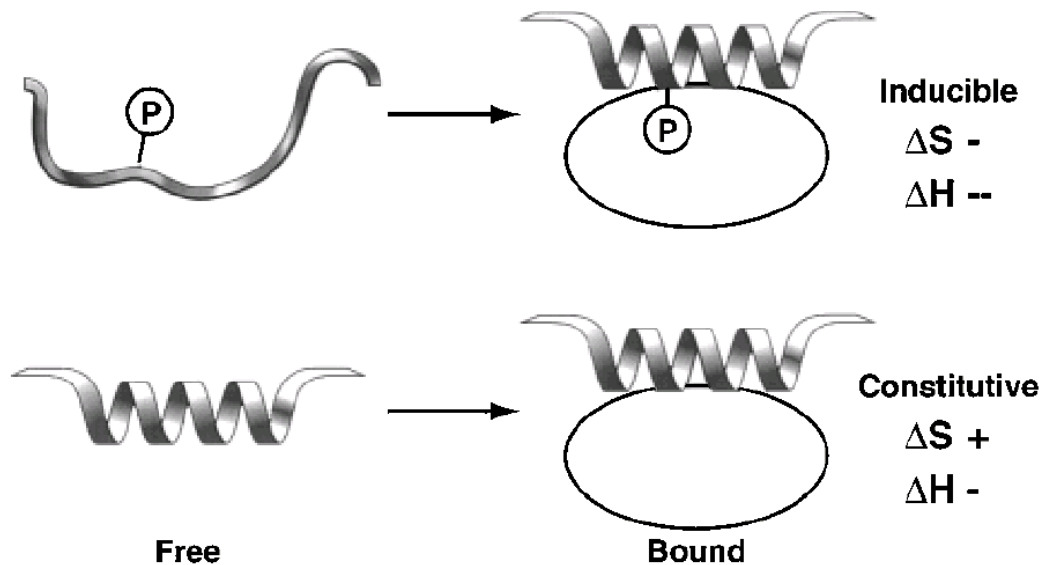


Figure 1.5: Schematic diagram showing how differences in intrinsic structural propensities of a protein domain can determine whether binding is constitutive or inducible [31].

than 50 residues are highly prevalent [4]. In the four eukaryotic genomes surveyed, more than 30% of sequences are predicted to have disordered regions of this length and, in *Drosophila*, a staggering 17% of proteins are predicted to be wholly disordered.

It has long been recognized that local folding of proteins is frequently coupled to DNA binding [6]. In some instances, mutual cooperative folding of both the protein and DNA has been observed. Induced folding transitions have been observed in the binding of zinc finger proteins to DNA and appear to function to increase binding affinity or specificity. Thus, binding of the zinc finger domain of the nuclear receptor RXR (retinoid X receptor) to DNA leads to induced folding of the dimerization region, which is dynamically disordered in the free protein [6].

The importance of protein folding and unfolded proteins in the cell is illustrated by the large variety of chaperone molecules that aid in the productive folding of proteins, most likely by binding to unfolded or incompletely folded states to prevent aggregation, or by unfolding misfolded forms. Part of the function of chaperones is to bind unfolded proteins or molten globules: in some cases, it has also been found that parts of the chaperones themselves must be unfolded in order for the chaperone to function correctly [6].

The folding process for any protein can be thought of as a binding reaction, as it involves the binding of distant parts of the polypeptide chain as tertiary structure is formed; however, in some cases, there is a requirement for external factors as well, in order for folding to be successful. An example of functionally relevant changes in the folding state of a protein concerns the photoactive yellow protein (PYP): activation of PYP converts it from a folded protein to a molten globule structure that is functional in signaling [6].

Various rationales for the employment of unstructured proteins in eukaryotic cells have been put forward in the past years: IUPs offer important advantages in cellular signaling and regulation, with their inherent flexibility which allows their local and global structure to be modified in response to different molecular targets, or interacting with multiple cellular partners and allowing fine control over binding affinity.

It may also be that the relative instability of intrinsically unstructured proteins could impose an additional level of control in cellular signaling and transcriptional processes, in which a response must be rapidly turned on and just as rapidly turned off [10]. Unstructured protein domains may be less sensitive to environmental perturbations and, therefore, may impart stability to complex regulatory networks that might otherwise be overly sensitive to temperature or other changes in cellular conditions [6].

One of the most compelling rationales for the participation of unstructured proteins in binding interactions in particular was provided by Shoemaker et al [32]: by analogy with the folding funnel mechanism of protein folding, the authors envisage that an unstructured protein would have a greater capture radius than a compact, folded protein with restricted conformational flexibility.

They propose a fly-casting mechanism, whereby the unfolded polypeptide binds weakly at relatively long distances and then folds as it reels in its target. The fly-casting mechanism predicts an increased rate of binding with respect to a fully folded protein, which may well be important when the cellular concentrations of a regulatory protein and its target are low, as is the case for many signaling and transcriptional processes.

1.6 Properties of the disordered state

Disordered states of proteins can be either collapsed (molten-globule-like) or extended: regions of proteins that are intrinsically unstructured under physiological conditions differ in amino acid composition from typical globular proteins, being characterized by amino acid compositional bias, low sequence complexity and high predicted flexibility [5].

Indeed, such proteins appear to occupy a unique region of charge-hydrophobicity space [29, 3]: the first analysis of Uversky et al. [33] considered a list of 91 natively unfolded proteins, with largest number (32) among them between 50 and 100 residues long and net charge at pH 7.0 as high as +59, or as low as -117, or close to zero.

A comparison was made against the same properties of a set of 275 protein sequences from the Swiss-Prot sequence database [12], selected among small globular monomeric proteins of 50 to 200 amino acid residues, with no disulfide bonds and with no known interaction either with natural ligands or with membranes.

Data shown in fig. 1.6, which includes an enlarged set of disordered proteins [33], are consistent with the conclusion that the combination of low mean hydrophobicity and relatively high net charge represent an important prerequisite for the absence of regular structure in proteins under physiologic conditions, thus leading to natively unfolded proteins: the solid line separating the two groups of proteins represents the border between intrinsically unstructured and native proteins, which allows a rough estimation of the boundary mean hydrophobicity value, $\langle H \rangle_b$, below which a polypeptide chain with a given mean net charge $\langle R \rangle$ will be most probably unfolded:

$$\langle H \rangle_b = \frac{\langle R \rangle + 1.151}{2.785} \quad (1.1)$$

The validity of these predictions has been successfully shown for several proteins [29]: this means that degree of compaction of a given polypeptide chain is determined by the balance in the competition between the charge repulsion driving unfolding and hydrophobic interactions driving folding.

By analysis of the Swiss-Prot protein database [12], the authors were able to find 130 different, nonhomologous proteins with sequences sharing low mean hydrophobicity and relatively high net charge which they predicted to be natively unfolded.

Many globular proteins are unfolded by extremes of pH and substantial evidence indicates that this is caused by charge-charge repulsion. However, some globular proteins do not unfold under conditions of extreme pH.

It is likely that the outcome is determined by the balance in the competition between the charge repulsion driving unfolding and hydrophobic interactions driving folding. Thus, the situation is analogous to that with natively unfolded proteins.

It is known that unfolded proteins normally have very short lifetimes in the cell. Thus, it is most probable that natively unfolded proteins are significantly folded in their normal cellular milieu: natively unfolded proteins in vivo are likely to be stabilized by binding of specific targets, ligands (such

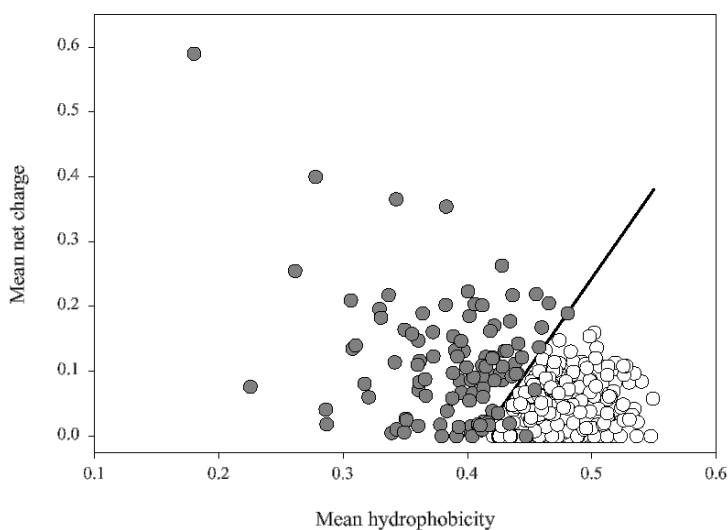


Figure 1.6: Comparison of the mean net charge and the mean hydrophobicity for the set of 275 folded and 105 natively unfolded proteins (gray circles) analyzed by Uversky et al [33, 29].

as a variety of small molecules, substrates, cofactors, other proteins, nucleic acids, membranes, etc.).

Moreover, for the majority of proteins listed, the existence of pronounced ligand-induced folding has been established. Consequently, in contrast to *in vitro* experiments with purified protein, natively unfolded proteins probably have considerable structure *in vivo* as the result of their interaction with their natural ligands.

The combination of low mean hydrophobicity and high net charge leads to natively unfolded conformation. This suggests that any interaction of natively unfolded protein with natural ligand that will affect its mean net charge, mean hydrophobicity, or both, may change these parameters in such a way that they will approach those typical of folded native proteins.

Unfortunately there are limited experimental data reported to confirm these predictions, since calculating the joint mean net charge and mean hydrophobicity of complexes of natively unfolded proteins with their natural ligands is rather difficult to implement.

However, evolutionary persistence of the natively unfolded proteins represents additional confirmation of their importance and raises intriguing questions on the role of protein disorder in biologic processes.

The results of the analysis summarized here are reasonable: amino acid sequences of proteins that have been shown to have little regular structure

under physiologic conditions differ significantly from the those of normal globular proteins, due to the combination of low mean hydrophobicity and relatively high net charge. High net charge leads to charge-charge repulsion, and low hydrophobicity minimally means less driving force for a compact structure.

It is clear that there are several ways in which such a specific sequence can lead to lack of a normal tightly packed globular structure. For example, in the case of α -synuclein the residues responsible are clustered mostly in the C-terminal region, and the isolated N-terminal region is predicted to fold. In other cases, the destabilizing residues are more uniformly distributed along the sequence.

The fully unstructured states are especially intriguing: where SAXS data are available, such domains appear to be highly elongated in solution and their hydrodynamic properties resemble those of a random coil, not the compact molten globule states formed during the unfolding of many globular proteins [6]. This feature, together with the unusual amino acid composition and distribution characteristics of intrinsically unstructured proteins may help them to evade temporarily, at least in eukaryotic cells, the proteolytic degradation machinery.

In summary, numerous proteins are intrinsically unfolded under physiological conditions and this is leading to a new view of bio-molecular recognition. No longer can binding be viewed as simply a lock and key event or as an interaction involving rigid macromolecular surfaces. Coupled folding and binding is seen to be common in interactions between biomolecules and appears to provide important advantages, especially in multicellular organisms.

Recent developments, both in theoretical and experimental models for unfolded macromolecules, are leading us to a deeper understanding of the nature of biological molecules and their interactions.

Chapter 2

Geometric approach to protein folding

The discovery of the structure of the DNA molecule has led to a description of the biology in complex living organism, based on chain molecules that store and replicate information and provide a molecular basis for natural selection [9].

Using the RNA molecule as an intermediary, the information contained in the genes is translated into protein molecules, which adopt a limited number of related structures, folding into thousands of native state structures under physiological conditions [34].

A protein molecule is large and has many atoms: in addition, the water molecules surrounding the protein play a crucial role in its behavior. Under the schemes of classical molecular dynamics, each protein is treated with all the details of the sequence of amino acids, their side chain atoms, and the water molecules. With such approaches, one can get a useful amount of informations on the chemistry of processes in which amino acids are involved. On the other hand, one may lose a unified way of understanding apparently disparate phenomena related to proteins.

Yet no simple unification has been achieved in a deeper understanding of the key principles at work in proteins. We restrict ourselves to globular proteins which display the rich variety of native state structures. Other interesting and important classes of proteins such as membrane proteins and fibrous proteins are not considered here.

A different approach to understanding proteins is presented in this chapter. The focus is on understanding the origin of protein structures and how they form the basis for both functionality and natural selection. The model points to a unification of the various aspects of all proteins based on symmetry and geometry, which are shown to determine the limited menu of folded

conformations that a protein can choose from for its native state structure, highly conserved throughout natural evolution [35]. These structures are in a marginally compact phase in the vicinity of a phase transition, and are therefore eminently suited for biological function.

Proteins are well-designed sequences of amino acids which fit well into one of these predetermined folds and they are prone to misfolding and aggregation leading to the formation of amyloids, which are implicated in debilitating human diseases such as Alzheimer and spongiform encephalopathies.

In the following sections we introduce the description of a protein as a thick polymer chain and highlight the differences in its phase diagram with respect to the usual string and bead model. Then a comparison of the predictions obtained from the simple tube model against experimental data available on protein native state structures is presented as well as a more refined model in which the tube picture is reinforced with the geometrical constraints that arise in the formation of hydrogen bonds.

2.1 Tube model of a thick polymer

Fluid and crystalline phases of matter can be understood from the behavior of a simple system of hard spheres [9]. The standard way of ensuring the self-avoidance of a system of uniform hard spheres is to consider all pairs of spheres, and require that their centers are no closer than their diameter.

Generalizing to a one-dimensional object, one must consider a line or a string, with space associated with each point along the line, leading to a uniform tube of radius of cross section (thickness) Δ , with its axis defined by the line.

The generalization of the hard sphere constraint to the description of the self-avoidance of a tube of nonzero thickness is done considering all triplets of points along the axis, and ensuring that their radii are bigger than the thickness [36].

This prescription entails discarding pairwise interactions and working with effective three-body interactions. One may visualize a tube as the continuum limit of a discrete chain of tethered disks or coins of fixed radius separated from each other by a very small distance. The inherent anisotropy associated reflects the fact that there is a special local direction at each position defined by the locations of the adjacent objects along the chain.

An alternative description of a discrete chain molecule is a string and bead model in which the tethered objects are spheres. The key difference between these two descriptions is the different symmetry of the tethered objects: upon compaction of a chain of spheres, each individual sphere tends

to surround itself isotropically with other spheres, unlike the tube situation, in which nearby tube segments need to be placed parallel to each other.

The tube and a chain of tethered spheres exhibit quite distinct behaviors with one exception in the presence of an attractive self-interaction favoring compaction. The chain and the string and bead model behave similarly in the limit of vanishing ratios of the radii of the coin and sphere to the range of attraction, where one gets a featureless compact phase. For a tube the simplest situation occurs in the swollen phase, where the finite size effects are not important, since they continue to adopt open conformations.

For a short tube, there are many more conformations that can be accommodated in the spherical topology than in the cylindrical topology without any accompanying sacrifice in the attractive interaction energy. There is a confluence of three distinct types of structures (the swollen conformations, the semicrystalline phase which one gets for a long tube due to its inherent anisotropy, and the featureless compact conformations), leading to quite remarkable finite size effects: a marginally compact phase is obtained with a reduction in the degeneracy [9].

Helices, hairpins, and sheets are ground states, with a parallel placement of nearby tube segments. The marginally compact phase is poised in the vicinity of a phase transition to the swollen phase.

The building blocks of protein structures are helices, hairpins, and almost planar sheets (fig. 2.1). Short tubes, with no heterogeneity, in the marginally compact phase form helices with the same pitch to radius ratio as in real proteins [37] and almost planar sheets made up of zigzag strands [9].

It is interesting to note that the helix is a very natural conformation for a tube and occurs without any explicit introduction of hydrogen bonding. Recent work on the denatured state of short amino acid sequences has suggested that the polyproline II helix might be the preferred structure in that phase, even though it does not entail the formation of any hydrogen bonds [9].

The tuning of the two length scales – the tube thickness (Δ) and the range of the compacting interactions (R) – to be comparable to each other happens automatically for proteins. The sizes of the amino acid side chains determine both the tube thickness and the range of interactions: steric interactions lead to a vast thinning of the phase space that protein structures can explore [15].

Physically, the notion of a thick chain follows directly from steric interactions in a protein: one needs room around the backbone to house the amino acid side chains without any overlap. The same side chains that determine the tube thickness also control the range of attraction: in fact, the outer atoms of the side chain interact through a short range interaction, screened by water.

Rapid folding of small proteins can be understood in terms of the inherent

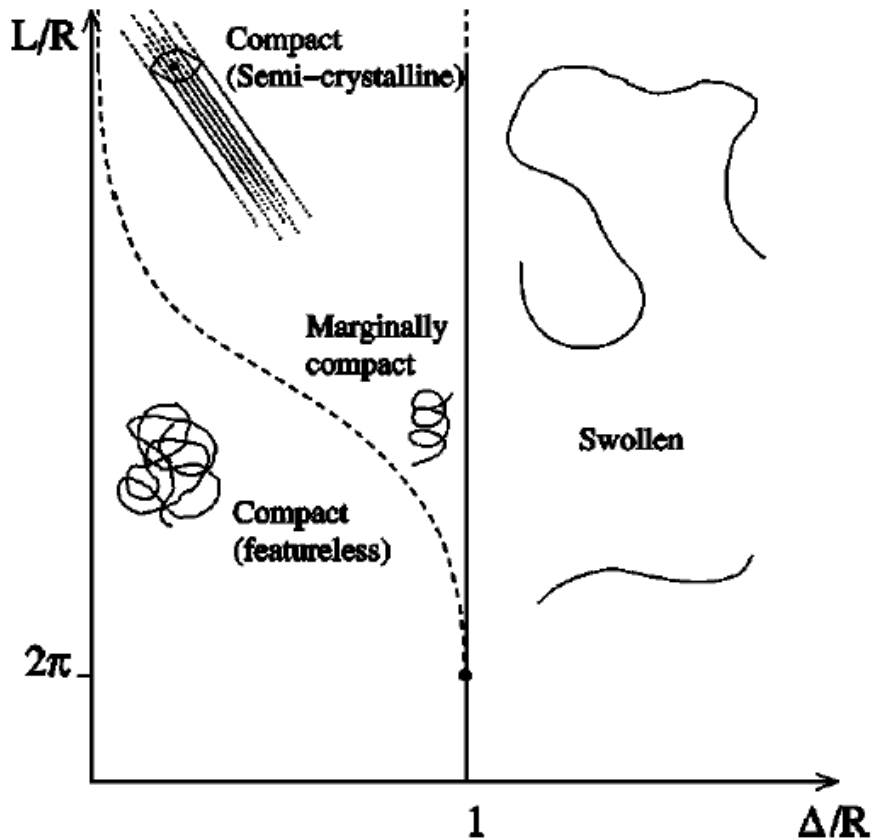


Figure 2.1: Zero-temperature phase diagram of a tube in the continuum with a self-attraction promoting compactness: the marginally compact phase is highlighted by the dashed line, accompanied by entropy reduction, with the choice structure being a helix with a well-defined pitch to radius ratio [37]. Other structures such as hairpins and sheets are present in the marginally compact phase for discrete chains [9].

anisotropy of a tube and the self-tuning of the two key length scales, the tube thickness and the range of the attractive interactions. In the marginally compact phase, in order to take advantage of the attractive interactions, nearby segments of the tube have to be parallel to each other and right up against each other: the helix and the sheet are characterized by such parallel space filling alignment of nearby tube segments.

In proteins, such an arrangement serves to expel the water from the protein core. As shown by Pauling and co-workers [38, 39], hydrogen bonds provide the scaffolding for both helices and sheets and place strong geometrical constraints stemming from quantum chemistry.

2.2 Refined tube model for proteins

The tube model in its simplest formulation can be used to describe any chain molecule with an effective thickness. A more refined model can be introduced to specialize to polypeptide chains. The tube geometrical constraint, a local bending energy penalty e_R , an overall hydrophobicity e_W , and effective hydrogen bonds between C^α s are the elements characteristic of the model.

The phase diagram and the associated structures for short homopolymers – chains made up of just one type of amino acid - of length 24, resulting from Monte Carlo simulations, are depicted in fig. 2.2.

In keeping with the behavior of the simple tube model discussed earlier, in the vicinity of the swollen phase one finds distinct assembled tertiary structures, quite similar to real protein structures, on making small changes in the interaction parameters e_R and e_W .

