

UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Fisica



Tesi di Dottorato di Ricerca in Fisica
Ph.D. Thesis in Physics

**Protein structural dynamics and thermodynamics
from advanced simulation techniques**

Supervisor:
Dr. Pietro Faccioli

Candidate:
Giorgia Cazzoli

DOTTORATO DI RICERCA IN FISICA, XXVI CICLO
Trento, 20th December 2013

Contents

Introduction	ix
1 Proteins	1
1.1 DNA, RNA and proteins	1
1.2 Protein folding	3
1.3 Proteins and life	4
1.4 The protein folding problem	5
1.5 The statistical view and the energy landscape	6
1.6 The study of the protein folding	8
2 Experiments and simulations in protein folding	9
2.1 Experiments and theory: two sides of the same coin	9
2.2 Experimental methods in protein folding	9
2.3 Computer simulations applied to protein folding: an overview	12
2.3.1 Methods	12
2.3.2 State of the art of the protein folding simulations . . .	13
3 Theoretical Models	17
3.1 Coarse-grained vs. all-atom models	17
3.2 Description of water	19
3.3 Realistic all-atom force field	22
3.4 $G\bar{o}$ model	23
3.5 $G\bar{o}$ model with the addition of non- $G\bar{o}$ interactions	27
4 Monte Carlo method applied to protein folding	31
4.1 The Monte Carlo approach: an introduction	31
4.1.1 The method	31
4.1.2 The Metropolis algorithm	34

4.2	Application to protein folding	35
5	Dominant Reaction Pathways approach	37
5.1	The overdamped Langevin equation	37
5.2	Path integral representation and its application to the DRP problem	38
5.3	rMD	41
5.4	MD vs. rMD-DRP	43
6	Case of study 1: Folding process of the immunity proteins IM7 and IM9	45
6.1	Introduction to the problem	45
6.1.1	Intermediate states during folding	45
6.1.2	Experiments and simulations try to characterize IM7 and IM9 folding	47
6.2	Methods	48
6.2.1	PDB files preparation	48
6.2.2	Parameters for rMD simulations	49
6.2.3	RMSD, or <i>Root-mean-square-deviation</i>	49
6.2.4	Calculation of the fraction of native contacts	49
6.2.5	Definition of the “Kinetic free energy”	50
6.3	Results	50
6.3.1	Characterization of the folding process in IM9 and IM7	50
6.3.2	IM7 mutant	55
6.3.3	Characterization of the initial unfolded state	58
6.4	Conclusions of this comparative study	62
7	Case of study 2: Thermal adaptation of marine ciliate pheromones	65
7.1	Introduction	65
7.1.1	Why thermal adaptation is important?	65
7.1.2	Charcterization of the pheromones produced by <i>Eu-</i> <i>plotes nobilii</i> and <i>Euplotes raikovi</i>	66
7.1.3	Pheromone thermal adaptation: an open debate...	69
7.1.4	...and our hypothesis: the role of CYS-CYS bonds	69
7.2	Methods	71
7.2.1	Monte Carlo simulations	71
7.2.2	Input files describing starting structures	72
7.3	Results	72

7.3.1	Fraction of native contacts as a function of the temperature	72
7.3.2	Specific heat and the fluctuation of the fnc	75
7.3.3	Localization of CYS-CYS bonds	78
7.3.4	The mutant <i>En</i> pheromone	79
7.4	Conclusions of the study	80
8	Case of study 3: Atomic-level characterization of conformational changes in Serpin family	83
8.1	Serpins: a general introduction	83
8.1.1	Serine protease inhibitor	83
8.1.2	Serpins: Structure and plasticity	83
8.1.3	The serpin mechanism	85
8.1.4	Serpins' diseases and misfolding	87
8.1.5	Characterization of the serpin conformational change from experiments and simulations	88
8.2	Our work overcomes the limits in serpin reaction investigation.	90
8.2.1	The challenge: is it possible simulate conformational changes in serpins?	90
8.3	Methods	92
8.3.1	MC simulations	92
8.3.2	rMD-DRP simulations	93
8.3.3	RMSD	94
8.3.4	Euclidean distance	94
8.4	Results	94
8.4.1	MC results and analysis	94
8.4.2	Complete all-atom DRP simulations of the latency transition in serpins.	99
8.4.3	Kramers-Arrhenius analysis of reaction kinetics	103
8.4.4	Characterization at the atomistic detail of the conformational changes in serpins during latency transition.	105
8.4.5	Energy analysis of PAI-1 WT latency transition	113
8.5	Conclusions of this study	116
9	General conclusions	117
A	The generalized Born model	119

B $G\bar{o}$ model: the unfolding temperature and the native contact energy	123
Acknowledgements	141

τὸν αὐτὸν τρόπον καὶ οὗτοι τὰς διαφορὰς αἰτίας
 τῶν ἄλλων εἶναι φασιν. ταύτας μέντοι τρεῖς εἶναι
 λέγουσι, σχῆμά τε καὶ τάξιν καὶ θέσιν: διαφέρειν
 γὰρ φασὶ τὸ ὄν ῥυσμῶ καὶ διαθιγῆ καὶ τροπῆ
 μόνον: τούτων δὲ ὁ μὲν ῥυσμὸς σχῆμά ἐστίν ἢ δὲ
 διαθιγῆ τάξις ἢ δὲ τροπῆ θέσις: διαφέρει γὰρ τὸ
 μὲν Α τοῦ Ν σχήματι τὸ δὲ ΑΝ τοῦ ΝΑ τάξει τὸ δὲ
 Ζ τοῦ Η θέσει.

“...the differences [of the atoms] are the causes of everything else. These differences are three: shape, arrangement and position; because what is differs only in contour, inter-contact and inclination. Of these contour means shape, inter-contact arrangement and inclination position. Thus, for example, A differs from N in shape, AN from NA in arrangement and Z from N in position.”