The striking similarity between the observed structures and real protein structures suggests that the model captures the essential ingredients responsible for the limited menu of protein native structures. These structures are the stable ground states in different parts of the phase diagram. Furthermore, conformations such as the $\beta - \alpha - \beta$ motif and the zinc-finger are found to be competitive local minima.

The specific structure depends on the precise values of the local radius of curvature penalty (a large penalty forbids tight turns associated with helices resulting in an advantage for sheet formation) and the strength of the hydrophobic interactions (a stronger overall attraction leads to somewhat more compact well-assembled tertiary structures). The topology of the phase diagram allows for the possibility of conformational switching, leading to the conversion of an α -helix to a β topology on changing the hydrophobicity parameter, in analogy with the influence of denaturants or alcohol in experiments.

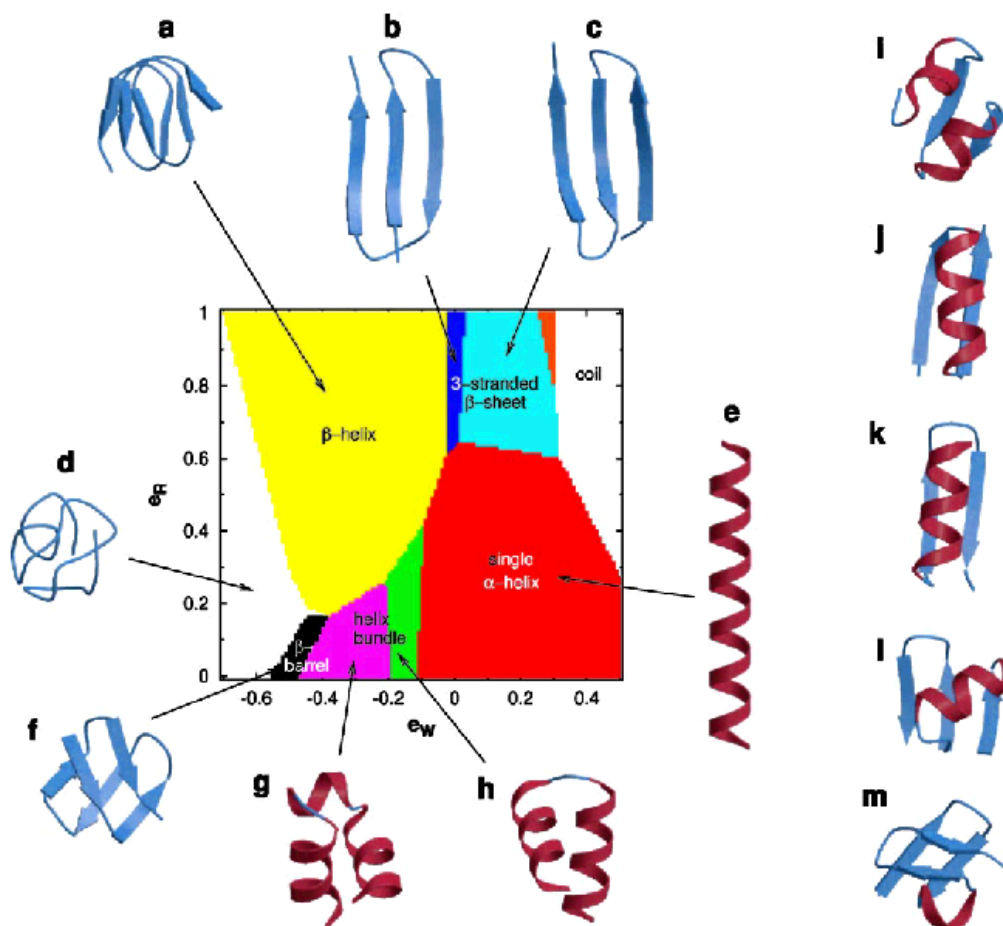


Figure 2.2: Phase diagram of ground-state conformations obtained from Monte Carlo simulations of a chain with 24 C^α atoms, where e_R and e_W represent the energy penalty due to the local radius of curvature and the solvent mediated interaction energy (structures drawn using Molscript [40] and Raster3D [41]).

2.3 Presculpted energy landscape

The standard approach of protein physics is to assume an overall attractive short range potential which serves to lead to a compact conformation of the chain in its ground state. In the absence of amino acid specificity or when one deals with a homopolymer, there is a huge number of highly degenerate ground states comprising all maximally compact conformations with high barriers between them (fig. 2.3).

The ground state degeneracy and the height of the barriers grow exponentially with the length of the homopolymer. The role played by sequence heterogeneity is to break the degeneracy of maximally compact conformations, leading to a unique ground state conformation which, of course, depends on the amino acid sequence. Yet, for a typical random sequence, the energy landscape is still very rugged.

A model protein moving in such a rugged landscape can be subject to trapping in local minima and may not be able to fold rapidly, so that glassy behavior may ensue due to such trapping.

Bryngelson and Wolynes [42] suggested that there is a principle of minimal frustration at work for well-designed sequences, in which there is a fit between a given sequence and its native state structure, resulting in a funnel-like landscape [43]. This promotes rapid folding and avoids the glassy behavior: given a sequence of amino acids, with side chains and surrounding water, one obtains a funnel-like landscape with the minimum corresponding to its native state structure.

The model calculations show that the large number of common attributes of globular proteins [44] reflects a deeper underlying unity in their behavior. A consequence of this hypothesis is that the main features of the energy landscape of proteins result from the amino acid common features of all proteins.

This landscape is (pre)sculpted by general considerations of geometry and symmetry (fig. 2.3): for each of the local minima the funnel-like behavior is achieved already at the homopolymer level in the marginally compact part of the phase diagram (fig. 2.2).

The already mentioned self-tuning of the two key length scales to be comparable to each other, and the interplay of the three energy scales – hydrophobic, hydrogen bond, and bending energy – in such a way as to stabilize marginally compact structures, provide the close cooperation between energy gain and entropy loss needed for the sculpting of a funneled energy landscape.

Recent work has shown that the rate of protein folding is not too sensitive to large changes in the amino acid sequence, as long as the overall topology

of the folded structure is the same [45]: a presculpted landscape greatly facilitates the design process. Even within crude design schemes, which take into account the hydrophobic (propensity to be buried) and polar (desire to be exposed to the water) character of the amino acids, is sufficient to carry out a successful design of sequences with one or the other of the structures shown in fig. 2.2.

The matching of the hydrophobic profile of the designed sequence to the burial profile, measured by the number of neighbours within the range of the hydrophobic interaction, leads to the correct fold in a Monte Carlo simulation. The sequence HPPHHPHHPHPPPPPHHPHHPHPPPP, with $e_R = 0.3$ for all residues, $e_W = 0.4$ for contacts between H and H, and 0 for other contacts, has as its ground state the two-helix bundle structure (fig. 2.2, e) whereas HPHHHPPPPHHPHHPHPPPPHHHPP prefers the $\beta - \alpha - \beta$ motif (fig. 2.2, j,k).

It is interesting to note that the $\beta - \alpha - \beta$ motif is only a local minimum in the phase diagram of a homopolymer but is stabilized by the designed sequence: many protein sequences adopt the same native state conformation, and once a sequence has selected its native state structure, it is able to tolerate a significant degree of mutability – except at certain key locations – with multiple protein functionalities that can arise within the context of a single fold [46].

There are several attractive features of the picture based on the tube-protein hypothesis. First, protein structures lie in the vicinity of a phase transition to the swollen phase which confers on them sensitivity, especially in the exposed parts of the structure, to the effects of other proteins and ligands.

The flexibility of different parts of the protein depends on the amount of constraints placed on them from the rest of the protein. From this point of view, it is easy to understand how loops, which are not often stabilized by backbone hydrogen bonds, can play a key role in protein functionality.

The existence of a presculpted energy landscape with broad minima corresponding to the putative native state structures, and the existence of neutral evolution demonstrate that the design of sequences that fit a given structure is relatively easy, leading to many sequences that can fold into a given structure.

This freedom facilitates the accomplishment of the next level task of evolution through natural selection: the design of optimal sequences, which not only fold into the desired native state structure, but also fit in the environment of other proteins.

A useful protein can interact with other proteins without being subject to the tendency to aggregate into the amyloid form. This suggests that

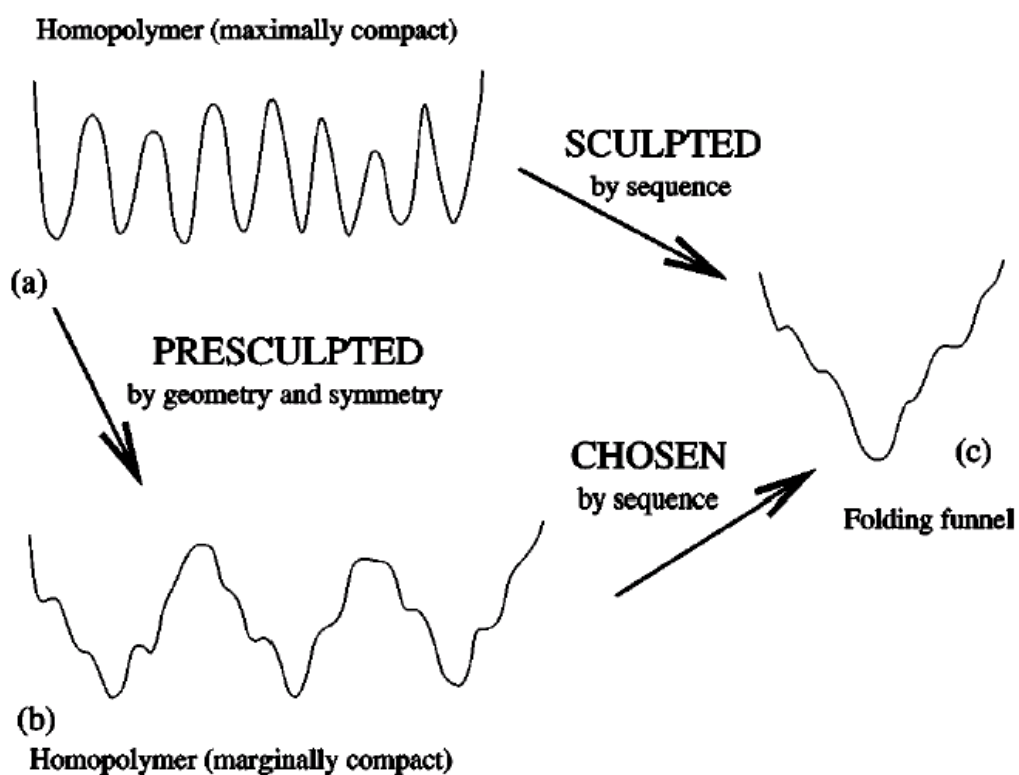


Figure 2.3: One-dimensional sketch of energy landscape. On the horizontal axis a rough distance between different conformations in the phase space is represented: the barriers in the plots refer to the free energy needed to travel between two adjacent local minima.

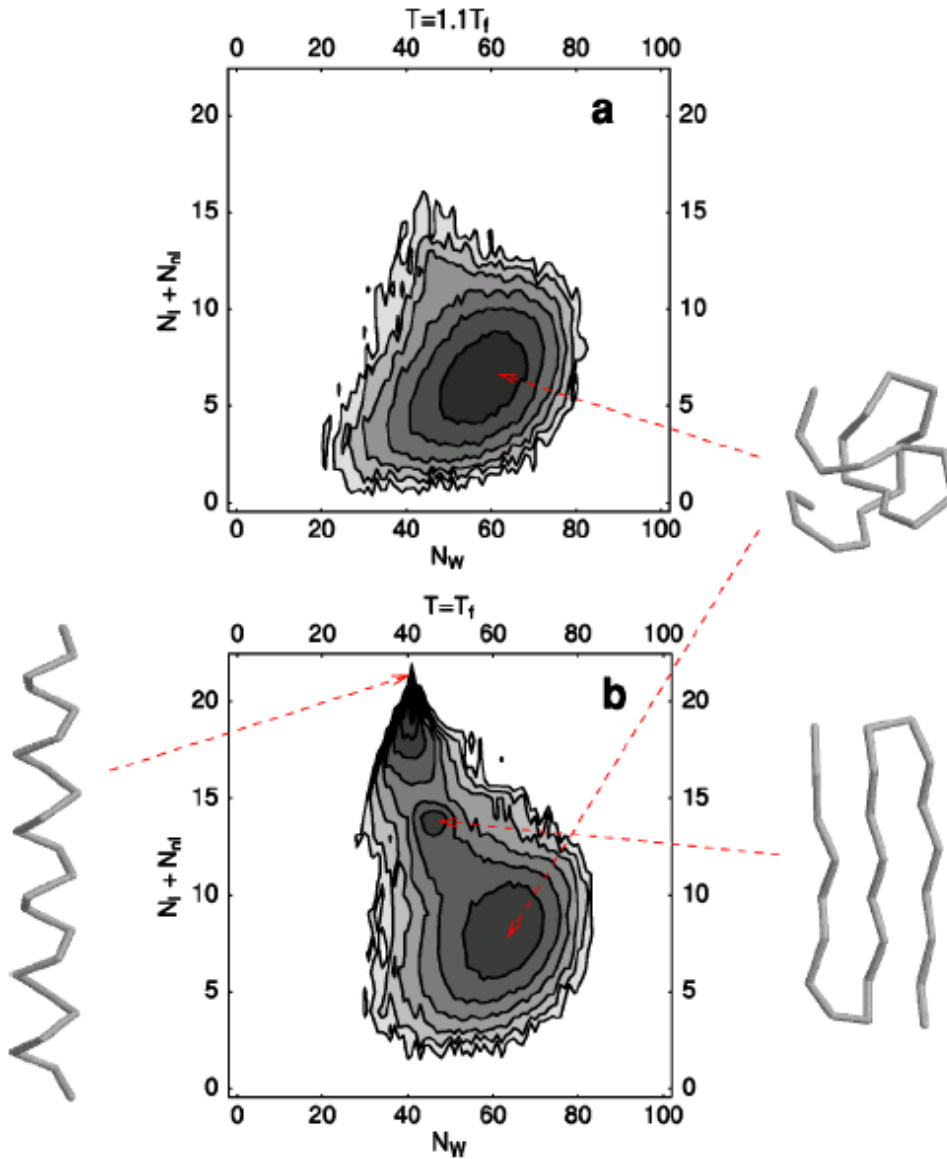


Figure 2.4: Contour plots of the effective free energy at high temperature (top) and at the folding temperature (bottom) for a 24-residue long homopolymer ($e_W = 0.08$, $e_R = 0.3$). The effective free energy is defined as $F(N_l + N_{nl}, N_W) = -\ln P(N_l + N_{nl}, N_W)$, which depends on the number of hydrogen bonds and the number of hydrophobic contacts. The histogram is collected from equilibrium Monte Carlo simulations at constant temperature. The spacing between consecutive levels in each contour plot corresponds to a free energy difference of $k_B T$. The free energy minimum at higher temperature corresponds to the denatured state; typical conformations from each of the minima at the folding transition are shown.

protein engineering studies aimed at improving enzymatic function ought to be carried out in a two steps: the family of sequences that fold into a desired target structure needs to be selected and a finer design needs to be carried out in the context of the substrates and the other proteins that the target protein interacts with.

After one obtains a presculpted energy landscape with relatively few folds, protein might fold in a cooperative manner into native state conformation: there is the possibility of straightforward design of optimal sequences that fit into a desired structure in a marginally compact phase, having the flexibility needed for biological function. How then a given sequence is able to reach its native state conformation starting from its denatured conformation?

The denatured state is an ensemble of open conformations that the protein adopts when it is not under physiological conditions: recent work has underscored the possibility that the number of accessible conformations is severely reduced compared to a random chain [47], leading to biases in the chain direction that persist over the entire length of the protein [2]. Long-range structure, which cannot be removed by strongly denaturing conditions, could arise predominantly from local steric hindrance.

Just as there is a one-way correspondence between a sequence and its native state structure, there could exist a similar correspondence between the sequence and its denatured state: the denatured state can be thought of as an address of the native state conformation lying within its basin of attraction.

Unlike the native state, the denatured state has a larger entropy and comprises open conformations. Because of this, water plays a quite crucial role in the denatured state. Both the above factors lead to local interactions playing a more important role than nonlocal interactions in the denatured state.

It has been shown that denaturation by at least three different agents – truncation, urea, and acid – gives rise to essentially the same persistent native-state like topology [2].

An interesting consequence of the type of denatured state described above along with the existence of the presculpted landscape is the possibility of disordered proteins [10], that are in temporally fluctuating denatured form but which fold in the presence of distinct substrates to carry out multiple functionalities.

In the present picture these sequences need appropriate stabilizing substrates to fold and without that the protein is denatured. Given that finite size effects are severe for proteins, the presence of different substrates – leading to different boundary conditions – would not only favor one competing structure over the others, but also result in folding to that structure. The

simultaneous existence of the distinct folds in the energy landscape allows the protein to choose from among them depending on the precise nature of the stabilizing influence.

2.4 Summary

The model described here introduces a unifying view on proteins, naturally leading to a finite number of protein folds. This number grows with the size of the protein, but is limited by the fact that proteins beyond a characteristic length form either autonomous domains or amyloids [9].

The inherent anisotropy, due to the tube-like description, is able to capture some aspects of the secondary structure motifs arising in proteins. Protein structures are modular in form, being simple assemblages of helices and strands connected by tight turns.

This unified picture leads to a single free energy landscape with two distinct classes of structures. The amyloid phase is dominated by β -strands linked to each other in a variety of forms whereas the native state structure menu is an assembly of α -helices and β structures.

Pauling and coworkers [38, 39] considered the protein backbone and explored the structures consistent with both the backbone geometry and the formation of hydrogen bonds, predicting that helices and sheets are the structures of choice. Ramachandran and Sasisekharan [15] stressed the role of excluded volume and steric interactions between the adjacent amino acids in reducing the available conformational phase space, with the two significantly populated regions of the Ramachandran plot corresponding to the α -helix and the β strand.

Even though hydrogen bonds and sterics are not related to each other, they are both promoters of helices and sheets. The marginally compact phase of short tubes has helices and sheets as its preferred structures: hydrogen bonds serve to enforce the parallelism of nearby tube segments, a feature of both helices and sheets, while sterics emphasizes the nonzero thickness of the tube and serves to place it in the marginally compact phase.

The marginally compact phase is a finite size effect: this may explain why proteins tend to be relatively short, at least compared to conventional macromolecules, including DNA. While sequences and functionalities of proteins evolve, the folds that they adopted, which in turn determine function, seem to be determined by physical laws and are not subject to evolution.

Chapter 3

Binding and Folding

The mechanism through which an IUP and its protein target bind exploits hydrogen bonding and electrostatic interactions between amino acids of the chains. The number of interactions can be large, leading to a high negative binding enthalpy balancing the entropy loss and allowing binding, which becomes thermodynamically favourable [10].

The aim of the present study is to clarify the effect of the contact interactions between an IUP and its target protein on the energy landscape of the first one, biasing its ground state towards motifs out of a restricted menu of folds, depending on the geometric properties of the pattern. The folding upon binding, typical of the disorder-order transition of several IUPs, is the way nature actively performs structural design for a specific biological function to be accomplished: contact interactions between the disordered protein and the binding partner have to be carefully selected by evolution since a proper geometrical pattern can bias efficiently the folding towards the desired structure.

We will try to implement a similar design, even if from a coarse-grained perspective, by choosing a contact pattern capable to bias the folding: to perform our task we will use an approach to the protein folding problem simply based on geometrical considerations. The IUP is constrained to move within a cubic box and its partner is represented in a coarse-grained framework, through three contact points lying on the inner part of the box bottom face: the geometrical arrangement of the points on the surface is tuned in order to bias the folding towards the target fold, thus trying to implement a molecular recognition mechanism. The bottom wall of the box is exploited to simulate the steric hindrance opposed by the protein partner to the approaching IUP, subject to a substantial loss of conformational entropy.

The contact pattern is defined given the coordinates of the binding centers, the radius of the contact interaction with the atoms along the homopoly-

mer chain: the one-to-one correspondence between polymer beads and contact centers switches on specific interactions, and the model system thus becomes a heteropolymer (homopolymer with contact bias).

3.1 Methods

The model used in this study is based on the geometrical approach to protein folding outlined in chapter 2: a coarse-grained, off-lattice, polymer model where only alpha carbon atoms are used to describe the protein's backbone, placed at the bond length of 3.8 Å along the chain.

The use of C^α atom only to represent chain molecules is not restricting, since it has been shown [48, 49] that one can approximately reconstruct the locations of all the backbone atoms, and even of the side chains ones, with just the knowledge of the C^α positions, through automatic procedures generating full protein coordinates, given the amino acid sequence.

Hydrophobicity – the entropic driving force for the binding of non polar solutes within solvent phases containing water – is incorporated by means of an effective pairwise attraction between hydrophobic amino acids of magnitude e_W .

The special local direction at each amino acid along the chain, defined by the position of the neighboring residues, is captured by employing the tube-like description of chapter 2, which leads naturally to the emergence of secondary motifs [37, 50]. The model penalizes sharp local turns of the backbone by means of a bending energy penalty of magnitude e_R .

The energetics and geometry of hydrogen bonds are encapsulated in the model based on a statistical analysis of protein native structures [8, 9]. Amino acid unspecific constraints have been found on the relative orientation of the intrinsic, Frenet, coordinate systems associated with the C^α atoms of amino acids between which hydrogen bonds are formed. The independence of such constraints on the types of the hydrogen bonded residues leads to a significant simplification.

The sampling of polymer's configurations is carried out through Monte Carlo simulations with the Metropolis acceptance rule for commonly used moves in stochastic chain dynamics [51], like the standard pivot and crankshaft conformational rearrangements.

The pivot algorithm [52, 53, 54, 55] acts selecting a random bead along the chain, excluding both extremities. Then a random direction in space is chosen and one part of the chain is rotated, pivoting on the selected hinge, along a gaussian distributed angle with zero mean around the direction previously defined.

The crankshaft move [53, 54] selects instead two beads along the polymer and then rotates the enclosed part of the molecule along a random angle, gaussian distributed with mean zero as before.

The first one is known to be highly efficient in decorrelating chain conformations for self-avoiding walks, in the case of long-range properties of polymers [53, 52]. The second one is faster, as required to local moves.

Beside these two standard chain rearrangements we employed two other moves: a reptation-like move in order to make the decorrelation of chain conformations easier, and an overall translation of the chain along a randomly chosen direction, at regular time intervals.

The reptation move was devised as a quick means for decorrelating polymeric chains when not bound to the effective pattern: a bead is chosen randomly along the chain and then the shortest part of the homopolymer is translated rigidly to match the opposite end.

This is a very simplistic scheme adapted from the slithering-snake algorithm for polymer models on a lattice [56]. Although quite expensive computationally, and with a high probability of rejection, it is able to rapidly decorrelate chain configurations, at least in the phase where the polymer is not bound to the substrate.

The overall translation of the modeled molecule is performed only to avoid the polymer to get stuck in local regions and corners of the cubic box, thus increasing the probability of rejection for both local and non-local moves.

The Metropolis acceptance/rejection test is employed with the usual thermal weight $e^{-\mathcal{E}/T}$, where \mathcal{E} is the energy of the conformation and T is an effective temperature. Monte Carlo simulations performed to study the order-disorder transition are applied to a polymeric chain constrained within a cubic box of side 50.0 Å where the polymer is able to perform the conformational rearrangements mentioned above (pivot, crankshaft, reptation, translation). The walls of the box are an infinite energy barrier, needed to confine the polymer to the vicinity of the contact pattern, thus allowing an efficient sampling of the pattern-bound configurations.

Model's parameters defining the ground-state fold are listed below (table 3.1). Energy parameters of the homopolymer model [8] are kept fixed, whereas only the new parameters taking into account the geometric and energetic properties of the contact pattern are modified.

Monte Carlo simulated annealing [57] is used to explore the ground-state of the system, formed by the homopolymer together with a fold-specific contact pattern, inside the simulation box. Simulations at constant temperature, carried out with the Metropolis acceptance criterion, are used to study the thermodynamic behaviour of the system. Both types of simulations, using

Energy parameters of the homopolymer model [8] (arbitrary units)	
Local hydrogen bonds (e_l)	-1.0
Non local hydrogen bonds (e_{nl})	-0.7
Cooperative hydrogen bonds (e_{co})	-0.3
Effective hydrophobicity (e_w)	-0.08
Curvature penalty (e_r)	0.3

Parameters of the model with substrate	
Box side	50 Å
Number of substrate points	3
Range of effective contact interaction (r_b)	2.0 Å
Energy of contact (e_b)	$3 e_l$

Table 3.1: Parameters of the model

the above described moves, have been extensively performed on the system. The more efficient and reliable technique of parallel tempering [58], also called replica exchange method, has been implemented and adopted on the homopolymer, in order to confirm thermodynamical properties obtained under the single temperature scheme.

The reptation move has roughly 30 % acceptance rate when applied together with the pivot algorithm when the polymer is not adsorbed on the substrate. As noticed above, even if it is computationally expensive it is able to rapidly decorrelate chain conformations. This was confirmed in a preliminary set of tests performed on the homopolymer within the box, without any substrate. Energy autocorrelation function were computed for simulations carried out both with the reptation move and without it: by measuring the corresponding correlation times, we were able to provide an estimate of the efficiency of the reptation move.

Several frequency ratios for different moves has been tested for the rearrangements of the polymer configurations. It has been noticed that high rates of pivot moves are more efficient to sample β -like conformations, the ones when non local hydrogen bonds are predominant. On the other hand, crankshaft moves are faster to build up helical conformations, where local hydrogen bonds are dominant. In fact it is known that pivot is able to quickly decorrelate global quantities for the self-avoiding walk, whereas it is not as efficient for local properties [53, 54].

We tried to keep sampling efficiency as well as speed during the simulations: one of the best schemes to explore the space of conformations uses pivot at 20%, reptation at 10%, crankshaft at almost 80%, with few moves left for the random translation of the chain.

The increased efficiency obtained using large frequencies of non local moves is reasonable for small chains, i.e. built-up by less than 64 beads. This has been tested by numerical studies on the dynamical properties of the pivot and the slithering snake algorithm for lattice polymer models [52, 55, 56]. It was shown as well that the computational cost of non local rearrangements grows more rapidly than for local ones with the growth of the size of the system.

In order to avoid the polymer to get stuck in local regions and corners of the cubic box, a random translation of 2.5 Å is attempted on the chain with frequency 0.1 %.

The move is accepted provided that each bead of the chain stays inside the simulation box in the absence of the contact pattern: the space inside the box is isotropic, and the energy does not change for overall translations of the chain. When a substrate is present inside the box, the additional Metropolis test on energy needs to be satisfied instead.

In the following sections we will write “target structure” or “target fold” to mean the particular fold from the menu of fig. 3.2 that we wish to observe, given a set of interaction centers, which might be or not in correspondence with selected polymer beads.

3.2 Tuning model’s parameters

In this section we list and try explain the rationale lying behind the choice of the parameters used within the simulations.

The size of the simulation box is chosen to be 50 Å. The box side has therefore an intermediate length between the average size of a random walk with twenty-four steps of 3.8 Å, and the fully stretched polymer. Full stretching and full rearrangement for the polymer inside the box are thus allowed during the simulation.

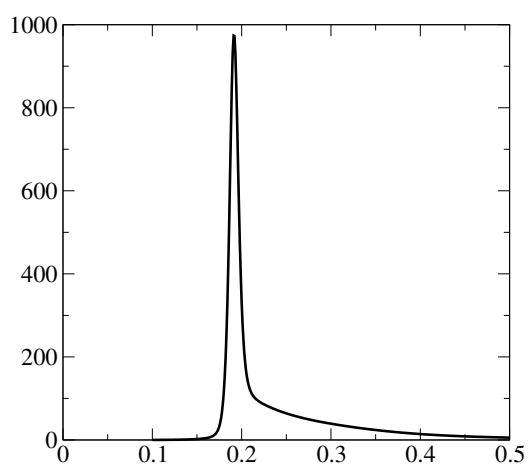
The presence of the simulation box does not appreciably modify the thermodynamical properties of the system with respect to the simulation in the bulk. A comparison of the results for the contour plots of the effective free energy obtained within the simulation box (fig. 3.1) with those computed in a previous study (fig. 2.4, chapter 2) shows the consistence of the two.

Figure 3.1 shows the plot of the specific heat for the process of folding in the absence of any substrate with the walls of side 50.0 Å whose temperature dependence was computed from several simulations at constant temperature using the multiple histogram technique [59, 60, 61]. The sharp peak in the plot is related to the folding to the α -helix, which is the ground state at the values set for the effective hydrophobicity and curvature penalty.

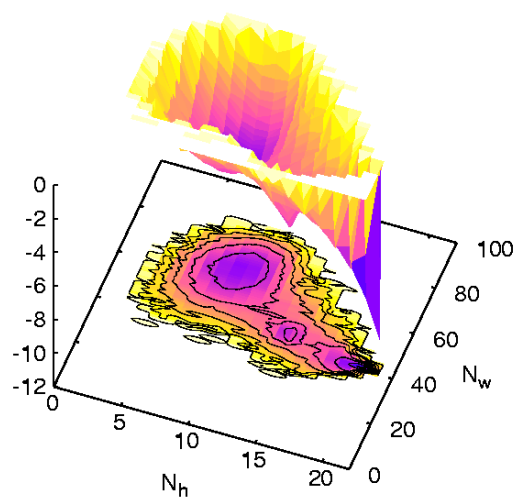
Surface and contour plots of the effective free energy at the folding temperature $T = 0.192$ given by the peak in the specific heat is shown on the right 3.1(b). Histograms $P(N_w, N_h)$ used in the plots have been collected from equilibrium Monte-Carlo simulations keeping the temperature constant. The effective free energy is defined here as a function of the number of hydrophobic contacts N_w (y axis) and of the total number of hydrogen bonds N_h (x axis), $F(N_w, N_h) = -\ln P(N_w, N_h)$. The spacing between consecutive levels in each contour plot corresponds to a free energy difference of $k_B \tilde{T}$, where \tilde{T} is the temperature in physical units.

The thermodynamic properties of the system obtained by the use of several single temperature computations have been fully confirmed within the more reliable scheme of parallel tempering, which has been implemented as well on the same system. The replica exchange method is particularly efficient especially at temperatures near the folding transition or below it.

A substrate modeled with few centres of an effective contact interaction



(a) Specific heat versus temperature.



(b) Free energy at the folding temperature.

Figure 3.1: Plot of the specific heat (3.1(a)) and plot of the effective free energy surface with contour lines (3.1(b)), as it has been defined within the text – of the polymer inside the simulation box at the folding temperature, without contact interactions.

is employed to bias the binding and subsequent folding of the polymer. The binding partner of the IUP is represented in a coarse-grained framework, through few contact points lying on one of the faces of the box. The geometrical arrangement of the points on the surface is tuned in order to bias the folding towards a specific structure: we are thus trying to implement a one-to-one correspondence between the pattern and the target fold, i.e. fold specificity through a molecular recognition mechanism.

The number of points of the contact pattern is fixed to three. This is the least amount of points necessary to define a plane, and the least that proved to be capable to bias relevantly the folding towards the target structure.

In few cases even two interaction centers only were able to produce a structure different from the α -helix. This has been observed with a two-helix-bundle, which has never occurred in simulations without the pattern at the chosen values of effective hydrophobicity and curvature penalty. Nonetheless these structures appeared only in a tiny fraction of the total number of annealings used to test the pattern, before computing thermodynamic properties. Thus they were discarded as not relevant enough to promote the target fold. The contact pattern is chosen to lie on the inner surface of the simulation box: in this way we exploit the walls to simulate the steric hindrance opposed by the protein target to the disordered protein. Therefore a substantial conformational entropy loss of the the IUP is induced, while the polymer explores different configurations during the process of binding.

The contact pattern is defined given the coordinates of the binding centers, the radius of the contact interaction (r_b), the contact energy associated to each formed contact (e_b) and their specific target beads along the homopolymer chain, which may be one or more. Several schemes have been tried for the interaction between the polymer and the substrate: from totally unbiased interactions, where every bead bound to a centre lowers the energy of the chain, to only three target beads for each centre.

Finally, a one-to-one interaction, where only one specific bead along the chain can bind to its partner on the substrate. Forbidding all polymer beads to interact with every contact centres, we are actually switching on specific interactions, and the model system thus becomes a heteropolymer – a homopolymer with contact bias.

The choice of only three points, for the substrate responsible of the induced binding and subsequent folding of the polymer, avoids the trivial case of a $G\bar{o}$ model approach: in that case all the native contacts are biased. If we put a bias on n atoms, this requires that $\binom{n}{2}$ distances among the beads are fixed. If n becomes comparable with the length of the chain, the number of constraints is similar to the $G\bar{o}$ model. In the present case instead, folding

is only loosely guided.

Following Bogner and coworkers [62] – who used a simplified model on lattice with substrates of lower dimensionality – we select a two-dimensional substrate defined by the set of points belonging to the interaction pattern. This turns to be sufficient to bias the fold towards structures different from the α -helix. Depending on the value of the binding energy per contact (e_b) the new fold is a local minimum of the free energy with respect to the α -helix, which is often kept as the ground state from the balance between energy and the entropic contribute at constant temperature:

$$\Delta F = \Delta U - T\Delta S \quad (3.1)$$

where U is the internal energy of the chain.

Specific interactions severely constrain the number of accessible configurations of the polymer at a given temperature. Thus the entropy of disordered configurations is lowered: given the same amount of energy gain, a transition like folding towards the target structure may occur at a higher temperature.

On the contrary, if one is not able to hinder the occurrence of antagonist folds – like the α -helix, in the present case – then competition among several local minima is encountered. This results in a smaller free energy difference between the target and its closest antagonist, thus delaying the folding to lower temperatures, if not even avoiding it at all.

Successful visit of the target structure occurs when a properly designed substrate is present within the simulation box. On the contrary, during the simulations without the contact pattern it happens seldom to observe the competing folds listed in table 3.2. In fact, only the three-stranded β -sheet is met as a local minimum during very long simulations at constant temperature without a proper contact pattern and $e_w = -0.08$, $e_r = 0.3$.

Hydrophobicity and curvature penalty partly account both for the physiological conditions of the solvent (e.g.: pH, concentration of denaturants) and to a limited extent for the chemical nature of the chain (e.g.: steric hindrance of residues side chains). The value of effective hydrophobicity e_w is -0.08 and the curvature penalty e_r is 0.3 (arbitrary units). The corresponding native state is an α -helix, still other local minima are present (table 3.2).

The specific values adopted have been chosen for the small difference in energy of the different folds reported in table 3.2 and for the presence of a broad three-stranded β -sheet local minimum in the energy landscape, beside the α -helical ground state [8]: the free-energy landscape is suitable to be modified in favour of the local minimum of type β by a proper contact bias, as it will be shown later.

Moreover the thermodynamic properties of the homopolymer with such values have been deeply investigated in a previous study [8], thus allowing

a direct check in the possible effects of the box boundary conditions on the behaviour of the system, if any. As noticed above and previously shown in figure 3.1(a), the presence of the restraining walls has no appreciable effects on the thermodynamic properties of the polymer.

The specificity of the pattern towards the selected target fold depends heavily on the radius of the interaction. Different values for r_b has been tried, ranging from 2.0 to 3.8 Å, which is the distance of nearest-neighbouring α -carbons. The latter allows folding into several structures, whereas the first one is rather specific, though it is still sufficiently large to let the polymer explore for a sufficiently long time the configurations in the vicinity of the free energy minimum.

The choice fell on $r_b = 2.0$ Å: this value allows us to bias different target beads towards the same interaction centre of the substrate. In fact, the self-avoidance imposed on the chain does not allow two beads to lie within a distance smaller than 4.0 Å [8].

Another key parameter to be tuned is the energy of contact interactions: most extensive simulations has been performed using twice and three times the value of local hydrogen bond energy (e_l), which is -1.00 in the model [8].

As it will be shown later, the value of the energy for contact interactions between beads lying on the polymeric chain and atoms of the substrate has a direct influence on the absorption temperature. Checking the temperature dependence of the average number of contacts between the polymer and the substrate, one can see that the adsorption corresponds to the first peak in the plot of the specific heat function against temperature occurs, accounting for a transition from the free polymer to the bound state.

The value $e_b = 3 e_l$ is a strong bias needed to make the target fold become the ground-state, when it happens to have a relatively high energy in comparison to the α -helix. This is the case of the three-stranded β -sheet, fig. 3.2. However, even using $e_b = 2 e_l$ higher energy folds are observed as local minima, whereas in the absence of the pattern they were not observed during a sufficiently long simulation at constant temperature.

3.3 Substrates for binding and folding

The aim of the work discussed in this thesis relies heavily upon the design of a geometrical pattern for the interaction points of the substrate. This needs to be appropriately designed, in order to bias the process of binding first, and folding afterward, of the disordered protein – here represented as a coarse-grained polymer of twenty-four beads.

The selection of the proper pattern is done in few steps:

$e_w = -0.08$	$e_r = 0.3$
---------------	-------------

Motif	Energy
α -helix	-23.68
two-helix-bundle	-22.96
β -sheet	-17.64
greek-key	-19.12
β -helix	-20.12
β -barrel	-19.10
$\beta - \alpha - \beta$	-20.64
zinc-finger	-19.18

Table 3.2: Fold energy

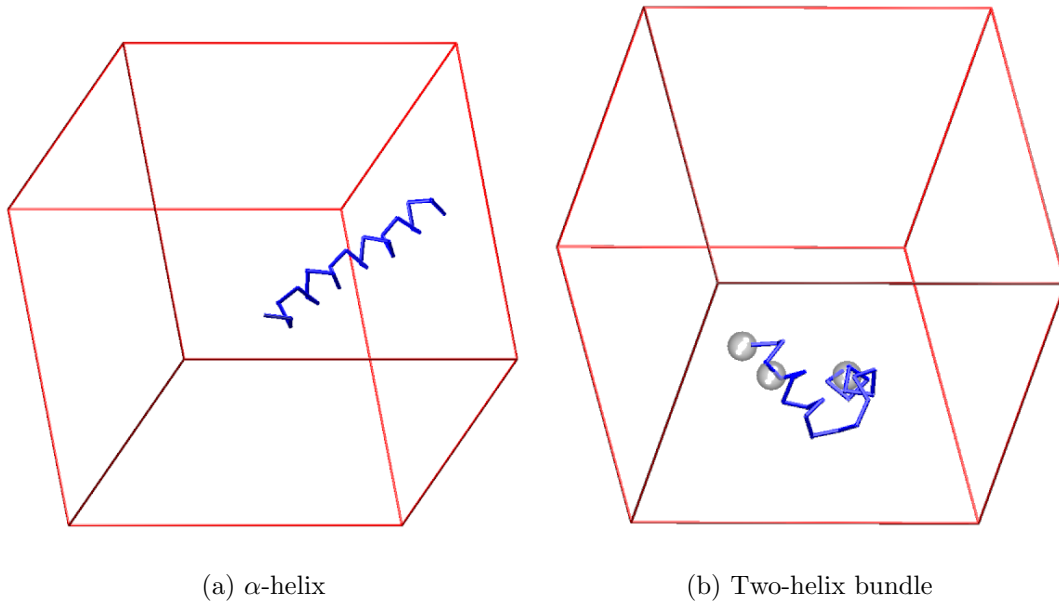


Figure 3.2: Ground states of the model system inside the simulation box in the absence of contact interactions and in the presence of a substrate geometrically tuned for a two-helix-bundle fold.

- The target conformation of the polymer that is wanted to be the ground state of the system is analyzed and three beads are selected. Selection of the proper beads is performed by choosing the three atoms in such a way that the rest of the polymer is left on one side of the plane that they define.

In this way the three interaction centres may be modeled as spheres lying on one side within the simulation box. The wall of the box restrains the configurations space accessible to the polymer, while approaching the substrate. This accounts for the conformational entropy loss accompanied by the process of binding.

- Once the geometry of the substrate has been modeled, an extensive search of the structures adopted by the homopolymer, favoured by that pattern has been done. We employed Monte Carlo simulations, both using annealing and constant temperature runs, during which we allowed all the beads along the polymer to bind to every interaction centre on the substrate.

In this way one can directly verify the entropic selectivity of the substrate against the selected fold, which is not accessible with simple geometric arguments.

- The last step is the selection of the correspondence between interaction centres of the substrate and beads of the chain to accomplish the interacting scheme of the disordered protein.