Aristotle, Metaphysics, Book 1, section 985b

this case the protein is inactive and the capability of carrying out its function breaks down, causing diseases that are directly correlated with the lack of activity of the specific protein. But there is also another consequence of the misfolding: these inactivated proteins tend to form aggregates, within cells and tissues, associated with dysfunction and cell death, explicative is the case of Parkinson and Alzheimer diseases as a consequence of amyloid aggregates, but also other neurodegenerative diseases can happen in response of aggregation of other proteins such as neuroserpins.

For this reason the study of all the issues related to proteins, their folding and behavior in response to the environment, described in a detailed manner, are essential for a deep comprehension of what can cause a misfolding and of how it is possible to prevent it.

A number of experimental and theoretical techniques have been applied in these years in order to give new insights into these topics. Among these techniques computer simulations are becoming interesting for simulating the process that leads the protein to its final state. Indeed, thanks to the advances in method and hardware, the approach turns out to be able to describe folding and conformational changes occurring in systems with increasing size and reactions at slower and slower time-scale.

In this work we apply advanced simulation methods, namely Monte Carlo simulations using the Metropolis algorithm in a coarse grained model and all-atom biased Molecular Dynamics simulations centred on a path integral based method called Dominant Reaction Pathways (DRP) approach, in order to study three different families of proteins, whose behavior, in aspects such as on-pathway intermediate states or unfolding thermodynamics or dynamics until equilibrium, has not yet been completely clarified. We decided to approach these issues by analyzing, for each investigated family, homologous proteins, evolutionarily correlated, very similar in structure, but with different behavior in respect of the focused characteristics.

The first treated issue is correlated directly with the topic of folding and focuses on two proteins, IM9 and IM7. These proteins belong to colicin immunity protein family produced by *Escherichia coli* and, despite the high degree of structural similarity, seem to fold in a different manner that involves, respectively, a two-state process, namely a folding process without the presence of populated intermediate states, and a three state mechanism with a well populated on-pathway intermediate. Our aim has been try to verify the presence of the intermediate, characterize it by describing the structures that are present in this state and understand the type of interactions that stabilize such state.

Then, another studied aspect is represented by evolutive adaptation of proteins to environment, in particular to cold temperature, where only cold-adapted organisms can survive and be active. The study investigates the

different thermodynamic behaviors showed by two classes of pheromones that are very similar in structure but live respectively in arctic and temperate water. We performed computer simulations in order to deep analyze the set of features that these organisms develop in response to adaptation to the environment.

Finally a study of the dramatic conformational changes that take place in a superfamily of proteins, the serpins, has been carried out. The serpins are peculiar proteins since their native state is a metastable state and only this conformational change allows them to reach the final stable configuration. Despite the large diffusion and importance of these proteins, whose misfolding, modification in time-scale related to the conformational change and aggregation can cause severe diseases, little is known regarding the events that lead to the final state. Indeed, experiments can only suggest possible mechanisms that underlie such activity and computer simulations fail to describe the whole process, because of the big size of the protein, more than 300 residues, and the very slow processes involved, from hours to weeks, completely beyond the capacity of the simulation methods to date. We tried to study this challenging problem by investigating types of serpins with a high degree of similarity in structure but with difference in the amino acid sequence in order to achieve a complete explanation of the residues that affect the dynamics and of the events that occur during the reaction.

The three specific problems described in this thesis cover a broad range of topics, involving both dynamical and thermodynamical aspects. The concepts, principles and methodologies which are used in the literature to approach these problems are usually applied to small globular proteins. On the other hand, in our thesis we consider systems which are somewhat peculiar, in this perspective, because of their large size, or because of the existence of topological constraints, or because of the presence of intermediate states during the folding process. This will allow to investigate whether general picture developed for small globular proteins, such as the funneled energy landscape theory, is still holding true for this larger class of systems.

The work is structured as follows:

Chapter 1 - Proteins: After a short introduction about proteins, their origin and function, the chapter has been focused on the protein folding problem.

Chapter 2 - Experiments and simulations in protein folding: How is it possible to study proteins both from experimental and theoretical point of view? This chapter tries to answer this question.

Chapter 3 - Theoretical Models: Description of the principal models and approximations for describing the protein, the solvent and the interac-

tions.

Chapter 4 - Monte Carlo method applied to protein folding: After a short introduction about Monte Carlo methods the Metropolis algorithm and the choices applied to the present work are described.

Chapter 5 - Dominant Reaction Pathways approach: The ratchet-and-pawl MD (rMD) algorithm and the DRP approach are presented. These concepts are both applied in the Molecular Dynamics simulations performed in this work.

Chapter 6 - Case of study 1: Folding process of the immunity proteins IM7 and IM9: The folding of two homologous proteins, the immunity proteins IM7 and IM9, is studied with the help of rMD-DRP simulations, in order to verify the presence of an intermediate state and, once observed, completely characterize this state for what concerns structures and interactions. The importance of this study has to be included into the more general picture that deals with concepts like funneled energy landscape and minimal frustration, since some works performed on IM7 and IM9 challenge the native-centric view. Our investigation suggests that both IM7 and IM9 fold driven by native interactions.

Chapter 7 - Case of study 2: Thermal adaptation of marine ciliate pheromones: *Euplotes nobilii* and *Euplotes raikovi* pheromones are produced by ciliate organisms living in cold and temperate water, respectively. As a consequence of the adaptation to different environments the pheromones show also different thermodynamic behaviours that are investigated with the help of Monte Carlo simulations. A role played by the location of CYS-CYS bond along the chain has been found as possible cause of stability of *Euplotes raikovi* pheromones at the increasing of the temperature.