In fact the geometric properties of the substrate are not enough alone to guarantee proper selectivity towards the desired fold, and we are compelled to bias the interactions between only selected beads along the chain and certain centres on the substrate.

One-to-one correspondence between an atom on the polymer and a centre on the substrate is an interaction of the type used by Gō models. This is the most specific kind of interaction one can adopt in such a simplified scheme, and this actually provides an efficient way of designing specific folds on a substrate.

In order to observe the different folds, with different relative frequencies than in the simulation without the fold-specific contact pattern, one has to carefully design the geometric properties of the pattern, which should suit the geometric arrangements of the target beads on the target structure.

At the same time one can also perform a negative design, in the case of biased interactions with specific target beads, by hampering the folding into

the α -helix, exploiting the geometric specificity of the addressed structure. Let d_{ij} and d_{ij}^α be the native distances of polymer target beads i and j in the desired fold and in the α -helix. Then a necessary condition to be satisfied in order to hinder the fold into an α -helix is

$$|d_{ij} - d_{ij}^\alpha| > 2r_b \quad (3.2)$$

This criterion may account for the energetic selectivity of the pattern: the desired fold will be the ground state of the free-energy landscape at zero temperature. This is not sufficient, in general, to avoid the α -helical fold, which is quite ubiquitous and difficult to hamper. Often the formation of only one contact between the polymer and the pattern is enough to bias the folding towards an α -helix.

The entropy contribute of the single helical configurations to the free energy landscape is crucial. This is out of reach from the tuning of the geometrical placement of interaction centres. In those cases, the steric hindrance due to the box walls needs to be fully exploited, by carefully choosing the identity of the target beads on the polymer to be associated with the interaction centres on the substrate.

The different folds listed in table 3.2 are ground states in the energy landscape of the homopolymer, for suitable values of the effective parameters e_w and e_r . Only the three-stranded β -sheet is a local minimum for the particular values selected in this study.

Interaction of the homopolymer in the box with several fold-specific patterns have been performed, in order to test the efficiency of the number of contact points and the geometric specificity of the chosen pattern towards the selected structure. Some of those substrates are shown in table 3.3. The relevant coordinates are only the ones on the plane xy : the height z is fixed, because all the interaction centres belonging to the substrate lie on one inner surface of the simulation box; the centre of the simulation box is the origin of the reference system.

Pictures showing the typical kind of folded structure obtained using the same values of e_w and e_r and the different type of substrates listed in table 3.3 are shown in the following pages (drawn using VMD [63], Molscript [40] and Raster3D [41]).

3.4 Unbiased interactions with the substrate

Introducing suitable contact interactions, we managed to observe the folding into several structures of the list of table 3.2, even without any specific bias.

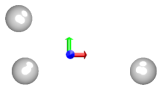
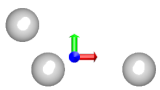
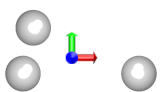
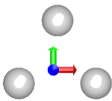
Target fold	Coordinates	Target bead	Geometry
β -sheet	-10.00 7.81 8.65 0.00 -9.05 0.00	8 18 24	
two-helix-bundle	0.00 0.00 10.90 0.00 -3.07 5.28	20 1 24	
zinc-finger	-1.16 0.00 11.99 0.00 0.00 5.15	15 19 24	
$\beta - \alpha - \beta$	-4.93 0.00 4.93 0.00 0.00 8.00	12 21 18	

Table 3.3: Different substrates for polymer binding and folding, with x and y axes shown in colour.

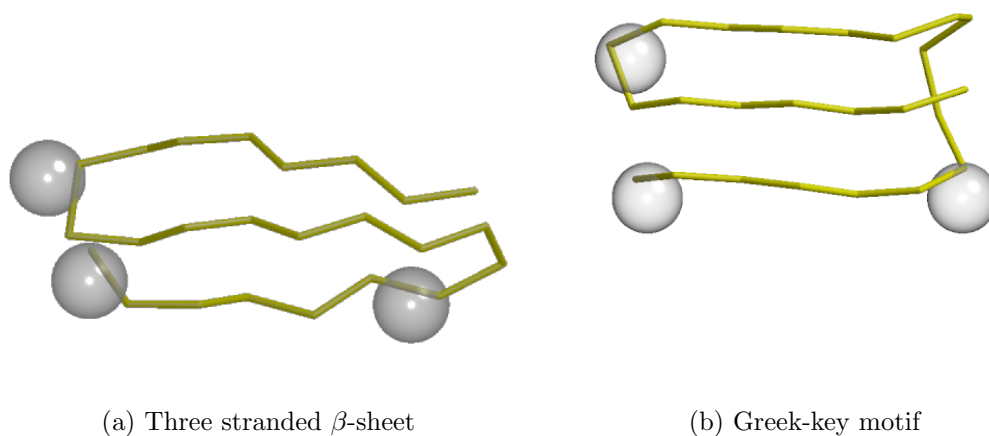


Figure 3.3: Ground state (3.3(a)) and local minimum (3.3(b)) in the effective free-energy landscape of the system with the substrate designed for the three-stranded β -sheet.

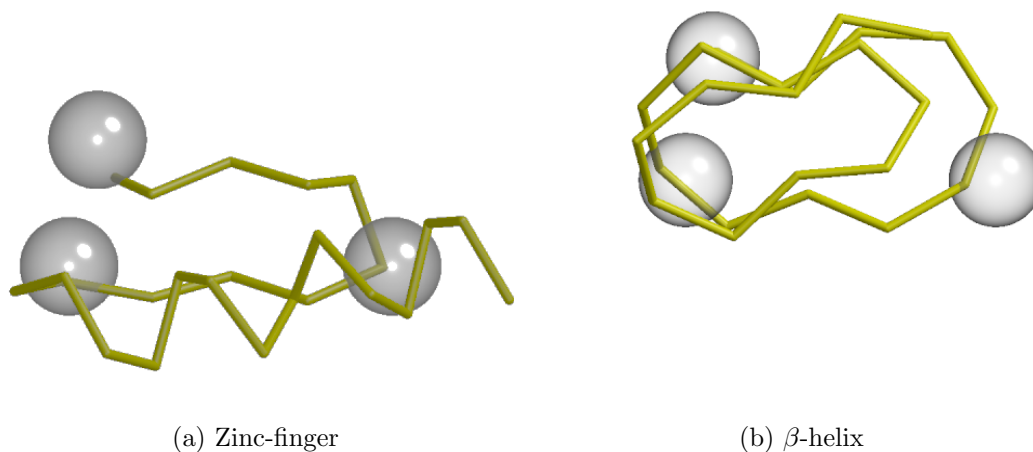


Figure 3.4: Competing minima in the effective free-energy landscape of the system with the substrate for the zinc-finger.

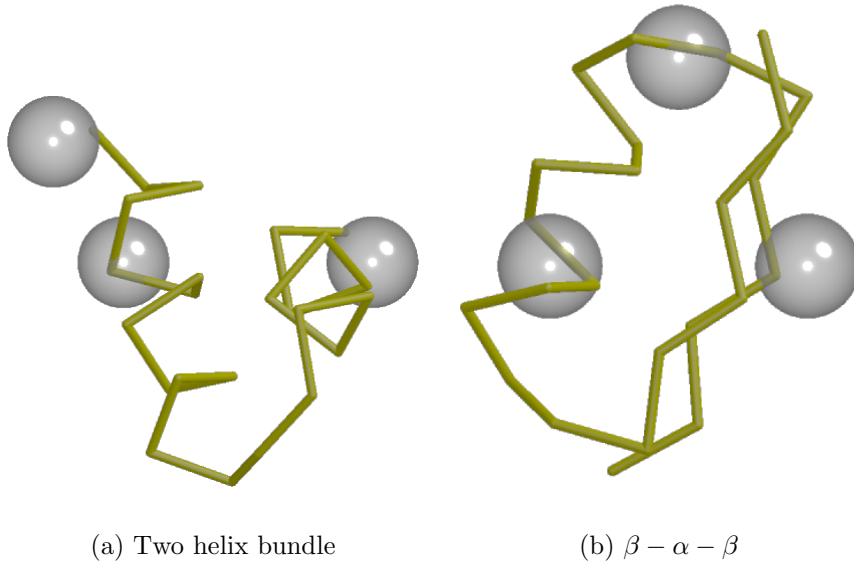


Figure 3.5: Ground states of the system with the substrates designed for the two helix bundle and the $\beta - \alpha - \beta$ folds respectively, shown in table 3.3.

Nevertheless the α -helix is always the resulting ground state structure in the case of totally unbiased interactions.

We show the results obtained with several simulations performed using the Monte Carlo simulated annealing technique on the polymer system with the substrate modeled for the zinc-finger (table 3.3).

Pictures 3.6 show that different kinds of folds are in fact compatible with the particular geometry that had been chosen to bind the β -hairpin of the zinc-finger type conformation 3.6, thus leading to several possibilities for the process of binding to the substrate.

The binding pattern – which models the partner of the disordered protein – is shown using transparent spheres, whose radius is equal to the range of their interaction with the corresponding amino acids of the IUP. The structures corresponding to the various local minima have been drawn using Molscrip [40], VMD [63] and Raster3D [41].

If we turn to the system with the substrate designed for the target fold of type $\beta - \alpha - \beta$ as listed in table 3.3, and we perform a thermodynamic analysis by means of a long simulation with the technique of parallel tempering, we are left with a result similar to the previous one. Completely unbiased interactions between the substrate and the polymer let the α -helix fit the geometry of the pattern, exploiting the three contacts to gain a considerable

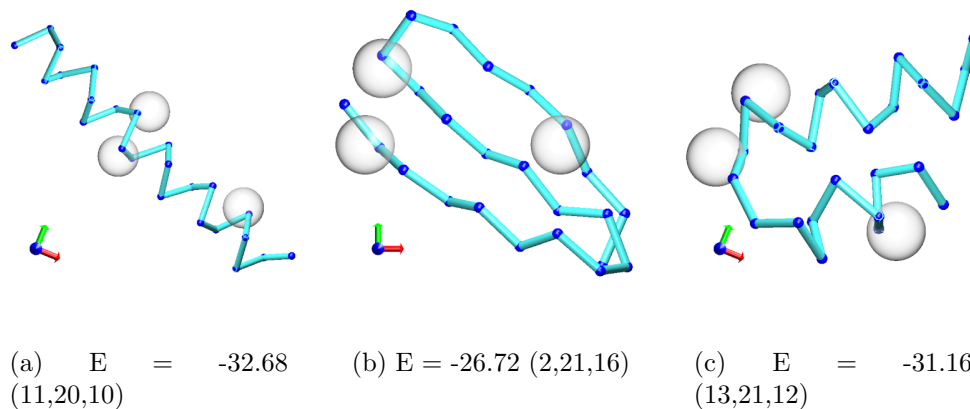


Figure 3.6: Local minima explored with several simulated annealings of the system with the substrate designed for the zinc-finger as of table 3.3: the ground state is the α -helix (identity of polymer beads bound to the substrate points – in the same ordering as of table 3.3 are reported within parenthesis).

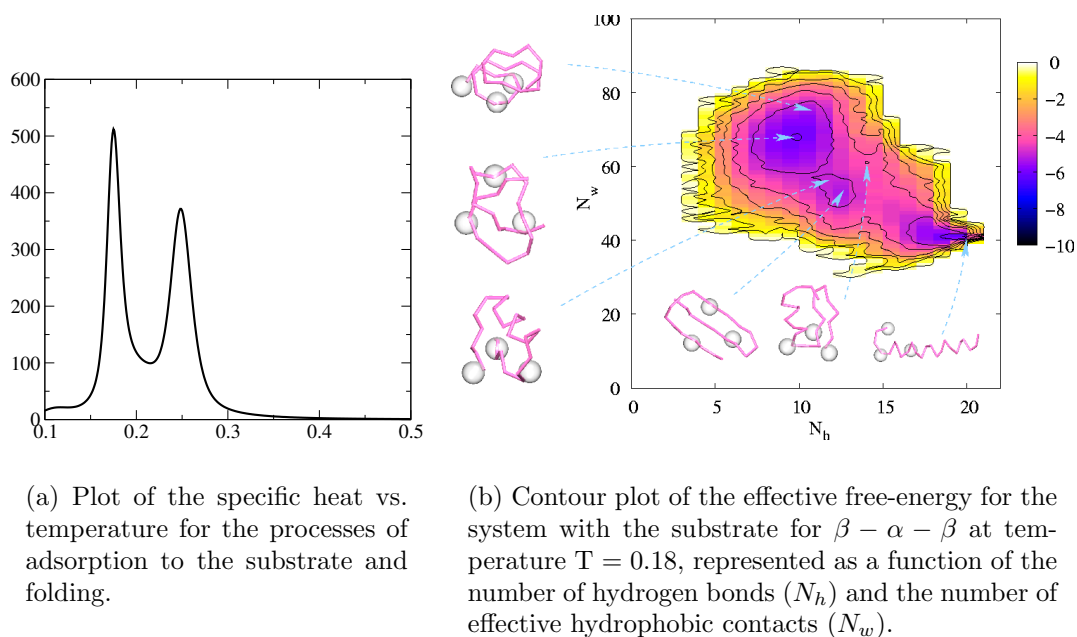


Figure 3.7: Specific heat and contour plot of effective free energy for the polymer in the presence of the substrate designed for the fold $\beta - \alpha - \beta$ slightly above the folding temperature.

amount of energy and thus win the competition among local minima. The specific heat plot of the system shows two peaks. The one at higher temperature corresponds to the processes of binding, that is adsorption of the polymer to the substrate; the second one marks the folding transition into the ground state, which is the α -helix in this case.

3.5 Enhancing specificity of effective contacts

Totally unbiased interactions, between the polymeric chain and the substrate mimicking the partner of the modeled interaction typical of disordered proteins, leave too much freedom to the molecule. It may adopt in this way several competing configurations. Then we make one step further, in order to reduce the space of bound configurations.

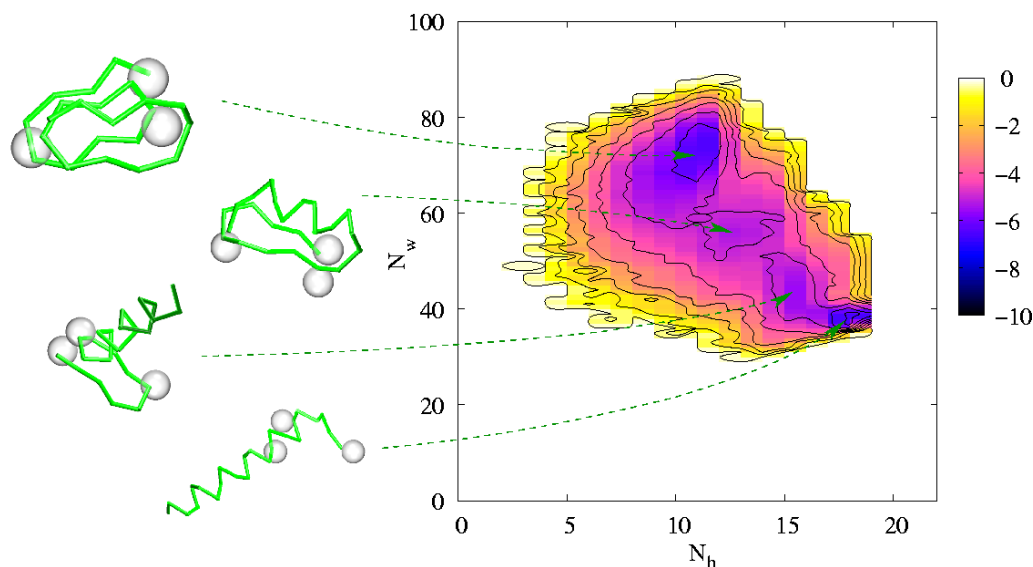


Figure 3.8: Contour plot of the effective free energy for the system interacting with the substrate of the zinc finger at a temperature below the folding, with partially unbiased interactions.

In order to model a more specific kind of interaction between the disordered protein and its functional partner, we proceed with the selection of a subset of atoms along the polymer. These ones will let the polymer gain energy upon binding.

This is actually more adherent to the biological event of binding and folding occurring in known cases of mutual influence between unstructured

proteins and their partners. Specificity is a key ingredient of the way the process occurs, realized through electrostatic interactions and hydrogen bonding. It can be accomplished only between selected atoms and residues: between charged and polar residues, which turn out to be more abundant in unstructured proteins (see chapter 1). In this way the homopolymer model becomes a heteropolymer one.

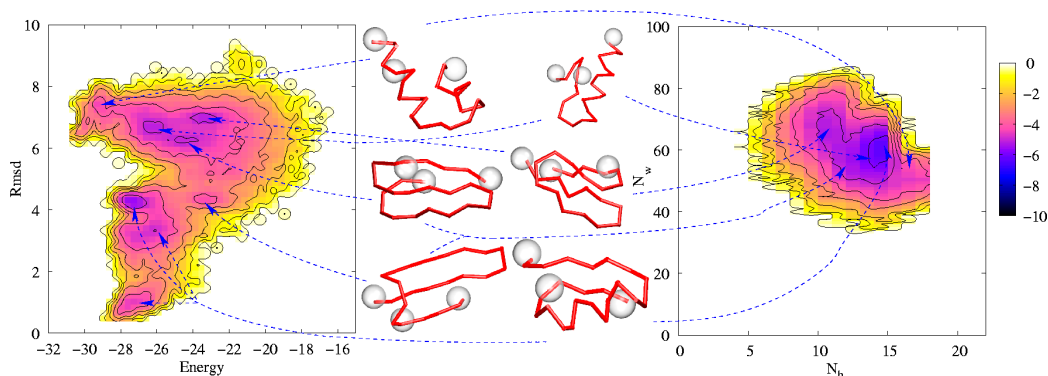


Figure 3.9: Contour plots of the effective free energy vs. energy and rmsd against $\beta - \alpha - \beta$ (left) and vs. the number of hydrogen bonds and effective hydrophobic contacts for the substrate of the two helix bundle partially unbiased.

The number of elements within the subset of the polymer's beads is chosen to be equal to the number of contact centres on the substrate, although this can be varied arbitrarily. This choice has been made just for simplicity, keeping in mind the purpose to go further to very specific one-to-one interactions.

The contour plot of figure 3.8 plots the free energy as a function of the total number of hydrogen bonds (N_h) and the number of effective hydrophobic contacts (N_w), in the case of the substrate designed for the zinc-finger. There is an evident decrease in the number of local minima, with respect to the case of the totally unbiased interactions, shown in the previous section.

This is of course what one expects: by specifying the correspondence between the interaction centres of the substrate and the beads of the polymer chain, one drastically reduces the entropy of some configurations, which turn out to be depleted from the sampling. At the same time the target fold takes advantage of the diminished entropy of some antagonist fold, thus increasing its stability.

In some cases, the partially unbiased scheme adopted – three selected beads along the chain being capable of a gain in energy by their binding to

the substrate – is already a good approximation to the design problem faced throughout this work. However the variables chosen to show the free energy plot of figure 3.8 are sometimes misleading, since they may hide several types of conformations, within the same local minimum of the free energy surface.

In fact, several different configurations may fit a similar number of hydrogen bonds and hydrophobic contacts: a deeper investigation on the variety of structures hidden within a free energy contour plot like the one of figure 3.8 is thus needed.

This has been done by computing the root mean square deviation (RMSD) of the configurations visited during the simulations against some reference structures. In general two folds are enough. They were selected carefully for each pattern geometry: the first choice is of course the target fold that we wish to obtain through the folding, the second one an antagonist structure.

The RMSD has been calculated using the algorithm of Kabsch [64, 65], which is by far the most used when referring only to the backbone. In the present case, the level of coarse-graining applied makes easy the comparison of the three-dimensional structures by letting the RMSD become a good measure for the resemblance of two conformations.

As model system to treat, we discuss the results of the simulations performed under the parallel tempering scheme for the pattern geometrically arranged for a two helix bundle. We do not keep now the bias of polymer's atoms towards a specific interaction centre of the substrate, leaving the possibility of a global reshuffling of the three contacts listed in table 3.3.

The contour plots of figure 3.9 show the effective free energy as a function of the energy and the RMSD against the structure of the $\beta - \alpha - \beta$ (plot on the left): this is one of the antagonist folds of the two helix bundle, that were encountered using this substrate. On the right side, the effective free energy is plotted in the space of the total number of hydrogen bonds and the number of effective hydrophobic contacts, as usual.