Chapter 8 - Case of study 3: Atomic-level characterization of conformational changes in Serpin family: Serpins are proteins that, because of their big size and very slow dynamics associated to their function are challenging to study with MD simulations and experiments cannot give a complete analysis since the high degree of flexibility found in the protein. We present the first all-atom simulation carried on with rMD-DRP approach and the complete characterization of the mechanism related to serpins. Moreover, important implications for what concerns medical research field, in particular in drug design, are drawn from this detailed analysis.

Chapter 9 - General conclusions: General conclusions regarding the work are presented.

Chapter 1

Proteins

1.1 DNA, RNA and proteins

Inside the living cells the agents that perform cellular functions are *proteins*, an example is reported in Fig.1.1. G. J. Mulder, who discovered them in 1830s, called them proteins, from the Greek word *πρωτεῖος* that means “of the first importance”, as they were considered to be the most important of cellular materials [1].

The origin of proteins has to be investigated in the genetic molecules that are contained in the cells. In the *eukaryotic* cells these molecules are located in the well defined nucleus and are called *chromatin*. A type of chromatin are the *chromosomes*, each chromosome consists of one DNA molecule, festooned at regular intervals with bead-like proteins called *histones* [2]. The DNA molecule is constituted by sequences, the *genes*, that carry the code for building one type of protein. In particular, the protein synthesis can be summarized in the following diagram [2]:

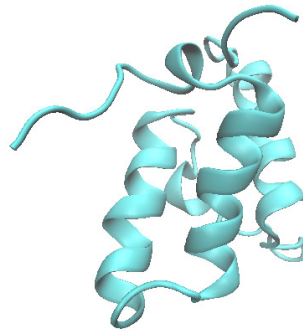


Figure 1.1: Protein IM9, PDB code 1IMQ.pdb



The first step of transcription occurs inside the nucleus: a specific gene is copied to an *RNA* molecule, which is then transported through pores in the nuclear membrane to the cytoplasm. There it is translated into protein molecules by particular organelles called *ribosomes* [2].

On the contrary, in case of *prokaryotic* cells, where in totally lack of a

defined nucleus the genetic molecules float free in the cytoplasm, translation of a transcript begins before the transcript is complete [2, 3].

In the initial stage that comes after their production proteins are unstructured long-chain molecules consisting of a backbone made up of *amino acids*, also called residues, connected sequentially via a peptide bond, see Fig 1.2. For this reason the chain is called a *polypeptide chain* [2].

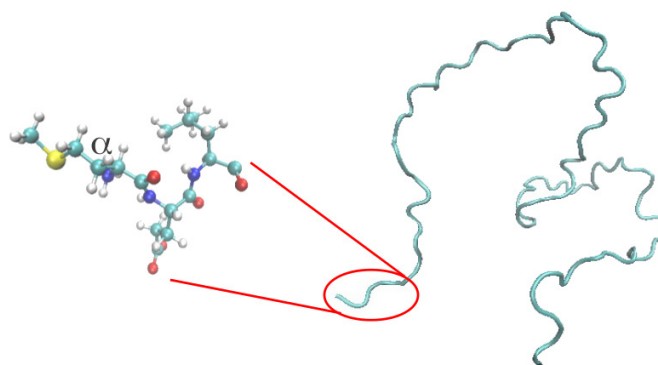


Figure 1.2: Protein in its initial configuration as a sequence of amino acids. The particular of the amino acids linked together is also reported.

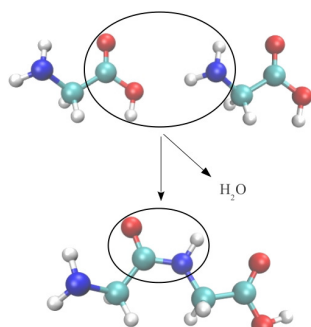


Figure 1.3: Formation of the peptide bond between two amino acids.

Amino acids are molecules that can exist also alone and are made up of a central α -carbon, highlighted in Fig. 1.2. Four groups are attached to the central carbon: amino group ($-NH_2$), atom N is conventionally reported in blue, as in Fig. 1.2, a carboxyl group ($-COOH$), the oxygen is colored in red, a hydrogen atom, in white, and a fourth arbitrary group ($-R$), that determines the difference in the 20 various types of amino acids used for building proteins [4]. Between two amino acids a peptide bond, that is a covalent bond, can be formed: this occurs when the carboxyl group of one residue reacts with the amino group of the next residue in sequence with the release of a molecule of water, as it is summarized in Fig. 1.3.

Only after their synthesis in the ribosomes, proteins are released in the cytoplasm, where the process called “*protein*

folding” can take place.

1.2 Protein folding

In the aqueous environment of the cytoplasm the polypeptide chain folds into its “*native state*”, whose geometrical shape is important for carrying out successfully the biological function. Locally, the chain curls up into α -helices, or braids into β -sheets, as showed in Fig. 1.4. These structures are called *secondary structures*. All the helices and the sheets inside a protein are arranged in a three-dimensional architecture called *tertiary structure*, namely the native state of the protein [2].

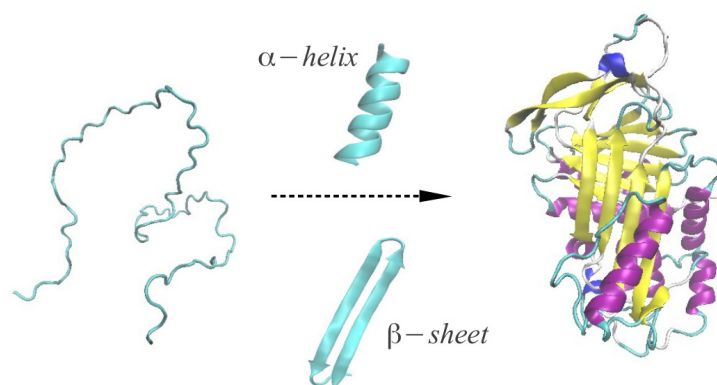


Figure 1.4: From a chain of amino acids, on the left, until the native state, on the right. The possible secondary structures, α -helix and β -sheet are also presented. The reported native state is that of serpin Plasminogen activator inhibitor-1, PDB code 1oc0.pdb

An important driving force for protein folding is the hydrophobic effect [2, 5]. Indeed, water molecules seek to form hydrogen bonds with each other or with other polar molecules but are hindered in this action by the presence, in water, of nonpolar molecules. The water molecules try to push them out of the way and the nonpolar molecules appear to avoid contact water, in this sense we can say that they are *hydrophobic* [2]. However, other types of interactions can contribute to protein folding. The Coulomb interactions between charged amino acids and van der Waals interactions may have a role during protein folding. In particular, it has been suggested that van der Waals interactions are involved in forming the hydrophobic core of the protein. This hypothesis is supported by studies that evidence that there is a strong difference between the protein core and that of nonpolar liquids. As

a consequence, it seems that increased protein stability is correlated with an increased number of van der Waals contacts [6, 7, 5]. Another factor that is expected to be involved in stabilizing protein folding is the peptide hydrogen bond. The interest in this type of H bond arises from studies about the stability of the helix backbone in water: the work performed by Shellman [8] concluded that the helix was marginally stable in water. Moreover, it has been showed that the carboxyl and amino groups interact with water as well as with each other [9, 5].

Only when protein has reached the native state, at the end of the protein folding, it is possible for the protein to carry out its particular function.

1.3 Proteins and life

Proteins are versatile [4]: in fact, they are able to assume different shapes and consequently functions. As a result, proteins are essential for life for their application to a vast range of processes in living beings [4].

In order to understand their basic role some examples of protein functions are proposed: proteins are building blocks of many biological structures, this is the case of epidermal keratin or collagen in bones and cartilages, but proteins can also transport and store other species, from electrons to macromolecules, or, as hormones, transmit information and signals between cells and organs [4]. Proteins are able, as antibodies, to defend the organism against intruders and they are also essential components of muscles, converting chemical energy into mechanical one allowing the movements in animals[4]. Proteins can control the functions of other proteins in order to assure their correct and balanced action on the path, in which they are involved.

As a consequence of this ubiquitous participation in almost every processes that are essential for life, protein science constitutes a topic of more and more broad interest for its repercussions in Medicine [4]. Indeed, the lack or malfunction of proteins is the reason of many pathologies directly correlated with the type of protein involved. For example, we can consider the case of antithrombin, a protein belonging to the class of serpins. In normal conditions antithrombin is able to control the function of the enzyme thrombin but in presence of mutations this activity breaks down and cases of venous thrombosis are observed [10]. Then, proteins that fail to fold correctly give rise to aggregations that characterize many neurodegenerative disorders, including Huntington, Alzheimer, Creutzfeld-Jacob ('mad cow'), or motor neuron diseases [4, 11, 12]. Finally, attack the proteins of pathogens such as HIV, SARS and hepatitis or to block the synthesis of proteins in bacteria are both strategies to fight infections in the field of drug design [4].

1.4 The protein folding problem

In view of the central role played by proteins in biology it is very important to study and analyze all the mechanisms that take proteins from the chain of amino acids to a well defined three-dimensional structure in the native, stable state. The capability to predict the sequences of movements that lead to the final state, the presence of intermediates, and the configurations that are visited or on the contrary avoided during the folding, can give an insight into the understanding on why, under particular conditions, proteins can misfold and as a consequence fail to carry out their functions.

The problem now arises from the challenging aspect of this issue. Kendrew, speaking of the first protein, myoglobin, ever resolved, said [13]:

The most striking features of the molecule were its irregularity and its total lack of symmetry.

That can be considered true for most of the proteins. However, thanks to a series of experiments carried out by Christian B. Anfinsen in the 50s [4] the conclusion has been that all the information needed to reach the native state is encoded in the sequence of amino acids [14, 15], although the existence of the so-called molecular chaperones, which help the proteins fold in the cellular milieu, has been known. Indeed, it has been considered that these molecular assistants should not add any structural information to the process [4].

After Anfinsen's experiments Levinthal recognized the difficulty of a molecule searching at random through the large number of unfolded configurations to find the folded structure in a biologically relevant time [16]. This searching would in fact take an enormously long time, since, indeed, each bond connecting amino acids can have several possible states, for example 3, so that a protein of 101 amino acids, it is only a quantity chosen for the purpose of the calculation, could assume 3^{100} configurations. Even if the protein is able to sample new configurations at the rate of 10^{13} per second, or 3×10^{20} per year, it will take 10^{27} years to try them all, while proteins can fold in seconds or less [17]. This paradox has been overcome by postulating the notion of a protein folding pathway [16, 18], although also the possibility of multiple parallel paths towards the folded state has been taken into account [16]. In these last years a new way for approaching to the protein folding problem has emerged based on the statistical characterization of the energy landscape[16].

1.5 The statistical view and the energy landscape

When experiments on proteins are performed, the samples containing a macroscopic number of molecules can be considered as thermodynamic systems with well defined properties. A single protein molecule is, instead, too small to be considered a true thermodynamic system and thermodynamic functions are subject to large thermal fluctuations [2]. The same when we deal with the observation of protein folding: it is only possible to measure the fraction of molecules having a certain average configuration and there are large fluctuations about the average configuration. For this reason, when we study the properties of a protein molecule, they have to be understood in a statistical sense and the folding has to be viewed as occurring through an ensemble of structures rather than through only a few uniquely defined structural intermediates [19, 2].

Taking into account this observation, the protein folding is a process that typically occurs at constant pressure and temperature. Under these conditions the *Gibbs free energy* is the natural thermodynamic potential for describing the process[16]:

$$\Delta G = \Delta H - T\Delta S, \quad (1.1)$$

where H indicates the enthalpy while S is the entropy. The free energy is able to describe the protein-solvent system as a function of the configuration of the protein. The form of the free energy as a function of the protein conformation is called the *energy landscape* [16].