The left plot reveals a wealth of local minima which is not captured by the standard plot used until now, showing several different structures visited during the simulation at temperature $T = 0.17$, which lies below the peak of the specific heat for that system, found at $T = 0.173$.

The striking result one can see from figure 3.9 is that folding occurs favouring the $\beta - \alpha - \beta$ conformation despite the two helix bundle, without a one-to-one bias of the effective interaction. This is due to the higher entropy of the $\beta - \alpha - \beta$ in comparison to the two helices, as one can immediately guess by looking at the energies of the two structures in table 3.2.

The target fold turns out to be the ground state at zero temperature, and the free-energy landscape has been depleted by the several competing local minima typical of the rugged landscape of the homopolymer with unbiased

interactions with a three-beads substrate.

Unfortunately, the number of allowed configurations, compatible with the reduced volume of the conformational space visited by the polymer during a relatively long MC simulation ¹, is still enough to produce a low folding temperature, compared to the one observed in the polymer free to move in the bulk without any contact bias.

We are thus left with a first preliminary conclusion to this problem of design, performed through effective interactions, mimicking partner recognition and binding of the intrinsically unstructured proteins modeled insofar. Specific bias on the interactions occurring among polymer's and partner's atoms is not strictly necessary to select only a subset of the presculpted folds of the homopolymer. Still, in order to obtain folding to the specific target, a strict bias is nonetheless essential, to prevent antagonist element from prevailing, due to entropic reasons which cannot be easily controlled, at least within such a simplistic scheme.

3.6 One-to-one interactions

Given the difficulties of the simple scheme adopted here, explained in the previous section, we introduce here a higher degree of specificity in the effective interaction between the polymer and the substrate.

This has been accomplished by letting each interaction centre of the pattern biasing the binding of one and only one atom of the homopolymer. Thus we have a precise correspondence between the polymer and the pattern, ruled not only by the geometry of the substrate, but by a sequence encoded by the substrate's centres as well.

If we ask ourselves to what extent is this bias realistic – as far as our aim is to model the biology of binding and folding of IUPs – we have to keep in mind that neglecting sequence is a lack of the present scheme. This severely limits the capability of predicting the real behaviour of the molecule, as already experienced with the unbiased and the relatively unbiased patterns used until now.

On the contrary, it is well known [4, 1] that the sequence of the disordered proteins encodes “disorder” to some extent, for instance with a relative abundance of charged and polar residues with respect to the average of known globular proteins [29]. This could be a major feature when the unstructured protein approaches a molecular partner by binding to specific sites, guided by chemical affinity.

¹ $N_{moves} \approx 10^8$ sweeps. A “sweep” is an amount of elementary Monte Carlo moves equal to polymer's length.

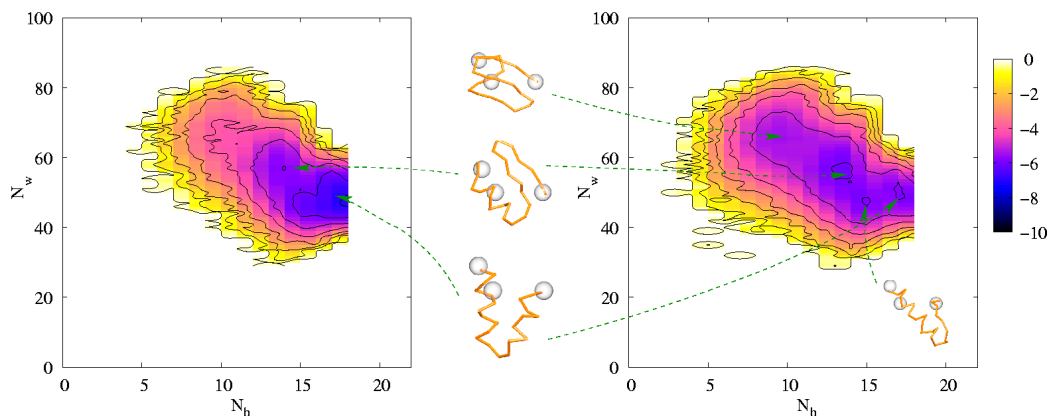
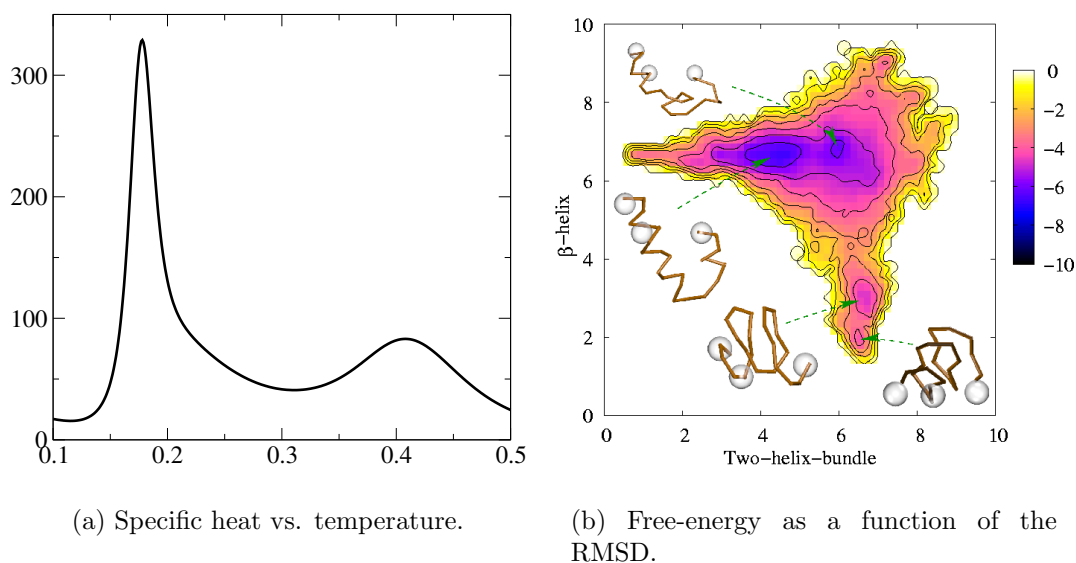


Figure 3.10: Two helix bundle substrate fitting the target structure in the effective free-energy versus hydrogen bonds and effective hydrophobic contacts at 95 % of the folding temperature (left) and slightly above it (left).



(a) Specific heat vs. temperature.

(b) Free-energy as a function of the RMSD.

Figure 3.11: Specific heat and contour plot of effective free energy for the two-helix bundle at 98 % of the folding temperature: the root mean square deviation (RMSD) against one representative fold chosen for the two helix bundle and one for the β -helix, among those visited during the simulation

By inspection on the case of the system with the substrate for the two helix bundle 3.11, now using the specific bias shown in table 3.3, we were able to suppress the occurrence of antagonist folds of type $\beta - \alpha - \beta$. Moreover, the folding to the target occurs at a temperature higher than in the case shown in the previous section, where the interaction was not yet so specific.

Now that we put the one-to-one bias for the interaction between polymer and substrate, the relative entropy of the two-helix-bundle is increased, gaining it from the absence of the former antagonist: the peak in the specific heat related to folding is now at $T = 0.178$.

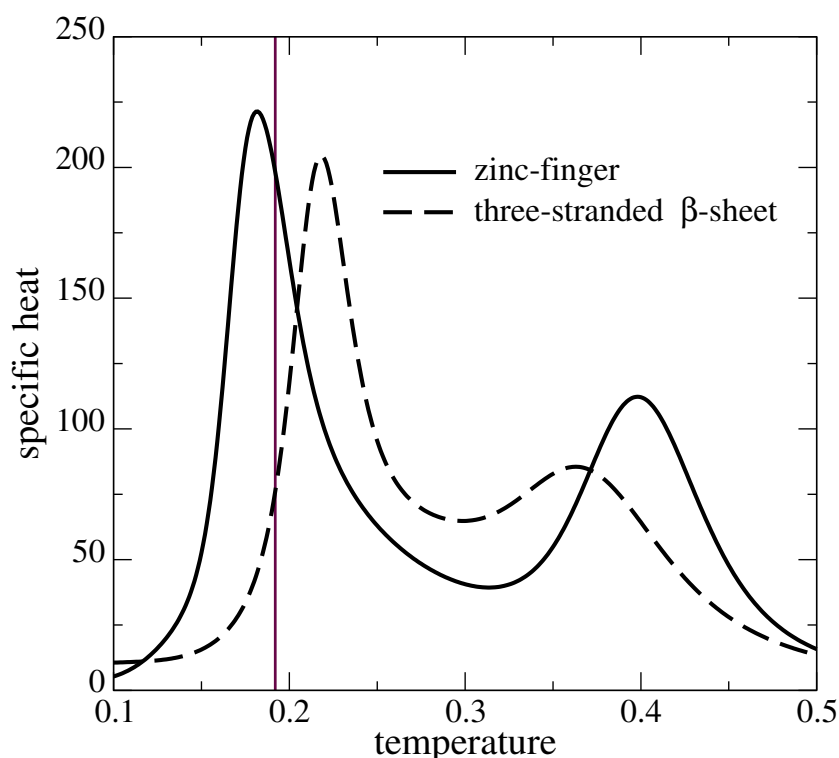
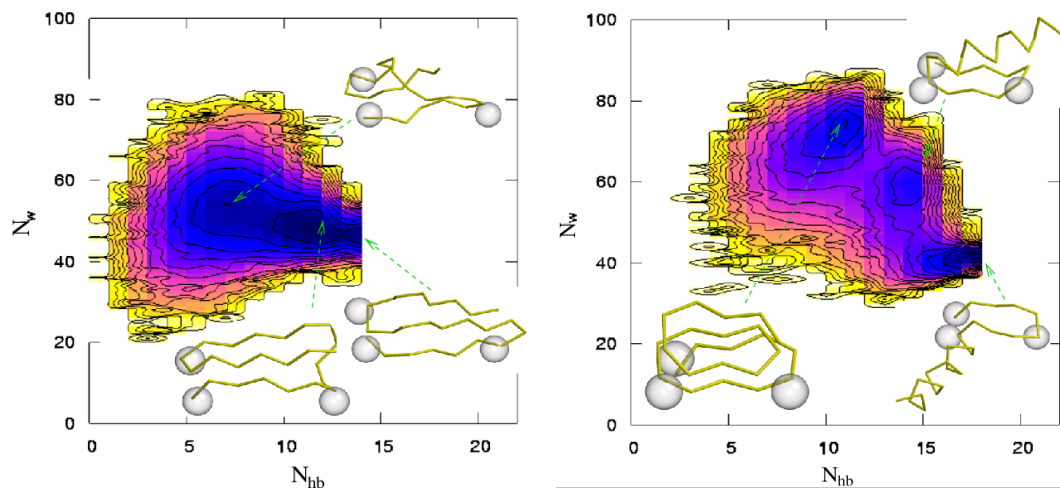


Figure 3.12: Comparison of the temperature dependence of the specific heat for the process of binding and folding of the homopolymer in the presence of the pattern designed for the three-stranded β -sheet and the zinc-finger. The vertical line marks the folding transition temperature of the chain in the absence of the substrate.

Plots of the specific heat as a function of temperature (Fig. 3.11(a)) show two peaks unlike the single peak observed in the absence of the binding. The higher temperature peak is associated with the adsorption transition of the polymer chain to the bottom wall containing the binding pattern while

that at the lower temperature marks the folding transition into the target structure.



(a) Free energy landscape of the three-stranded β -sheet at the folding transition ($T_f^\beta = 0.218$).

(b) Free energy landscape of the zinc-finger at the folding transition ($T_f^z = 0.182$).

Figure 3.13: Effective free energy as a function of the total number of hydrogen bonds and hydrophobic contacts.

Also in the case of the zinc-finger and the three-stranded β -sheet, the plot of the specific heat shows the same behaviour (3.12). The relative positions of the peaks in the different cases can be understood as follows. Different binding patterns have different entropies for the disordered adsorbed state which is populated at intermediate temperatures. This entropy is lower for the three-stranded β -sheet (the residues selected for the specific interaction with the pattern are more spread out along the chain than in the zinc-finger and the conformational freedom is more restricted as a result) so that the adsorption (folding) transition ought to take place at a temperature lower (higher) than for the case of the zinc-finger on ignoring the free energy difference between the two folded states and assuming that the average energy of the disordered state is not affected by the nature of the pattern.

In the two different free energy landscapes of fig. 3.13 the logarithm of the normalized histograms is plotted as a function of the total number of hydrogen bonds (N_{hb}) and of the effective hydrophobic contact (N_w). Data points have been collected from equilibrium Monte Carlo simulations at the temperatures characteristic of the folding transitions for the two systems:

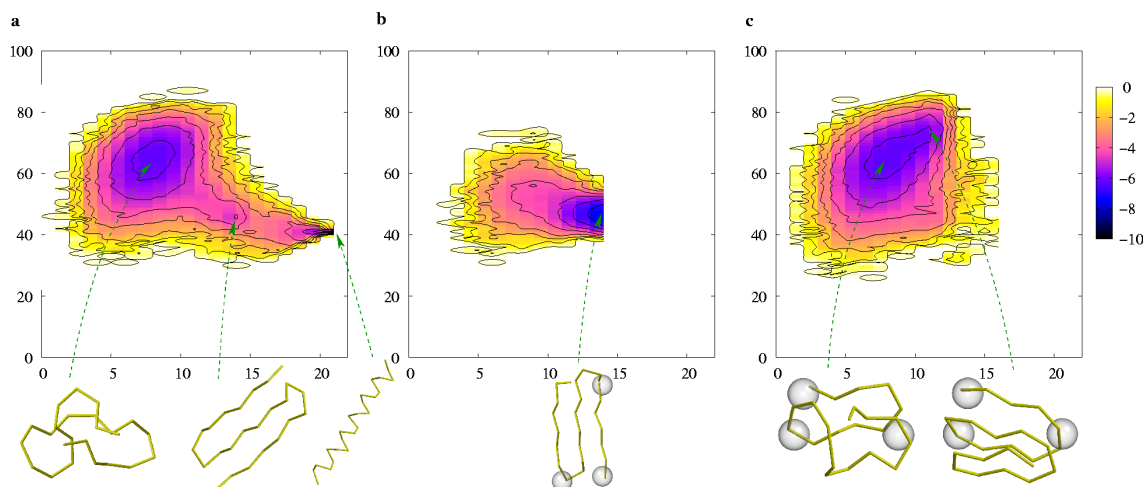


Figure 3.14: Contour plot of the effective free energy at temperature $T = 0.2$ for the three different cases; no pattern (left), three stranded β -sheet pattern (center), zinc-finger pattern (right).

$$T_f^\beta = 0.218, T_f^z = 0.182.$$

The relatively big difference in the two folding temperatures arises from the manifestly different behaviour of the adsorbed phase. This is due, of course, to the difference in the geometrical arrangement of the substrate adopted. The reasons for the differences in the effective free energy landscape experienced by the polymer may be investigated further by the analysis of the landscape at a temperature intermediate between the two mentioned above.

Fig. 3.15 underscores the role of the binding pattern in shaping the energy landscape of the IUP. Contour plots of the effective free energy as a function of the total number of hydrogen bonds and the total number of hydrophobic contacts are compared at the same temperature, $T = 0.2$, for three different cases. No binding pattern is present on the left, then the three-stranded β -sheet pattern is in the centre, and the zinc-finger pattern on the right.

In the first case the chosen temperature is slightly above the folding transition (see Fig. 3.12). Note that even in the absence of the binding pattern the IUP is kept confined within the cubic box, and the denatured disordered state is the most populated one.

The ground state (single α -helix) is populated as well, and the three-stranded β -sheet conformation appears as a competitive local minimum as it is the case for a completely free chain [8].

In the three-stranded case, $T = 0.2$ is quite below the folding transition

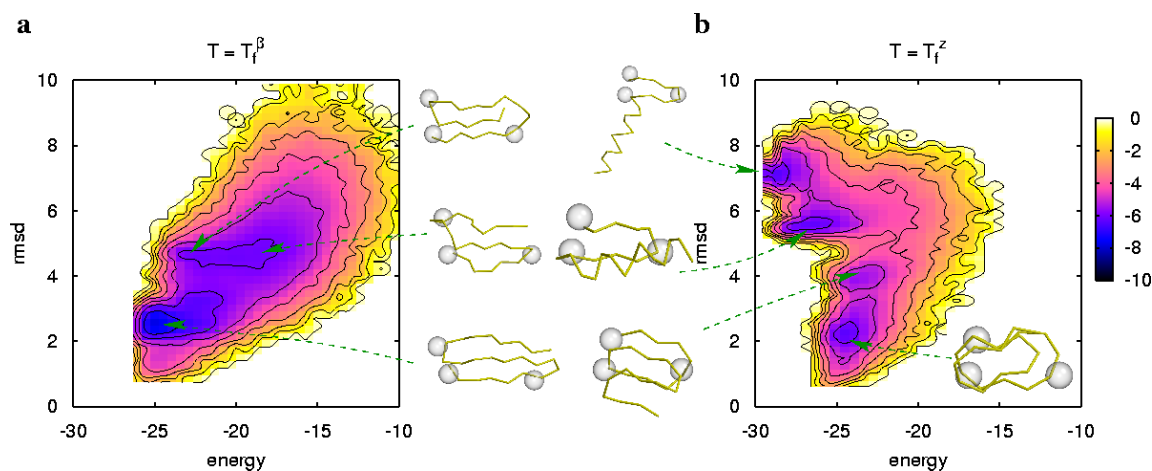


Figure 3.15: Contour plot of the effective free energy at the folding temperature for the system with the substrates designed for the three-stranded β -sheet (left) and for the zinc-finger (right). The RMSD of the latter is computed against one of the antagonist structures resembling a β -helix.

temperature so that only the free energy minimum corresponding to the three-stranded β -sheet target conformation is populated. The single helix conformation is entirely absent, because it is incompatible with the chosen binding pattern.

The free energy of the two different low temperature ordered states, the three-stranded conformation in the presence of the pattern and single helix in its absence, is similar. The folding transition temperature is increased in the presence of the pattern, since the entropy of the disordered state is reduced by adsorption to the binding substrate with respect to the bulk case, whereas the intra-chain energy of the disordered state is not significantly influenced by the presence of the pattern. We also found a competing conformation similar to a greek-key motif (fig. 3.3(b)), which fits equally well the contact pattern. However, the three-stranded target fold is the global free energy minimum at the folding temperature.

In the case of the zinc finger the situation is more complex. The free energy of the ordered target state, the zinc-finger like conformation, is lower than that of a single helix or a three-stranded conformation. In fact, the zinc-finger is not populated in the absence of the pattern, and the folding transition temperature ($T_f^z = 0.182$) is relatively well below $T = 0.2$.

As a consequence, at the latter temperature the denatured disordered state is mostly populated. Note that the properties of this state are not

affected by the presence of the pattern for the coordinates used in the contour plot.

Again, the single helix conformation is not compatible with the binding pattern. The energy landscape might be more complex also because the binding pattern could be less specific for the zinc-finger than for other kind structures chosen among the fold menu of the homopolymer model: other pre-sculpted minima different from zinc-finger are observed and compete entropically with it.

At intermediate temperatures such as $T = 0.2$, a β -helix like conformation is actually the only one competing with the denatured state, as it is shown in the contour plot. At the transition temperature the target zinc-finger conformation is observed, along with a similar conformation in which the helix is detached from the hairpin (data not shown).

One would then expect that the introduction of heterogeneity in the model would serve to increase the specificity of the binding pattern towards the target conformation.

Chapter 4

Gaussian models for protein function

Several studies in the past decades have shown the validity of the normal modes approach to extract useful information on the large-scale functional movements of proteins near their native state conformation [66, 67, 68, 69, 70].

Molecular dynamics simulations of biomolecules performed using detailed all atoms potentials provides a wealth of information on chemical reactions, but part of the insight resides in the time scale of the large amplitude, concerted displacements of atoms [71]. However, some aspects of these motions can be obtained within simple coarse-grained approaches and to a limited extent from the analysis of the Hessian matrix: in fact, it has been shown that low-frequency motions provide the major part of the norm for those global motions, whereas the fastest ones account for spatially localized fluctuations [72, 73].