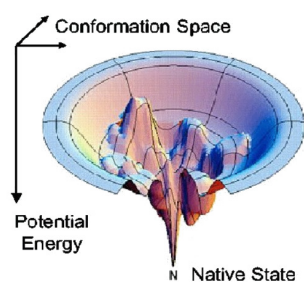


Figure 1.5: Representation of the funnel of the energy landscape. Figure is from [20]

The weak point in Levintal's paradox is that it assumes that all conformations are equally probable during the path from the unfolded state towards the native configuration. In this sense Levinthal's argument considers a free energy landscape that looks like a flat golf course with a single hole at the free energy minimum. However, the argument breaks down for a free energy landscape that looks like a funnel with the energy that decreases when the structures approach the native state, placed at the bottom of the funnel as a minimum, as it appears in Fig. 1.5 [16].

The funnel can also display ruggedness, namely traps that can be popu-

lated transiently by proteins and that are due to the competition between entropy and energy. In order to clarify this aspect the energy landscape of a two-state protein folding is reported in two dimensions in Fig. 1.6:

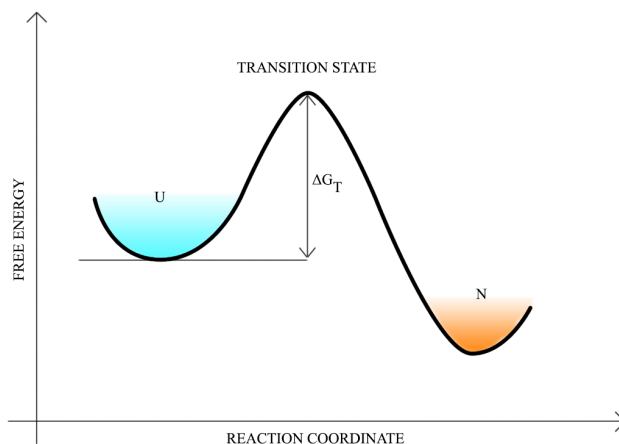


Figure 1.6: 2D representation of the energy landscape of a two-state protein folding.

The free energy barrier between the unfolded state, **U**, and native state, **N**, represents the conflict between the two contributions of enthalpy and entropy to the free energy. The unfolded state is characterized by high energy and high entropy, because it does not present ordered structures and, moreover, the apolar groups in the unstructured chain are forced to stay in contact with water. During the folding the entropy decreases because of the formation of ordered structures that constrain the movements of the chain by restricting the degrees of freedom of the protein. In this way the entropy hinders the prosecution towards the native state, as it is possible to evince by considering Eq. 1.1, while the enthalpy contribution decreases because of noncovalent bonds formations and the avoidance of unfavorable contacts by burying apolar groups inside the protein and away from water. The top of the free energy barrier is called “*transition point*” and represents the stage in which the enthalpy contribution drives the folding and the protein can slip to the final state, characterized by low energy and low entropy [16, 20].

So, the resulted picture presents the protein folding as a process that goes down the funnel-like energy landscape via multiple parallel pathways from the vast majority of individual non-native conformations to the native states placed around the bottom of the funnel. At any stage the protein exists as an ensemble of conformations and can be trapped transiently in many local energy minimum wells [16, 19, 21, 20].

This organization is a result from evolution and natural evolution would

select sequences in which the interactions are not in conflict but instead are cooperative for reaching a low-energy structure. The interactions are said to be “*minimally frustrated*” [16, 19, 21, 20].

The energy landscape theory provides a basis to understand the mechanism of protein folding and the most of proteins conforms to the concepts of funneled energy landscape and minimal frustration [21, 16, 22, 23]. However, it is important to stress that this theory should be considered only a model for describing protein folding and that there exists a part of proteins whose behaviour is not fully and correctly understood with the use of these sole principles. For example, this is the case of knots formation in proteins with knotted native topology as presented in a recent article [24, 25], .

Moreover, for a two-state folding an information that can be obtained by evaluating the free-energy gap between the transition point and the unfolded state, ΔG_T as reported in Fig. 1.6, is the rate at which the folding occurs, namely the rate for diffusing across the barrier. This is obtained by the Kramers formula:

$$k = k_0 e^{-\frac{\Delta G_T}{k_B T}}, \quad (1.2)$$

where T is the temperature at which the folding occurs, k_B is the Boltzmann’s constant and k_0 is a prefactor. This prefactor reads:

$$k_0 = \omega_a \omega_b m D / k_B T 2\pi \quad (1.3)$$

where ω_a and ω_b are the curvature of the free energy function at the bottom and top of the barrier, respectively, D is the particle’s diffusion constant, related to the friction coefficient γ through the Einstein relation $D = k_B T / m \gamma$ [26].

1.6 The study of the protein folding

A number of techniques have been developed in order to deep study and understand every step of the protein folding, both from experimental and theoretical point of view.

In the next chapter, after a brief introduction about experimental methods used to give insight into the issue of protein folding, a detailed description of theoretical techniques, with an emphasis on the approaches that we will use in this thesis, is proposed.

Chapter 2

Experiments and simulations in protein folding

2.1 Experiments and theory: two sides of the same coin

Thanks to the rapid and noteworthy improvements and developments occurred in the recent years for what concerns experimental apparatus, power of computational resources and softwares a more in depth understanding in the field of protein folding even in atomistic detail has been available, giving unprecedented insight into this challenging issue. It is also clear that only combining the experimental and theoretical aspects, it is then possible to reveal more details and overcome suggestions and hypothesis in order to reach a possible certain knowledge. Moreover, one method can give predictions that the other method is able to confirm. This cooperation will appear in all the projects presented in this thesis.

The purpose of the following paragraphs is to present the most common experimental techniques and the theoretical approaches applied to the study of proteins.

2.2 Experimental methods in protein folding

An experimental technique that can elucidate the conformation of proteins is *Nuclear Magnetic Resonance spectroscopy*, or *NMR spectroscopy*. This technique uses the magnetic properties of certain atomic nuclei revealing the atomic structure of macromolecules, and also of proteins, in solutions. It requires large amounts of material that has to be stable at room temperature under a rather long time of data acquisition. Moreover, the investigation with NMR of large molecules becomes challenging for signal overlap, limited

solubility and fast transverse relaxation that causes broader and weaker peaks[27, 28].

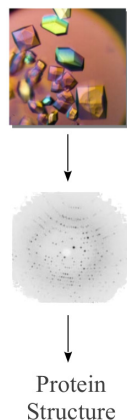


Figure 2.1: X-ray crystallography experiment.

X-ray crystallography is another experimental method able to reveal precise three-dimensional positions of most atoms in a protein. The use of x-rays provides an excellent resolution because the wavelength of x-rays is of the same order of length as that of a covalent bond ($\sim 1\text{\AA}$). The whole apparatus requires a source of x-rays, a detector that collects the x-rays diffracted, and a crystal of the protein of interest that diffracts the x-ray beam, as expressed in 2.1. This last point can be considered the drawback of the technique because of the difficulty of forming a crystal from proteins. Indeed, this process is very time consuming and requires large amount of proteins ordered and aligned, a challenge especially for flexible proteins [28].

NMR and x-ray crystallography allow to elucidate new protein structures each day. All the coordinates provided by these methods are collected at the *Protein Data Bank* [28].

The limitations imposed by NMR and x-ray crystallography can in principle be overcome by *Small-angle x-ray scattering*, SAXS, that, by collecting the patterns to a very small angles, typically a few degrees, allows to obtain the shape and size of the proteins in solution, but not the exact positions of the atoms inside the protein. *Wide-angle x-ray scattering*, WAXS, can in principle extend data present in SAXS but is a low resolution technique and, although it is sensitive in respect of protein secondary and tertiary structure, is not capable of describing the full three dimensional protein conformation and other studies have to be developed in order to clarify correlation between protein structural elements and WAXS profile regions [29].

An important role in mapping the structures of transition states and intermediates in protein folding at the level of individual residues is represented by a protein-engineering method, the Φ -value analysis. A mutation is inserted inside the residue chain and the phi value reads:

$$\Phi = \frac{(\Delta G_W^{TS-D} - \Delta G_M^{TS-D})}{(\Delta G_W^{N-D} - \Delta G_M^{N-D})} = \frac{\Delta\Delta G^{TS-D}}{\Delta\Delta G^{N-D}}, \quad (2.1)$$

where ΔG_W^{TS-D} is the difference in energy between the transition state and the denatured state in the wild type protein, ΔG_M^{TS-D} is the same difference in energy but for the mutant version and ΔG^{N-D} is the energy

difference between the native state and the denatured state, the indices W and M indicates also in this case the wild type and the mutant protein, respectively. This means that Φ value represents the mutation-induced change in the transition state free energy divided by the change in the equilibrium free energy of folding [30]. From this result it is possible to indicate which residues, if mutated, affect the reaction.

It is also important taking into account *Circular Dichroism*, or *CD*, that is based on differences in the absorption of left-handed polarized light versus right-handed polarized light by chromophores which are themselves chiral or are placed in chiral environments. In case of proteins, if the far UV region (240-180 nm or even lower) is analyzed, the measurements rely on the peptide bond absorption and the CD spectrum can give information about the content of secondary structures. If the near UV region (320-260 nm) is investigated then information about the tertiary structure of the protein can be obtained because this region is linked with the amino acid side chains or disulphide bridges [31].

A very useful technique in determining rigidity or flexibility of a protein along a path is represented by *Hydrogen-deuterium exchange technique coupled with mass spectroscopy* and schematized in Fig. 2.2.

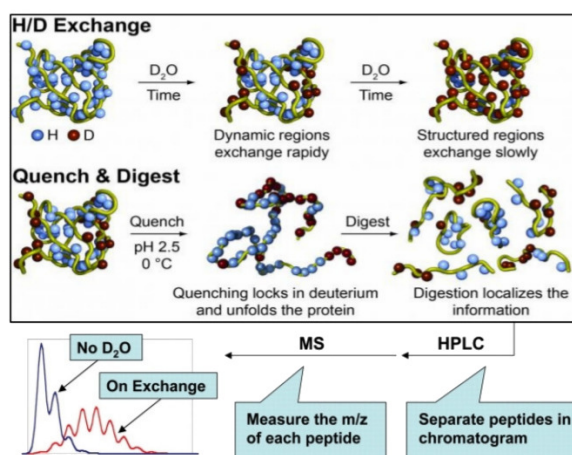


Figure 2.2: Scheme of a typical Hydrogen/Deuterium exchange technique coupled with mass spectroscopy. Figures are from [32, 33]

The studied protein is incubated in D_2O for a fixed time period, in which exchanges between hydrogens, in the protein, and deuterium, in the solvent, take place. After this period the exchange reaction is then quenched by dropping the pH to 2.5 and the temperature to 0°C . Then, the process continues with the digestion of the protein by pepsin. The digest is then loaded

on to a reverse phase HPLC column and eluted directly into an electrospray ionization mass spectrometer. As deuterium is one mass unit heavier than hydrogen, as a consequence of this the extent of deuterium uptake in each fragment can be determined from the mass shift relative the undeuterated sample. In this way it is possible to have information about the solvent-exposure of particular parts of the protein and the structures that are more flexible [34]. Indeed, the higher the quantity of deuterium in a fragment, the more flexible and more exposed to the solvent the fragment itself. This technique has provided important information in case of conformational changes in a particular superfamily of proteins, the serpins, studied in this thesis, and represents a valid reference for all the simulations performed and reported in Chapter Case of study 3: Atomic-level characterization of conformational changes in Serpin family. In case of conformational changes one can use also *FRET*, the *Froester resonance energy transfer*, that let to obtain the distance between two cromophores in a molecule by measuring the energy transferred between the two sites [35].