Dynamical trajectories of the atoms in the molecules can be decomposed along a orthogonal set of eigenvectors of the covariance matrix [74]; thus one is brought to interpret the functional, large amplitude motions of biological relevance for proteins as superpositions of the principal motions of a network of atoms.

A pioneering work developed by Tirion [75] paved the way for extremely simplified Normal Mode Analysis (NMA): detailed harmonic potentials are replaced by a single-parameter, spring-like potential between atoms found to be in contact in the native configuration.

Despite the extreme simplicity of this approach, the good agreement obtained with atomic mean square displacements and vibrational spectra of molecular dynamics simulations [75] opened the possibility for further studies, within the same approach [76, 77, 78, 79, 80, 81, 82, 83]: a good level

of consistency with more accurate analyses is achieved even treating proteins under coarse-grained schemes, as recently shown by Gaussian Network Models [76, 77, 78, 79, 80, 81], simple yet useful models that explore the collective motions of proteins, profitably adopted in a variety of biological contexts [84, 85].

In the present study two structures recently solved [86] are addressed, which belong to the family of truncated hemoglobins (trHbs), small heme proteins widely distributed in bacteria, protozoa and plants, forming a distinct group within the hemoglobin super-family [87, 88, 89].

Though having a simpler structure than the traditional globin fold, they still preserve the respiratory function, providing transport and storage of oxygen molecules. Furthermore they have been proposed to be involved also in other biological functions, such as protection against reactive nitrogen species, photosynthesis or to act as terminal oxidases [90, 91, 92, 87].

The low complexity of trHbs structure, compared to normal globin folds, might help the comprehension of the mechanisms used by these shorter molecules to bind small ligands to the heme iron atom (e.g.: O_2 , their main target, and CO , to which heme has a high affinity).

In particular, the presence of an apolar cavity system extending throughout the protein matrix of truncated hemoglobin from *Mycobacterium Tuberculosis* and homologous structures has been recently noticed [93, 89]: this tunnel connects the heme distal pocket to the protein surface, and may thus allow an efficient diffusion path for oxygen and other small molecules to the iron atom (fig.4.1).

The role of protein cavities has been deeply investigated in myoglobin (see [94, 95, 89] and references therein), both theoretically using computer simulations and experimentally suggesting pathways for ligands migration switched by a small number of substates, which can be allosterically converted to the stable conformations [96].

These issues are investigated here from a novel point of view, through a simple coarse-grained scheme in the spirit of the Gaussian chain models, with a twofold goal: understanding the mechanical processes involved in the functional movements of these key proteins and taking advantage of this new Gaussian framework, computationally fast and conceptually simple.

4.1 Structural characterization

The structures addressed in the present study are two truncated hemoglobins from the ciliated protozoan *Paramecium Caudatum* (PtrHb, pdb id: 1dlw) and the green unicellular alga *Chlamydomonas Eugametos* (CtrHb, pdb id:

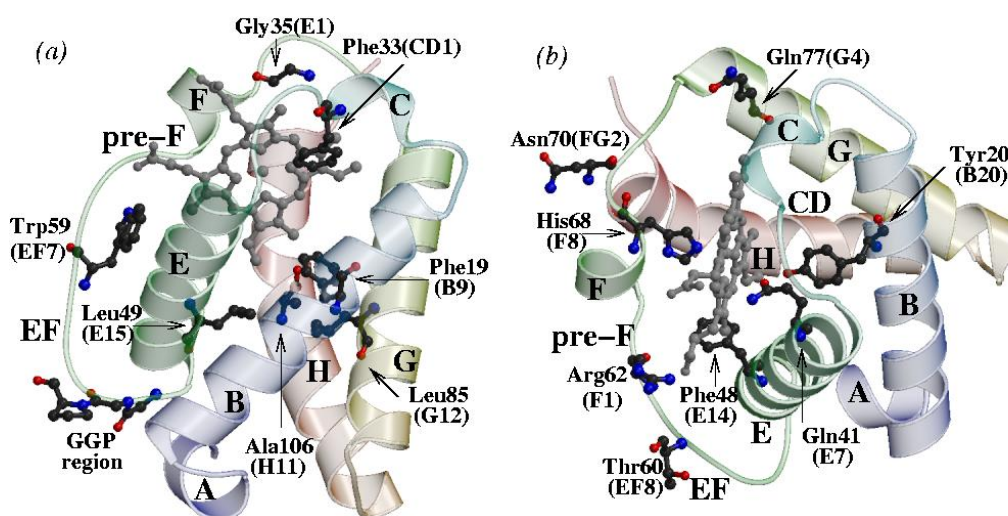


Figure 4.1: Truncated hemoglobin fold from *Paramecium Caudatum*: helices, coils and the main residues described in the text are labeled according to the standard nomenclature for globins. The two-over-two helical structure enclosing the heme group is clearly visible: (a) side view, (b) top view. Figure drawn using Molscript [40] and Raster3d [41].

1dly), solved at 1.54 and 1.80 Å resolution respectively, by Pesce et al. [86].

Similarly to the other proteins belonging to the trHb family, they display low sequence identity with hemoglobins from vertebrate and non vertebrate. This is smaller than 15% for PtrHb and CtrHb, due to substantial residue deletions at either N- or C- termini and in the C and D helical region of the globin fold [86].

More than 70% of the residues in the two structures belongs to helices, mainly of type α (above 67% in both proteins: only the short helix C is of type 3_{10}): this is a typical feature of the globin fold, which leads to guess a primary role of helices in the functional motions of these proteins, as well as in myoglobin and hemoglobin. Nonetheless several structural differences make truncated hemoglobins fall in a distinct group in the hemoglobin superfamily [89, 86].

Helices in the globin fold are traditionally indexed through capital letters A, B, C, D, E, F, G and H, while loops between them are named according to the nearby helices, and residues are numbered sequentially with each unit [97].

The structures taken in consideration here reveal the so called “two over two α helical sandwich” (fig. 4.1), in place of the classical “three over three”

observed in the globin fold [46]: in fact helix D is absent, while N-terminal A helix and proximal F helix are drastically reduced to only one turn.

A structure-based sequence alignment of PtrHb, CtrHb and other truncated hemoglobins with sperm whale Myoglobin, reported in [86], shows the strongly conserved residues among proteins in the trHbs family, mainly of three types:

glycine rich motifs especially at helices termini, which enhance structural flexibility (Gly-Gly motifs at the beginning of the AB and EF regions, and Gly-Arg/Lys in the pre-F region [86]);

hydrophobic residues on heme distal and proximal sides, which play the main role of shielding heme from solvent molecules, in order to prevent iron oxidation ;

heme binding residues stabilizing the porphyrin ring in the heme pocket; one particularly relevant is the proximal histidine, His 68, localized on helix F.

Strongly conserved residues on the distal side responsible for the shielding of the heme pocket from the solvent are mainly localized on helices B and E, as well as in the CD and EF loops: hydrophobic residues Phe A12, B9, CD1, E14 and Trp EF7, with their side chains pointing to the inner part of the molecule; Tyr B10, Gln E7, with side chains responsible for the stabilization of the ligand bound to heme [86, 89].

The hydrophobic residues identified in [86, 93] as the ones defining a cavity inside the molecule, linking the solvent exposed surface of the proteins to the heme group are positioned on the distal side. They are mainly localized on helices A (at the opening of the tunnel on the surface), B, E (limiting the distal side) and G.

On the proximal side of the heme pocket one finds the proximal histidine, in a strongly conserved position within hemoglobin (Hb) and trHb families: the imidazole ring of histidine allows it to act as either a proton donor or acceptor at physiological pH. In hemoglobins is essential its ability to buffer the H⁺ ions from carbonic acid ionization in red blood cells, allowing the molecule to exchange O₂ and CO₂ respectively at the tissues and at the lungs [98].

It will be shown how the small α helix F, which contains the proximal histidine F8, can play a leading role as a reference position for elucidating the functional motions of the protein regions around the heme pocket.

4.2 Theory

The model adopted in this study is the Beta Gaussian Model (β GM) presented in [84]. It is an extension of the Anisotropic Gaussian Model (ANM) [76, 77, 78, 79, 80, 81], a single parameter model apt to describe small amplitude fluctuations of residues around their native-state equilibrium.

Only alpha and beta carbon atoms (C^α , C^β) are treated: rather than the actual C^β , the latter is an effective centroid accounting for the directionality of the side chain, built for all residues but glycines and terminal ones; its position is determined by the coordinates of neighbouring α carbons [99, 84], according to the following relation:

$$\mathbf{r}_i^\beta = \mathbf{r}_i^\alpha + l_0 \frac{2\mathbf{r}_i^\alpha - \mathbf{r}_{i+1}^\alpha - \mathbf{r}_{i-1}^\alpha}{|2\mathbf{r}_i^\alpha - \mathbf{r}_{i+1}^\alpha - \mathbf{r}_{i-1}^\alpha|} \quad (4.1)$$

where $l_0 = 3 \text{ \AA}$, the vectors \mathbf{r}_i^α and \mathbf{r}_i^β hold the native coordinates of the α carbon atom and of the effective β centroid which belong to the i th residue. Expanding the displacement of the C^β from the equilibrium to leading order in the displacements of the C^α s one gets:

$$\delta\mathbf{r}_i^\beta \sim l_0 \frac{2\delta\mathbf{r}_i^\alpha - \delta\mathbf{r}_{i+1}^\alpha - \delta\mathbf{r}_{i-1}^\alpha}{|2\mathbf{r}_i^\alpha - \mathbf{r}_{i+1}^\alpha - \mathbf{r}_{i-1}^\alpha|} \quad (4.2)$$

The Hamiltonian of the system depends quadratically on the deviations of the C^α and C^β from their native positions, assumed to be the energy minimum in the configuration space (thus neglecting crystal effects on X-ray structures): the displacements of protein's atoms from the equilibrium position are supposed to be small enough to justify this approximation [66, 79, 74].

The Hamiltonian includes interactions between α and β carbons lying within a cut-off distance r_c , above which no pairwise interaction is allowed, as well as an effective interaction accounting for the strength of the peptide bond for nearest-neighbouring C^α s:

$$\mathcal{H} = \mathcal{H}^{peptide} + \mathcal{H}^{\alpha\alpha} + \mathcal{H}^{\alpha\beta} + \mathcal{H}^{\beta\beta} \quad (4.3)$$

where

$$\begin{aligned} \mathcal{H}^{peptide} &= \frac{\gamma_p}{2} \sum_i \sum_{\mu,\nu} \mathcal{M}_{i,i+1}^{\mu\nu}(\alpha, \alpha) \delta r_{i,\mu}^\alpha \delta r_{i+1,\nu}^\alpha \\ \mathcal{H}^{xy} &= \frac{\gamma_{xy}}{2} \left(1 - \frac{\delta_{xy}}{2}\right) \sum_{i,j} \sum_{\mu,\nu} \mathcal{M}_{ij}^{\mu\nu}(x, y) \delta r_{i,\mu}^x \delta r_{j,\nu}^y \end{aligned} \quad (4.4)$$

γ_p is the elastic constant accounting for the relative strength of the effective peptidic interaction between nearest-neighbouring α carbons;

γ_{xy} is the elastic constant for the contact interaction between carbon atoms of type x and y ($x, y \in \{\alpha, \beta\}$);

δ_{xy} is Kronecker delta to avoid double counting of the interactions between atoms of the same type;

$\delta \mathbf{r}_i^x$ is the displacement from the native position of the carbon atom of type x that belongs to the i th residue (μ and ν are the indexes of the Cartesian components);

$\mathcal{M}_{ij}(x, y)$ ($i \neq j$) is a (3×3) matrix, the off-diagonal super-element of the hessian matrix for the interaction between atoms of type x and y which belong to residues i and j :

$$\mathcal{M}_{ij}^{\mu\nu}(x, y) = \Gamma_{ij}^{xy} \frac{r_{ij,\mu}^{xy} r_{ij,\nu}^{xy}}{\mathbf{r}_{ij}^{xy} \cdot \mathbf{r}_{ij}^{xy}} \quad (4.5)$$

where Γ_{ij}^{xy} ($i \neq j$) is equal to 1 if the native separation of the corresponding atoms lies below the cut-off radius r_c , 0 otherwise; $\mathbf{r}_{ij}^{xy} = \mathbf{r}_i^x - \mathbf{r}_j^y$ is the vector of native separation of atoms of type x and y that belong to residues i and j respectively. Entries of diagonal super-elements are built according to the relation:

$$\mathcal{M}_{ii}^{\mu\nu}(x, y) = - \sum_{j \neq i} \mathcal{M}_{ij}^{\mu\nu}(x, y) \quad (4.6)$$

Since the position of the effective C^β and its displacement from equilibrium are fully determined by α carbons coordinates (equations (4.1) and (4.2)), by substitution of (4.1) and (4.2) in (4.4) one is left with an effective hamiltonian $\tilde{\mathcal{H}}$ which depends quadratically on C^α displacements from native state [84] (the index of atom type will be therefore dropped in the following equations for simplicity):

$$\tilde{\mathcal{H}} = \frac{\gamma}{2} \sum_{ij} \sum_{\mu\nu} \tilde{\mathcal{M}}_{ij}^{\mu\nu} \delta r_{i,\mu} \delta r_{j,\nu} \quad (4.7)$$

where γ_p and γ_{xy} ($x, y \in \{\alpha, \beta\}$) have been incorporated in $\tilde{\mathcal{M}}_{ij}$, expressed in units of the reference elastic constant γ .

Time dependent two-point correlation functions can be calculated within a Langevin dynamics leading to equilibrium with the Boltzmann factor $e^{-\beta \tilde{\mathcal{H}}}$ [77].

In the overdamped regime with the viscous damping factor f , the same for all residues [76], and white noise $\eta_i(t)$, the Langevin equation for our system is [100]:

$$f \frac{d}{dt} \delta r_{i,\mu}(t) + \gamma \sum_{j,\nu} \tilde{\mathcal{M}}_{ij}^{\mu\nu} \delta r_{j,\nu}(t) = \eta_i(t) \quad (4.8)$$

One can easily get from equation (4.8) the time dependence of cross correlations between couples of C^α s (the so-called “reduced” cross-correlations):

$$\langle \delta \mathbf{r}_i(t) \cdot \delta \mathbf{r}_j(0) \rangle = \frac{k_B T}{\gamma} \sum_k \frac{1}{l_k} (\mathbf{a}_{ik} \cdot \mathbf{a}_{jk}) e^{-l_k \frac{t}{\tau}} \quad (4.9)$$

$\tau = \frac{f}{\gamma}$ is the reference relaxation time, corresponding to an overdamped spring of elastic constant γ in a dissipative medium of friction f ; l_k are non zero eigenvalues of $\tilde{\mathcal{M}}$ and \mathbf{a}_k the corresponding eigenvectors.

Theoretical B-factors (measured in \AA^2) are obtained from the diagonal elements of the reduced covariance matrix (i.e. from the mean square fluctuations of C^α s around native-state equilibrium, after thermal equilibrium has been reached), through the relation:

$$B_i = \frac{8\pi^2 kT}{3} \frac{1}{\gamma} \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_i \rangle \quad (4.10)$$

Equation (4.10) will be used to fit the experimental B-factors and get an estimate of the reference elastic constant γ .

4.3 Tuning model parameters

In order to obtain reliable data for the structures under study, we compare theoretical and experimental results using the ranking correlation between the two data sets as a guideline to tune model parameters to their optimal values.

ANM was applied on the structures as well as β GM: in the case of the trHb, the theoretical temperature factors obtained with the β GM showed a higher value of Kendall’s non parametric τ [101](see below) against the experimental ones ($\tau = 0.45$ for ANM with $r_c = 13.0 \text{\AA}$, $\tau = 0.57$ for β GM with $r_c = 7.0 \text{\AA}$, in the case of 1DLW).

ANM works very well for bigger complexes, while for smaller proteins more details are required: the reason for the better agreement obtained by the β GM is to be found in the presence of the β centroids, which considerably increases the number of pairwise interactions and takes into account the directionality of the side chains in the contact map of α carbons.

As a consequence, the β GM needs a lower and more realistic cut-off radius r_c to reproduce experimental B-factors and molecular dynamics data, in comparison to those used by ANM [80], as already remarked in a previous study [84], even with small proteins like trHbs: hence the choice to use the β GM in the present work.

Here in particular, the best agreement between theory and experiment was found using a cut-off of 7.0 Å.

This choice is imposed by the difference in compactness between helical regions and coils, and it is critical in order to keep the density of effective contact interactions at the coarse-grained level comparable to the all atoms one.

Larger cut-offs cause contact density to be overestimated for the helical regions, leading to smaller values of B-factors, with respect to the experiment: the consequence is a marked difference between flexible and solvent exposed parts of the protein, compared to the less flexible and buried parts (fig. 4.2).

A key point was the tuning of γ_p , the ratio between the effective peptide bond and the $\alpha\alpha$ interaction: it accounts for the relative stiffness of the covalent bonds along the backbone as opposed to the weaker contact interactions between C^α pairs.

Summarizing the values for the parameters used in the calculation for both structures, $r_c = 7.0$ Å, $\gamma_p = 2.0$, $\gamma_{\alpha\alpha} = \gamma_{\alpha\beta} = \gamma_{\beta\beta} = 1.0$ (the last ones are in units of γ). The value of the reference elastic constant γ will be determined later, fitting the results of the model with the available experimental data.

4.4 Temperature factors and heme modeling

Truncated hemoglobins are heme proteins, the heme group being the active site of the molecule: there oxygen and carbon oxide bind to the sixth coordination position of the iron atom, which lies at the center of the tetrapyrrole ring and is bound to the imidazole ring of the proximal histidine F8 at the fifth coordination site (His 68, eighth residue of helix F in sperm whale myoglobin and in vertebrate hemoglobins, where nomenclature “F8” comes from [102]).

Figure 4.2 shows the plot of the α carbon atoms B-factors of the X-ray structure of truncated hemoglobin in *Paramecium Caudatum* and their corresponding mean square displacements derived from the β GM: most mobile regions are loops and turns between helices, which on the contrary display smaller fluctuations, in agreement with the results of an NMA study performed on deoxymyoglobin (Mb) [103, 104].

The significance of the correlation between experimental and theoretical

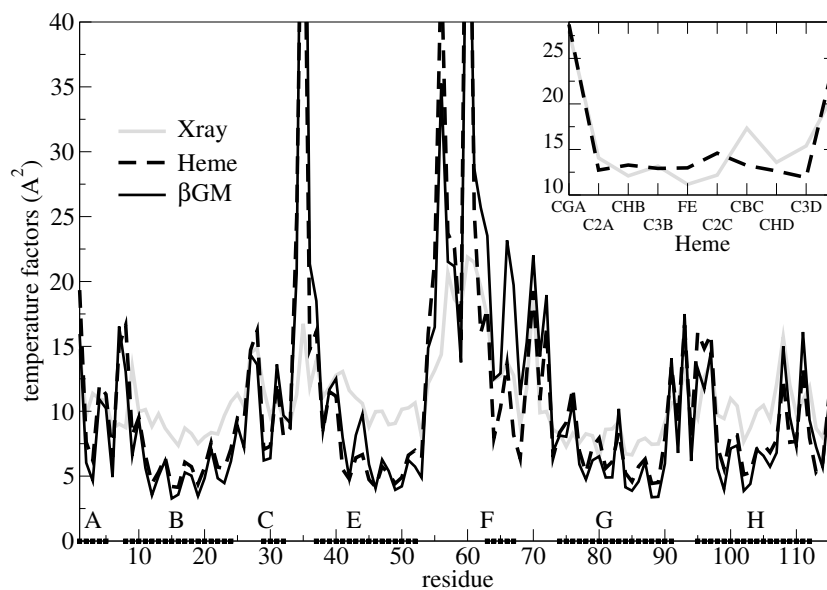


Figure 4.2: Theoretical (black) versus experimental (gray) X-ray B-factors for α carbons in PtrHb, related through equation 4.10. Theoretical B-factors including coarse-grained heme group are shown for comparison (dashed). Helical segments have been marked on residues axis. The inset shows theoretical versus experimental B-factors for coarse-grained heme, with pdb names of iron and carbons included in the coarse-graining.

values is deduced from Kendall's non parametric τ [101]. Since one does not know *a priori* the probability distribution of the experimental B-factors, a significance for the agreement between the two data sets cannot be computed from the value of the linear correlation coefficient. On the other hand, the rank correlation given by τ is independent from the distribution. Kendall's τ for *PtrHb* is 0.57 (0.52 with heme), for *CtrHb* is 0.37 (0.40 with heme), and $P_{null}(\tau) < 10^{-9}$ in all cases ($P_{null}(\tau)$ is the probability for two random sets of data to have a value of τ bigger than the one found between B-factors predicted by the model and calculated from X-ray structure. The number of residues is 116 for *PtrHb* and 121 for *CtrHb*).

The coarse-graining on the heme group includes the iron atom and nine carbons of the porphyrin ring (whose names are reported on the x axis of the inset in fig. 4.2), chosen in order to keep the number of contacts in the modeled system comparable to the number of heme native contacts with nearby residues, thus avoiding to have a loosely connected group as an artifact of the coarse-graining procedure.

Insertion of heme brings only one relevant change to the temperature factors plot (fig. 4.2): helix F has displacements from equilibrium considerably damped, as it was expected, being bound to the iron atom. A reduction in the fluctuations is shown also by the loops between helices C and F, to a lesser extent than in helix F.

The protein part of the reduced covariance matrix obtained including the coarse-grained heme was compared with the covariance matrix computed without modeling the tetrapyrrole ring. The two show a Kendall's parametric correlation $\tau \sim 0.81$ over more than thirteen thousands of points, which stands for a remarkable agreement between them: the coarse-grained heme in fact anticorrelates with the same parts of the protein as helix F, even if more weakly (data not shown). This is not surprising, since the iron atom and the proximal histidine F8 are in direct contact, so the motion of the heme group will be strongly correlated with the one of the F helix, following the proximal side in its deviations from native-state equilibrium; the inclusion of few more atoms under the coarse-grained scheme adopted here do not seem to significantly modify the correlations: further details are needed to extract significant informations on the dynamics of heme. The mechanical response of the protein upon binding of ligands on the iron atom is given by the properties of the network of backbone atoms: thus a good agreement with known behavior of globins may be achieved using gaussian models even without considering heme groups in the coarse-graining procedure [105].

The β GM heme B-factors plot is in substantial agreement with the experimental B-factors for heme (fig. 4.2, smaller plot). In fact the heme pocket is entirely surrounded by non polar residues: one of the main purposes of the

distal region is to screen the heme group from solvent interaction, in order to avoid iron oxidation [102].

The results for Kendall's ranking lie in the typical range of gaussian models [84], even with terminal residues included, and the confidence of the correlation is no doubt statistically significant: still there are some regions of the protein whose fluctuations are not well reproduced by the model, as shown in the plot of temperature factors (fig. 4.2).

The model overestimates interactions between α and β carbons belonging to secondary structures, resulting in local deviations from the density of the all atom picture. Hence displacements of residues belonging to helical regions are underestimated, since these are the most compact parts of the protein and it produces deviations in the profile of B-factors, whose values depend both on the assembly of secondary motifs [106] and on the local packing density [107].

Furthermore, electrostatics and solvent exposure for different residues are not taken into account by the simple approach of the model: electrostatic interactions localized on helices may modify the magnitude of the driving forces producing larger displacements from native state than expected.

POPS program (Parameter OPTimized Surface [108]) has been used to calculate the solvent accessible surface area per residue for PtrHb: the most exposed residues are the ones displaying the greatest average displacements from the native structure, as it was expected (figg. 4.2, 4.5). These small residues (Gly 35, the GGP region – Gly54, Gly55, Pro56 – Thr60, Gly61), located in loops CD, EF and to the pre-F region, allow larger flexibility to the polypeptide chain (glycines especially) and the bigger fluctuations predicted by the model are due to their diminished connectivity as well, being the most exposed to the solvent. This was expected, since the model totally neglects solvent exposure.

The simplified approach used here shows a remarkably better agreement with experiment, for buried regions, where the connectivity of atoms is greater and the solvent plays a minor role.

4.5 Results and Discussion

In order to identify the relevant motions of the protein the reduced covariance matrix plot (figure 4.3) of PtrHb modeled without the heme group is inspected (PtrHb will be the main target of the following discussion, the same considerations holding for CtrHb as well), normalized as follows:

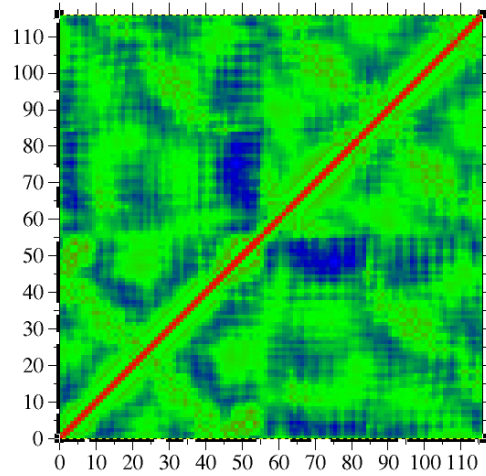


Figure 4.3: Normalized covariance matrix: trivial correlations due to contacts have been put to 0 (green); diagonal elements are equal to 1 (red); anti correlation range from 0 to the minimum value found, for Gln41 (E7) and His68 (F8), lower then -0.35 (blue). Helical regions have been highlighted.

$$c_{ij} = \frac{\langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle}{\sqrt{\langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_i \rangle \langle \delta \mathbf{r}_j \cdot \delta \mathbf{r}_j \rangle}} \quad (4.11)$$

Normalization is generally performed in order to allow a direct comparison between the cross-correlations predicted by the model and the ones obtained in computer simulations, e.g. from molecular dynamics, provided equilibration has been reached [109].

From the reduced normalized covariance matrix one is able to extract non trivial informations on the collective motions of the protein under study: these generally involve the regions of the molecule that show negative correlations.

Indeed it turns out that spatially closed parts of the molecule, i.e. residues in contact, undergo motions with positive correlation, as one would expect for contact-driven motions.

One can identify three main blocks in the covariance plot (fig. 4.3): the first one contains helices A, B, C, E, the loops between them and the EF loop; the second one includes clearly the preF-loop, heme bound helix F, as well as the first part of helix G, while the third block hosts the major part of helix G and the C-terminal side of helix H.

Most residues in the first block, especially the ones belonging to helical

regions A, B and E (distal side), show a remarkable anti-correlation with residues localized at the beginning of the second block, belonging to the proximal helix F and to helix G; in the third block the last turns of helix H is bent at the C-terminal to allow closer contacts with heme [86].

This division in domain of motions is similar to the one found in [103] for deoxymyoglobin, provided that one notes the effect of the bending of C-terminal side in helix H, which implies a correlated motion with the proximal side, as suggested by fig. 4.4, where normalized correlations between His68 (F8) and the rest of the protein are shown. Here the crucial role of small helix F in the dynamics of the protein is underlined, since it contains the proximal histidine, and the division of the protein in domains of motion as described above is made more evident.

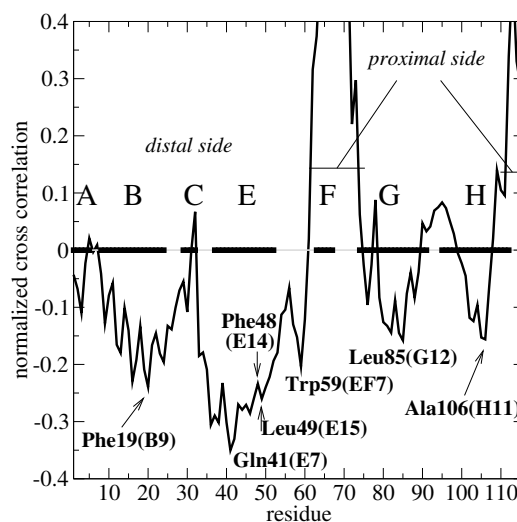


Figure 4.4: Normalized cross correlations between the proximal histidine F8 and the rest of the protein (c_{68j} , hence the peak raising to 1.0 at $j = 68$): residues displaying significant anti-correlations with His68 are labeled in the plot. Like Phe19 (B9), they are strongly conserved throughout the trHb family [86], being relevant to prevent solvent access to the heme pocket (E14, EF7 [86]), to stabilize heme bound ligand (E7 [86]) and to build the gate between the heme pocket and an apolar cavity running inside the protein matrix (G12, H11 [93]).

The covariance minima in the plot of figure 4.4 are particularly meaningful, being found between His F8 and other key residues of the protein. Phe19 (B9), which has a bulky side chain, is responsible for the screening of the distal cavity from the aqueous environment outside the molecule, and

is strongly conserved among trHbs; in a position occupied by the distal histidine in vertebrate Hbs we found Gln41 (E7), hydrogen bonded to Tyr20 (B10), which contribute to stabilize the heme-bound ligand [86] and form a hydrogen bonding network in the heme pocket, which is believed to be responsible for the different ligand rebinding kinetics displayed by PtrHb and CtrHb in comparison with Mbs and Hbs [110].

His68, taken here as a representative to deduce the motion of the whole proximal side from the covariance and correlation plots (figg. 4.3, 4.4), anti-correlates with hydrophobic Phe33 (CD1) as well, another strongly conserved residue among trHbs: together with the previous three residues they line precisely the distal cavity facing the heme group. The anti-correlation of the distal and proximal sides is a clear sign of the concerted motion which may allow the heme pocket to expand, thus making easier access to heme for ligands coming from the apolar cavity that links the inner part of the protein to the solvent [86], escaping the steric hindrance of the distal side residues.

Strong anti-correlation with the proximal histidine are displayed by Leu49 (E15), Leu85 (G12) and Ala/Val106 (H11) as well: these residues lie at the bottom of the distal cavity, at the interface between the tunnel running inside the protein matrix and the heme pocket [93, 111].

The anticorrelated motion of the proximal and distal sides is made more visible by inspection of the components of the eigenvectors corresponding to the first two slowest overdamped modes, plotted in figure 4.5.

Residues displaying the biggest deviations from their native positions are highlighted: they belong to loops between helices lining the heme pocket (CD and EF loops, pre-F region), and to helices enclosing the distal and proximal sides (helix B and E, helix F and H). These modes contribute substantially to the opening and closing of the distal side, in agreement with previous studies on globins [103, 104].

A detailed view of the conformations visited by the first mode is shown in figure 4.6, where the open and closed structures of the distal cavity are displayed, along with distal residues Tyr B10 and Gln E7.

From the covariance plot (fig. 4.3) and the component along the y axis of the second slowest eigenvector of figure 4.5 (although small, due to the normalization, which enhances most mobile regions like loops) one can notice the anticorrelation of the proximal histidine with the residues identified to line the passage leading to the heme pocket from the tunnel inside the protein (mainly Phe19, Leu85 and Ala106, already evidenced in fig. 4.4) [93, 89, 112, 111]. The anti-correlation between the two groups of residues hints at a possible mechanism for the passage of ligands to the heme pocket, through the enlargement of the gate: the presence of the apolar cavity has been proposed to contribute effectively to the fast rebinding of ligands on heme,

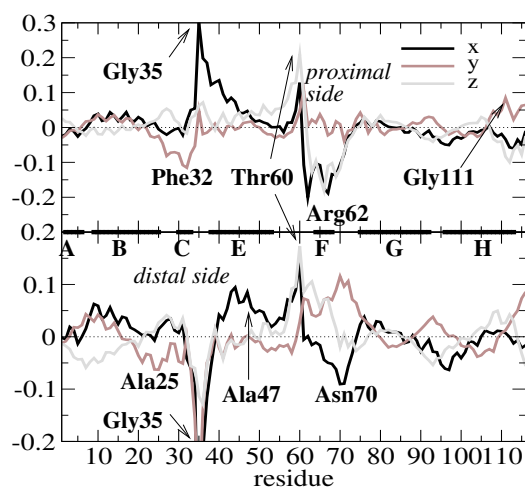


Figure 4.5: Components of normalized eigenvectors for the first two slowest modes of motion (1, top; 2, bottom; ratio of corresponding eigenvalues: 1.16), which bring a similar contribute to the dominant opening mechanism of the distal cavity, driven by the anticorrelated motions of the proximal (pre-F loop, helix F, loop FG and last part of helix H) and distal sides (helix C, CD loop and helix E especially). Residues with bulky side chains, strongly conserved in the family of trHbs and belonging to the hydrophobic cluster preventing solvent access to the heme pocket [86] are spatially located near the residues with biggest components, highlighted in the plot: Phe33 (CD1), Trp59 (EF7), Phe48 (E14). The latter acts as gating residue in trHbN from *Mycobacterium Tuberculosis* [111].

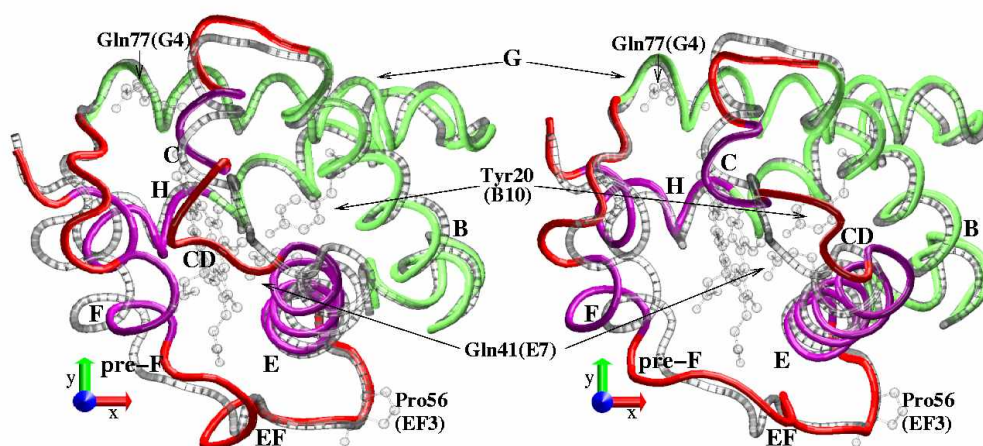


Figure 4.6: Open (left) and closed (right) conformations of the distal cavity, obtained by adding or subtracting the rescaled eigenvector of the first slowest mode to the native positions of α carbons (scaling factor: 20). Most mobile regions in the first mode are coloured in red (loops) and purple (helices). Heme group and native structure are drawn in gray, as well as heme bound ligand stabilizing residues Tyr B10 and Gln E7 and hinges of the distal side opening mechanism – Pro EF3 and Gln G4 (figure drawn using VMD [63] and Raster3d [41]).

together with the hydrogen bonding network in the distal side, as already pointed out [93, 110, 112, 89, 113].

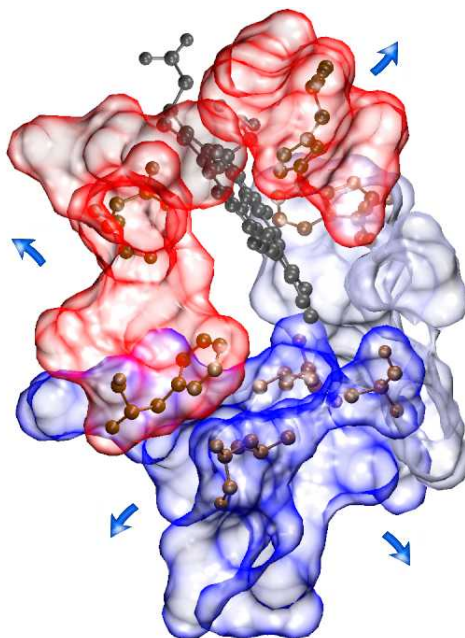


Figure 4.7: Schematic representation of near equilibrium motions of groups of residues delimiting the heme pocket, inferred from covariance analysis: solvent accessible surfaces of residues delimiting the apolar cavity (res. 6, 12, 16, 17, 49, 53, 85, 89, lower left), and the distal cavity (proximal side: res. 64, 68, 71, upper right; distal side: res. 19, 20, 32, 33, 41, upper left) are shown with a 1.4 Å radius probe. Ball-and-stick representation is used for His68 and the residues labeled in figure 4.4. The cluster in the lower right (res. 48, 51, 52, 59, 105, 109) defines a narrower cavity [112]. Figure drawn with VMD [63], rendered with Raster3d [41].

The decay of these essential modes is compatible with a combined motion of the main blocks similar to a pumping mechanism: according to the results obtained in this study, the native state conformation of the two truncated hemoglobins is such that small displacements of the atoms, due to stochastic interactions with the solvent, produce an anti-correlated motion of the proximal and distal sides, which line the heme pocket, bringing atoms back to equilibrium positions. These movements may facilitate the diffusion of small ligands such as O_2 and CO to heme through the protein tunnel, exploiting its volume variations [111].

4.6 Elasticity and time scale

An estimate of the reference elastic constant γ of the model can be computed fitting the experimental temperature factors of the X-ray structures with the theoretical ones, obtained from the mean square displacements of C $^{\alpha}$ s, according to equation (4.10). Following the method used in [77] to fit the data (i.e. by matching the areas of the surface enclosed by the two data sets) and averaging the values found for the two proteins yields $\gamma = 0.20 \text{ Nm}^{-1}$, with a tolerance of 0.05 Nm^{-1} between averaged values (the introduction of heme in the network of interactions leads to a decrease of the value of the reference elastic constant, since it enhances the local connectivity of the buried residues in the heme pocket). The order of magnitude obtained for γ agrees with estimated values for the elastic constant of single parameter models [75, 80, 84, 85].

The importance of friction due to the solvent in determining the rates of functional motions of proteins, as it slows down the relaxation times of large-scale displacements predicted by normal mode analysis, has recently been underlined [114]: in the framework of the Langevin dynamics introduced with equation (4.8), we estimate order of magnitude for the reference decay time τ of the first two modes of motion previously described, through an effective value for the friction coefficient f , chosen to be the same for all residues for simplicity. A lower limit for f is the value computed from an all-atom simulation in [115], whereas here whole residues are considered (although the effective radii associated with such an estimate are bigger than the Van der Waals radii of the atoms in the simulation, hinting at a collective character of the simulated displacements [115], the motions predicted by the slowest modes involve many more residues in distant parts of the protein and a larger value for the friction may be expected). As an upper limit, the friction relative to the whole proteins (both PtrHb and CtrHb roughly fit a cubic box of side 3.5 nm) moving in water at physiological conditions is calculated from Stoke's law (see [116], chapter 3). We obtained $f \sim 4 \div 70 \text{ pN m}^{-1} \text{ s}$ (similar ranges for the values of friction coefficients have been extracted from molecular dynamics simulations [117]).

The corresponding reference relaxation time τ in the Langevin dynamics of equations 4.8 and 4.9 lies within the range $0.02 \div 0.35 \text{ ns}$, while the relaxation time associated with eigenmode i will be $\tau_i = \frac{\tau}{l_i}$ (where l_i is the eigenvalue relative to that eigenmode): the two slowest eigenmodes display relaxation times for the related motions approximately within the range $0.2 \div 3.5 \text{ ns}$.

This range of time scale is compatible with CO rebinding kinetics of Mbs and Hbs, while PtrHb and CtrHb behave quite differently [110]: the expla-

nation proposed for the different behaviour relies on the hydrogen-bonding network formed in the distal cavity of these trHbs, which is absent in invertebrate globins and is beyond the possibility of the simple model used here, which underlines instead the common characteristics of globins and trHbs.

Conclusions and perspectives

We have shown in this thesis that the validity of a coarse-grained model for folding, based on simple geometrical rules for the hydrogen bonds and on the physico-chemical properties common to globular proteins, can be extended to the modelization of the interaction of intrinsically unstructured proteins with a suitable partner. Such modelization is based mainly on the spatial arrangement of the binding centers which specify the interaction pattern of the IUP with its binding partner, emphasizing once more the role of geometry.

The successful application of the model adopted in the present work to several aspects of protein behaviour include the ability of globular proteins to fold reproducibly, the limited number and the simple modular nature of native state folds, which are built of helices and almost planar sheets and are recovered as local minima in the free energy landscape of a homopolymer chain. The framework developed for understanding these common features of globular proteins lends itself as well to the description of apparently disparate phenomena such as the behaviour of intrinsically unstructured proteins, as shown in this thesis, sequence design or amyloid formation.

This suggests how considerations of geometry and symmetry are crucial in the protein folding problem. They lead not only to the existence of a limited menu of native state folds but also, in the case of an intrinsically unstructured protein, to the possibility of selecting the structure of choice from this predetermined menu by suitably adjusting the patterns of effective interactions mimicking the presence of the binding partner of the IUP.