All these techniques are powerful methods that allow doubtless to go deeper into the protein folding issue. However, a support for understanding some experimentally observed features or a method for giving even new insight into topics where experiments could hardly have a role, for example for too rapid dynamics or large and flexible proteins, is provided by computer simulations, that are used in this work in order to study some aspects related to the protein behavior and folding.

2.3 Computer simulations applied to protein folding: an overview

2.3.1 Methods

Computer simulations are thought to run on a single computer or a network of linked computers and are based on algorithms, that let to study the dynamics or the thermodynamics of the system. The algorithms for the simulation of proteins generally converge in the two main classes of methods: Monte Carlo (MC) and Molecular Dynamics (MD). Monte Carlo is a stochastic method that can be applied for studying systems at equilibrium but also dynamics [36]. The peculiarity of MC simulations is that it is not possible to get any information from the simulation itself about the real time needed for conformational changes that are happening in the simulation. On the contrary, the MC could be an effective and useful method for studying proteins' free-energy surface [37].

In an MD simulation Newtonian equations of motion for all particles in the protein structure model are numerically integrated, once the forces

between different particle types have been defined. The initial velocities are usually chosen from a Maxwell-Boltzmann distribution at the simulation's temperature. The time step for numerical integration is usually about a femtosecond (10^{-15} s) and should be about one or two orders of magnitude less than the system's smallest time scale [37].

In some cases the simulations of slow processes require a computational effort too high for the present computers, so techniques for speeding the processes have been developed. These are the case of the targeted MD, in which external forces guide the system towards the target state. A type of biased simulation is the ratchet-and-pawl MD (rMD) simulations, deeper explained in the dedicated section and coupled with the Dominant reaction pathways (DRP) approach, used in in this work, a molecular dynamics simulation that allows to choose the path that minimizes the error introduced by the bias.

In the following a presentation about the state of the art of computer simulations to date is proposed.

2.3.2 State of the art of the protein folding simulations

If Monte Carlo could represent the preferred choice for elucidating properties of systems at equilibrium because allows the protein to explore efficiently the phase space, for what concerns dynamics the situation becomes more complicate. It is also possible to apply Monte Carlo methods to the study of a dynamics of a process, as it will explained better in the dedicated Chapter, but this could be revealed not an efficient way for investigating such thematic, especially for big size proteins. As an example, we performed coarse grained Monte Carlo simulations based on a $G\bar{o}$ model for reproducing a conformational change occurring in a protein composed by more than 300 residues. Although the $G\bar{o}$ model introduces a bias until the final state weeks of simulation have been used and the applied description turned out to be not sufficient for taking into account all the effects during the process. For this reason in the following we will focalize on a more efficient simulation method for reproducing protein dynamics, either in folding or in conformational changes from one well structured state to another.

From the first reported all-atom molecular dynamics simulation of a protein in water in the picosecond timescale [38, 39] about 30 years have passed and the accessible timescale with such an approach has increased dramatically. Indeed, accurate all-atom molecular dynamics simulations at the millisencond timescale are now possible [40, 41]. Voelz et al. reported in 2010 the first simulation in implicit solvent of the folding of the 39-residue protein NTL9(1-39) with an experimental folding time of ~ 1.5 ms [42], employing a peculiar distributed computing project, called *Folding@Home*, and the *Markov State Models* approach.

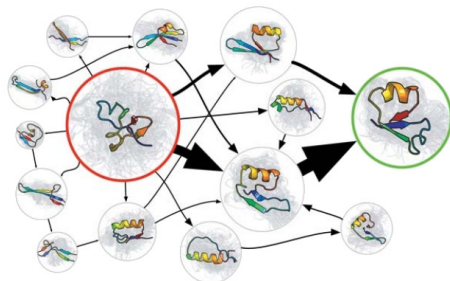


Figure 2.3: MSM for the NTL9 protein. States in better equilibrium are drawn larger, as the arrows for the more likely transitions. The Figure is reproduced with the permission from [42]. Copyright (2010) American Chemical Society.

Folding@home uses the idle resources of personal computers owned by volunteers from all over the world for running molecular dynamics simulations, [43], while Markov State Models (MSMs), an example is reported in Fig. 2.3, are a way for describing all the conformations a protein can explore as a set of states, namely distinct structures, and the transition rates between them. Moreover, MSMs are particularly important as they facilitate parallelization across many computer processors by allowing statistical aggregation of short, independent simulation trajectories, as used by computer networks Folding@home [43].

The Folding@Home distributed computing platform also allows, a few years later, to study through MD the folding of the 86-residue protein ACBP on the 10 ms timescale [44].

Another breakthrough in the issue of the protein folding is represented by the work of Shaw and coworkers, who developed a special-purpose machine, Anton, capable of running biomolecular MD simulations on a microsecond and even millisecond timescale [45, 46, 47, 48, 49]. Indeed, Piana et al. published the simulation of the folding of 76-residue ubiquitin, that occurs in milliseconds [50].

However, many biological processes, such as conformational changes of big-size proteins represent a challenging topic for accurate MD simulations because of the high computational effort required for following slow events and/or rearrangements of proteins with thousands of atoms. In the literature it is possible to find out some examples of these dares. This is the case of the Ras GTPase. The application of accelerated MD [51] allows the observation of the conformational transition related to these proteins [52, 53]. It is important to underline that the simulated system is made up by a chain with less than 200 residues. Another examples are represented by the studies reported on the rearrangements that occur in the protein kinases [54, 55] in microsecond- to millisecond-scale [56, 54], or in the protein GroEL, an ATP dependent molecular chaperone that, despite the big size, has been studied completely in a microsecond-long unbiased MD [57] or with a temperature-accelerated MD [58]. The latter technique has also been used in order to characterize the conformational changes related to HIV-1apt [58].