Remaining within a homopolymer description of the protein chain (the simplest possible one), we indeed have shown how a proper geometrical design of the interaction pattern proved to be crucial in the effective selection of the target fold, among the list of the presculpted minima predicted by the model for an isolated homopolymer chain. Still, in few cases the lack of more specific information, that could have been encoded in a heterogeneous sequence, prevents the chosen target fold from being the global free energy minimum. This is due to the competition from either the ground state conformation of single isolated chain, the α -helix, or other folds from the presculpted menu

that are still compatible with the spatial geometry of the interaction pattern.

The introduction of sequence heterogeneity, even a two-letter one within a simple hydrophobic-polar scheme, would probably help completing the successful discrimination of the correct target fold by means of geometric design procedure already employed in this thesis. This is indeed the direction of the ongoing work; the actual three-dimensional structure of the binding partner of a IUP is now used to model the binding sites of the substrate, and the actual sequence of the disordered segment is modeled adding heterogeneity, with already promising results.

In the last chapter of this thesis we presented, it has been shown how this simple coarse-grained approach can bring insights into the functional motions of two small proteins of the truncated hemoglobins family, PtrHb and CtrHb, near equilibrium vibrational properties of the structures modeled as a gaussian network of interacting α carbons and β centroids.

The key point in the analysis performed here is the information extracted from the covariance matrix in its reduced form and from the two slowest modes of fluctuation: negative correlations between residues set far apart in the tridimensional structure are particularly useful, being non trivial and hinting at the collective character of the motions.

This information has been used in the present work to confirm within such a simplified approach the mechanism which is believed to facilitate small ligands diffusion to the heme pocket and the iron atom. The cavity delimited by several key hydrophobic residues, providing a path from the surface of the protein to the heme pocket [86, 93, 113, 89], is able to enlarge its volume allowing the passage of small molecules to the distal side [110, 111], as it is inferred from the anti-correlations between the displacements of the opposite sides of the heme pocket.

Excitations, due to interactions between the molecule and the solvent, produce deviations from equilibrium followed by a decay towards the native state. The collective behaviour of the return back to equilibrium, produced by a superposition of overdamped motions, allow the volume of the inner cavities to vary accordingly.

Through a fit of the mean square displacements of α carbons from their minimum energy configuration with the experimental temperature factors for the two structures under study, a rough estimate of the order of magnitude of time scale for functionally relevant motions has been given, in reasonable agreement with known properties of globular proteins.

This suggests the validity of the simple gaussian approach as a means to get a fast picture of the near-native functional motions of globular proteins, yet in agreement with the results obtained using more accurate and computationally demanding tools.

The description given by the simple model used here does not provide atomic details, keeping the analysis at a coarse-grained level. Still the use of the effective β centroid for each residue, along with the C^α , helps in characterizing with more adherence to reality the displacements of residues side-chains, thus getting a closer agreement with more detailed approaches.

Bibliography

- [1] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.*, 338:1015–1026, 2004.
- [2] D. Shortle and M. S. Ackerman. Persistence of native-like topology in a denatured protein in 8 m urea. *Science*, 293:487–489, 2001.
- [3] V. N. Uversky. What does it mean to be natively unfolded? *Eur. J. Biochem.*, 269:2–12, 2002.
- [4] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratcliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. H. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and z. Obradovic. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19:26–59, 2001.
- [5] A. K. Dunker and Z. Obradovic. The protein trinity - linking function and disorder. *Nat. Biotechnol.*, 19:805–806, 2001.
- [6] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12:54–60, 2002.
- [7] P. Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27:527–533, 2002.
- [8] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. Geometry and symmetry pre-sculpt the free energy landscape of proteins. *Proc. Natl. Acad. Sci. USA*, 101:7960–7964, 2004.
- [9] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato. A unified perspective on proteins – a physics approach. *Phys. Rev. E*, 70:041905, 2004.

- [10] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293:7960–7964, 1999.
- [11] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
- [12] Swiss-prot sequence database. <http://www.expasy.org/sprot/>.
- [13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.
- [14] T. E. Creighton. *Proteins: Structures and Molecular Properties*. W.H. Freeman and Company, New York, USA, 1993.
- [15] G. N. Ramachandran and V. Sasisekharan. Conformations of polypeptides and proteins. *Adv. Protein Chem.*, 23:283–438, 1968.
- [16] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.
- [17] C. Sander and R. Schneider. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.
- [18] C. Chothia and A. Lesk. The relation between the divergence of sequences and structures in proteins. *EMBO J.*, 5:823–826, 1986.
- [19] C. Levinthal. Are there pathways to protein folding? *J. Chem. Phys.*, 65:44–45, 1968.
- [20] Y. Duan and P. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in water solution. *Science*, 282:740–744, 1998.
- [21] N. Gō and H. A. Scheraga. On the use of classical statistical mechanics in the treatment of polymer chain conformations. *Macromolecules*, 9:535–542, 1976.
- [22] J. Bryngelson, J. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins*, 21:167–195, 1995.

- [23] K. A. Dill, S. Bromberg, K. Yue, K. Fiebig, D. Yee, and P. Thomas. Principles of protein folding: a perspective from simple exact models. *Protein Science*, 4:561–602, 1995.
- [24] S. Kamtekar, J. M. Schiffer, H. J. Xiong, J. M. Babik, and M. H. Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685, 1993.
- [25] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278:82–87, 1997.
- [26] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic level accuracy. *Science*, 302:1364–1368, 2003.
- [27] O. Schweers, E. Schonbrunn Hanebeck, A. Marx, and E. Mandelkow. Structural studies of tau protein and alzheimer paired helical filaments show no evidence for β -structure. *J. Biol. Chem.*, 269:24290–24297, 1994.
- [28] S. Ohnishi and D. Shortle. Observation of residual dipolar couplings in short peptides. *Proteins*, 50:546–551, 2003.
- [29] V. N. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci.*, 11:739–756, 2002.
- [30] I. Radhakrishnan, G. C. Perez-Alvarado, D. Parker, H. J. Dyson, M. R. Montminy, and P. E. Wright. Solution structure of the kix domain of cbp bound to the transactivation domain of creb: a model for activator-coactivator interactions. *Cell*, 91:741–752, 1997.
- [31] D. Parker, M. Rivera, T. Zor, A. Henrion-Caude, I. Radhakrishnan, A. Kumar, L. H. Shapiro, P. E. Wright, M. Montminy, and P. K. Brindle. Role of secondary structure in discrimination between constitutive and inducible activators. *Mol. Cell Biol.*, 19:5601–5607, 1999.
- [32] A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by using the folding funnel. *Proc. Natl. Acad. Sci. USA*, 97:8868–8873, 2000.
- [33] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are natively unfolded proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, 41:415–427, 2000.

- [34] A. M. Lesk. *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press, Oxford, UK, 2000.
- [35] C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, 31:45–71, 2002.
- [36] O. Gonzalez and J. H. Maddocks. Global curvature, thickness, and the ideal shapes of knots. *Proc. Natl. Acad. Sci. USA*, 96:4769–4773, 1999.
- [37] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar. Optimal shapes of compact strings. *Nature*, 406:287–290, 2000.
- [38] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: two hydrogen bonded helical conformations of the polypeptide chain. *P. Natl. Acad. Sci. USA*, 37:205–211, 1951.
- [39] L. Pauling and R. B. Corey. Conformations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *P. Natl. Acad. Sci. USA*, 37:729–740, 1951.
- [40] P. J. Kraulis. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.
- [41] E. A. Merritt and D. J. Bacon. Raster3d photorealistic molecular graphics. *Methods in Enzymology*, 277:505–524, 1997.
- [42] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *P. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.
- [43] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.
- [44] J. D. Bernal. Structure of proteins. *Nature*, 143:663–667, 1939.
- [45] D. Perl, C. Welker, T. Schindler, K. Schroder, M. A. Marahiel, R. Jaenicke, and F. X. Schmid. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.*, 5:229–235, 1998.
- [46] L. Holm and C. Sander. Structural alignment of globins, phycocyanins and colicin a. *FEBS Lett.*, 315:301–306, 1993.
- [47] R. V. Pappu, R. Srinivasan, and G. D. Rose. The floppy isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *P. Natl. Acad. Sci. USA*, 97:12565–12570, 2000.

- [48] R. Kazmierkiewicz, A. Liwo, and H. A. Scheraga. Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a monte-carlo method. *J. Comput. Chem.*, 23:715–723, 2002.
- [49] L. Holm and C. Sander. Database algorithm for generating protein backbone and side-chain coordinates from a c alpha trace application to model building and detection of coordinate errors. *J Mol. Biol.*, 218:183–194, 1991.
- [50] J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato. Geometry of compact tubes and protein structures. *ComplexUs*, 1:4, 2003.
- [51] A. D. Sokal. Monte Carlo methods for the self-avoiding walk. *Nucl. Phys. B, Suppl.* 47:172–179, 1996.
- [52] N. Madras and A. D. Sokal. The pivot algorithm: a highly efficient monte carlo algorithm for the self-avoiding random walk. *J. Stat. Phys.*, 50:109, 1988.
- [53] A. D. Sokal. Monte carlo methods for the self-avoiding walk. In *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*. Oxford University Press, Oxford, UK, 1994.
- [54] J. Baschnagel, J. P. Wittmer, and H. Meyer. Monte carlo simulation of polymers: coarse-grained models. In *Computational soft matter: from synthetic polymers to proteins, Lecture notes*, volume 23 of *NIC series*, pages 83–140. John von Neumann Institute for Computing, Juelich, Germany, 2004.
- [55] T. Kennedy. A faster implementation of the pivot algorithm for self-avoiding walks. *J. Stat. Phys.*, 106:407–429, 2002.
- [56] L. Mattioni, J. P. Wittmer, J. Baschnagel, J.-L. Barrat, and E. Luijten. Dynamical properties of the slithering-snake algorithm: A numerical test of the activated-reptation hypothesis. *Eur. Phys. J. E*, 10:369–385, 2003.
- [57] S. Kirkpatrick, C. D. Jr. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [58] M. C. Tesi, E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington. Monte carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.*, 82:155–181, 1996.

- [59] A. M. Ferrenberg and R. H. Swendsen. New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635, 1988.
- [60] A. M. Ferrenberg and R. H. Swendsen. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, 63:1195, 1989.
- [61] A. M. Ferrenberg. Modern methods of analyzing monte carlo computer simulations. *Phys. A*, 194:53–62, 1993.
- [62] T. Bogner, A. Degenhard, and F. Schmid. Molecular recognition in a lattice model: an enumeration study. *Phys. Rev. Lett.*, 93:268108, 2004.
- [63] W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *J. Molec. Graphics*, 14:33–38, 1996.
- [64] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A 32:922, 1976.
- [65] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A 34:827–828, 1978.
- [66] T. Noguti and N. Gō. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature*, 296:776–778, 1982.
- [67] N. Gō, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. of Sci. USA*, 80:3696–3700, 1983.
- [68] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. of Sci. USA*, 80:6571–6575, 1983.
- [69] M. Levitt, C. Sander, and P. S. Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.: Quant. Biol. Symp*, 10:181–199, 1983.
- [70] M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181:423–447, 1985.
- [71] S. Hayward and N. Gō. Collective variable description of native protein dynamics. *Annu. Rev. of Phys. Chem.*, 46:223–250, 1995.

- [72] S. Hayward, A. Kitao, and N. Gō. Harmonic and anharmonic aspects in the dynamics of bpti: A normal mode analysis and principal component analysis. *Proteins Sci.*, 3:936–943, 1994.
- [73] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, 17:412–425, 1993.
- [74] T. Horiuchi and N. Gō. Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins: Structure, Function, and Genetics*, 10:106–116, 1991.
- [75] M. Tirion. Large amplitude elastic motions in proteins from a single-parameter. atomic-analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.
- [76] T. Haliloglu, I. Bahar, and B. Erman. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, 79:3090–3093, 1997.
- [77] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.
- [78] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman. Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.*, 80:2733–2736, 1998.
- [79] P. Doruker, A. R. Atilgan, and I. Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to α -amylase inhibitor. *Proteins: Structure, Function, and Genetics*, 40:512–524, 2000.
- [80] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.
- [81] P. Doruker, R. L. Jernigan, and I. Bahar. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.*, 23:119–127, 2001.
- [82] F. Tama and Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, 14:1–6, 2001.
- [83] F. Tama, M. Valle, J. Frank, and C. L. Brooks. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis

- and cryo-electron microscopy. *Proc. Natl. Acad. Sci. USA*, 100:9319–9323, 2003.
- [84] C. Micheletti, P. Carloni, and A. Maritan. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins: Structure, Function, and Bioinformatics*, 55:635–645, 2004.
- [85] M. Neri, M. Cascella, and C. Micheletti. Influence of conformational fluctuations on enzymatic activity: modelling the functional motion of beta-secretase. *J. Phys. Cond. Mat.*, in press.
- [86] A. Pesce, M. Couture, S. Dewilde, M. Guertin, K. Yamauchi, P. Ascenzi, L. Moens, and M. Bolognesi. A novel two-over-two α -helical sandwich fold is characteristic of the truncated hemoglobin family. *The EMBO Journal*, 19:2424–2434, 2000.
- [87] M. Couture, S. Yeh, B. A. Wittenberg, J. B. Wittenberg, Y. Ouellet, D. L. Rousseau, and M. Guertin. A cooperative oxygen-binding hemoglobin from *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA*, 96:11223–11228, 1999.
- [88] S. Yeh, M. Couture, Y. Ouellet, M. Guertin, and D. L. Rousseau. A cooperative oxygen-binding hemoglobin from *Mycobacterium tuberculosis*. stabilization of heme ligands by a distal tyrosine residue. *Proc. Natl. Acad. Sci. USA*, 96:11223–11228, 1999.
- [89] J. B. Wittenberg, M. Bolognesi, B. A. Wittenberg, and M. Guertin. Truncated hemoglobins: a new family of hemoglobins widely distributed in bacteria, unicellular eukaryotes, and plants. *J. Biol. Chem.*, 277:871–874, 2001.
- [90] M. Potts, S. V. Angeloni, R. E. Ebel, and D. Bassam. Myoglobin in a cyanobacterium. *Science*, 256:1690–1691, 1992.
- [91] M. V. Thorsteinsson, D. R. Bevan, M. Potts, Y. Dou, R. F. Eich, M. S. Hargrove, Q. H. Gibson, and J. S. Olson. A cyanobacterial hemoglobin with unusual ligand binding kinetics and stability properties. *Biochemistry*, 38:2117–2126, 1992.
- [92] M. Couture, H. Chamberlan, B. St-Pierre, J. Lafontaine, and M. Guertin. Nuclear genes encoding chloroplast hemoglobins in the unicellular green alga *Chlamydomonas Eugametos*. *Mol. Gen. Genet.*, 243:185–197, 1994.

- [93] M. Milani, A. Pesce, Y. Ouellet, P. Ascenzi, M. Guertin, and M. Bolognesi. *Mycobacterium tuberculosis* hemoglobin n displays a protein tunnel suited for o_2 diffusion to the heme. *The EMBO Journal*, 20:3902–3909, 2001.
- [94] M. Brunori and Q. H. Gibson. Cavities and packing defects in the structural dynamics of myoglobin. *EMBO Rep.*, 2:674–679, 2001.
- [95] F. Schotte, M. Lim, T. A. Jackson, A. V. Smirnov, J. Soman, J. S. Olson, G. N. Phillips Jr., M. Wulff, and P. A. Anfinrud. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300:1944–1947, 2001.
- [96] M. Teeter. Myoglobin cavities provide interior ligand pathway. *Protein Science*, 13:313–318, 2004.
- [97] M. F. Perutz. Regulation of oxygen affinity of hemoglobin: influence of structure of the globin on the heme iron. *Annu. Rev. Biochem.*, 48:327–386, 1979.
- [98] M. W. King. Biochemistry of amino acids. <http://web.indstate.edu/>, 2003.
- [99] B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.
- [100] M. Doi and S. F. Edwards. *The theory of polymer dynamics*. Clarendon Press, Oxford, UK, 1986.
- [101] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 1999.
- [102] L. Stryer. *Biochemistry*. W. H. Freeman and Company, New York, NY, 1995.
- [103] Y. Seno and N. Gō. Deoxymyoglobin studied by the conformational normal mode analysis. i. dynamics of globin and the heme-globin interaction. *J. Mol. Biol.*, 216:95–109, 1990.
- [104] Y. Seno and N. Gō. Deoxymyoglobin studied by the conformational normal mode analysis. ii. the conformational change upon oxygenation. *J. Mol. Biol.*, 216:111–126, 1990.

- [105] C. Xu, D. Tobi, and I. Bahar. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin t \leftrightarrow r2 transition. *J. Mol. Biol.*, 333:153–168, 2003.
- [106] C. Micheletti, G. Lattanzi, and A. Maritan. Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. *J. Mol. Biol.*, 321:909–921, 2002.
- [107] B. Halle. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA*, 99:1274–1279, 2002.
- [108] F. Fraternali and L. Cavallo. Parameter optimized surfaces (pops): analysis of key interactions and conformational changes in the ribosome. *Nucl. Acids Res.*, 30:2950–2960, 2002.
- [109] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65:31910–31919, 2002.
- [110] U. Samuni, D. Dantsker, A. Ray, J. B. Wittenberg, B. A. Wittenberg, S. Dewilde, L. Moens, Y. Ouellet, M. Guertin, and Joel M. Friedman. Kinetic modulation in carbonmonoxy derivatives of truncated hemoglobins. *J. Biol. Chem.*, 278:27241–27250, 2003.
- [111] A. Crespo, M. A. Martí, S. G. Kalko, A. Morreale, M. Orozco, J. L. Gelpi, F. J. Luque, and D. A. Estrin. Theoretical study of the truncated hemoglobin hbn: exploring the molecular basis of the no detoxification mechanism. *J. Am. Chem. Soc.*, 127:4433–4444, 2005.
- [112] M. Milani, A. Pesce, Y. Ouellet, S. Dewilde, J. Friedman, P. Ascenzi, M. Guertin, and M. Bolognesi. Heme-ligand tunneling in group i truncated hemoglobins. *J. Biol. Chem.*, 279:21520–21525, 2004.
- [113] M. Milani, P. Y. Savard, H. Ouellet, P. Ascenzi, M. Guertin, and M. Bolognesi. A tyrcd1/trpg8 hydrogen bond network and a tyrb10tyrcd1 covalent link shape the heme distal site of *Mycobacterium tuberculosis* hemoglobin o. *Proc. Natl. Acad. Sci. USA*, 100:5766–5771, 2003.
- [114] J. Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373–380, 2005.
- [115] S. Swaminathan, T. Ichiye, W. van Gunsteren, and M. Karplus. Time dependence of atomic fluctuations in proteins: analysis of local and collective motions in bovine pancreatic trypsin inhibitor. *Biochemistry*, 21:5230–5241, 1982.

-
- [116] J. Howard. *Mechanics of motor proteins and the cytoskeleton*. Sinauer Associates, Sunderland, MA, 2001.
- [117] K. Hinsén, A. J. Petrescu, and S. Dellerue. Harmonicity in slow protein dynamics. *Chem. Phys.*, 261:25–37, 2000.

Acknowledgements

I wish to thank my Supervisors – Amos, Antonio and Cristian – for their stable support during my study in the last few years, spent between Trieste and Padova, and above all for giving me the opportunity to face such interesting research topics. They provided a lot of useful hints to improve my work, and have always been encouraging when difficulties arose, which seemed to be overwhelming.

I am particularly grateful to Antonio, both for his never wearying patience while dealing with my rash surmises, at any time unshakably calm and collected, and for his kind hospitality in Padova.

I am deeply beholden to Gianluca, who upheld my undertaking at SISSA from the very beginning: I could rely daily on his steadfast friendship, essential against the venturesome everyday issues, despite the distance.

I gratefully acknowledge the supportive attitude of all the people of the Sector of Statistical and Biological Physics, who yielded invariably precious and helpful encouragements, which were crucial to afford my study.

I have also strongly appreciated the kind hospitality and welcoming of the friendly people I met during my sojourns at the Physics Department of the University of Padova.

Finally, I would like to warmly thank all the wonderful friends I had the good luck to meet in Trieste, to whom I am indebted more deeply than I could ever realize, indeed beyond any chance of payback, despite any effort of myself.