It is important to stress that this conformational change has been followed also by standard MD, this is a proof of the fact that this dynamics does not exceed the millisecond timescale. Moreover, the authors of the paper are not sure to be able to observe the whole mechanism although the use of the accelerated MD[58]. Then, studies about β_2AR performed in the microsecond-timescale MD simulations should also be taken into account [59].

This review draws a picture of the most important results achieved by computer simulations of protein dynamics and highlights the limits of applicability, for what concerns size of the system and time scale of the reaction, of these studies to date. These observations stress also the challenging aspect of a topic treated in this thesis: the simulation of conformational change occurring in serpin family, a reaction that can take even weeks involving more than 350 residues. These conditions are, as expressed in this paragraph, beyond the limits of all the performed simulations.

We referred to the $G\bar{o}$ model, that represents the interactions inside the protein. Indeed, all the algorithms are applied to a model of the system, which should be able to describe not only the interactions but also the constituents and the environment related to the object of the study. The next Chapters are dedicated to explain the details of the most common models and to present in depth the used Monte Carlo and rMD-DRP methods applied to the treated protein issues.

Chapter 3

Theoretical Models

3.1 Coarse-grained vs. all-atom models

An accurate fully *ab initio* description of proteins in a realistic environment remains well beyond the reach of computers and algorithms to date. As alternatives to this accurate electronic structure calculations a number of models at different level of detail, where not all the degrees of freedom are explicitly considered, have been proposed. There is a theoretical justification for these reduced models and can be found in the separation of time scales in macromolecular systems. To be more precise, the dynamics of a set of slow variables can regulate the behaviour of the system over long time scales. The remaining fast variables equilibrate around each point in the space spanned by the slow variables [60]. Quantum effects involve only fast time scales below the ps. All the dynamics occurring in time scales lower than the ps can be classically treated.

This paragraph proposes the two principal approximation in which a protein and its constituents, the amino acids, can be described: *all-atom* models and *Coarse-grained*, schematized in Fig. 3.1.

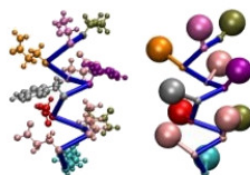


Figure 3.1: All-atom, left, and one-bead coarse-grained, right, description of protein. Figure is from [61]

The former takes into account all the atoms that constitute the amino

acids of the protein and is considered the smallest compromise in which only the electronic degrees of freedom are treated implicitly. This model is based on the Born-Oppenheimer approximation [60] and assumes that the lowest Born-Oppenheimer energy surface can be parametrized in terms of empirical transferable potentials.

For what concerns the latter, on the contrary, it includes models with a more and more increasing degree of resolution, starting from the one-bead approach until four or six-bead approximations. The justification of such models is found in studies that suggest that coarser resolution than complete atomic detail may be suitable for describing large scale protein motion and that, despite the fact that at the microscopic level proteins show a complex network of interactions among a large number of constituents, in the laboratory their dynamics follows a coherent dynamics at the mesoscopic level. Then, studies documented the success of such models in capturing the features of folding [60, 62, 63]. Regarding the coarse grained models, the one-bead model describes the amino acids as beads, as expressed in Fig. 3.2 a, each bead located at the place of the α -carbon, with the mass of the whole amino acid that represents. The beads are linked together via virtual bonds.

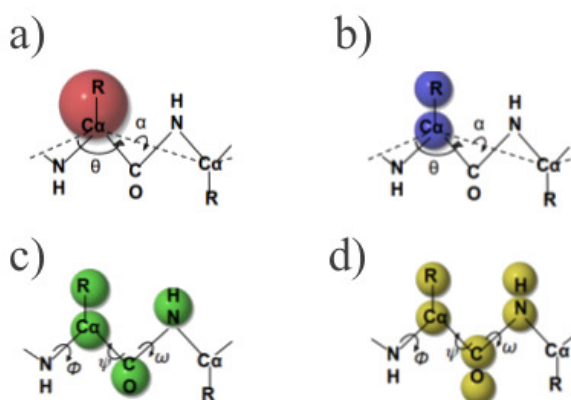


Figure 3.2: Coarse grained models: one-bead, a), two-bead, b), four-bead, c), six-bead, d). Figure is from [64]

Then, it is possible to increase the resolution by adding a second bead on the centroid of the side chain, 3.2 b, or, in the four-bead models, Fig. 3.2 c, the side chain is represented by a single bead, the R , whereas the backbone is described by all its constituents: N , the amide group, C_α , the central carbon, and C , the carbonyl group [65, 66]. Then, a finer representation is given by the six-bead model, Fig. 3.2 d, with simplified hydrogen bond

terms [65].

In conclusion the all-atom approach is clearly more detailed and is associated to a higher accuracy, and consequently desirable, than the coarse-grained one. Nevertheless, its weak point is that it is constrained by computer-power limitations, because more degrees of freedom are included. Instead of a few trajectories often run by all-atom simulations, it can be possible run hundreds of trajectories with a coarse grained approach, which allow extensive statistical analysis [67]. Moreover, a coarse-grained description makes more feasible simulations of systems of large size. However, coarse grained models, although in a more detailed version, cannot properly take into account particular side-chain motions [67].

As these last sentences show, the choice between the two approaches is not immediate and has to be applied on every case of study. For example, as it will be described more in detail in the chapters dedicated to the results, while for pheromones a coarse grained description is able to describe the behavior of the system under different temperatures, the approach fails to describe the conformational changes that occur in the serpin proteins when passing from the native metastable state to the final stable state. In this case only an all atom description with a realistic potential allows the study of this feature.

3.2 Description of water

Proteins live, change and act not in a vacuum but inserted into an environment that provides an important contribution in forming their structure and carrying out their function, a noteworthy example of this is provided by the inhibitory function of the serpin protein towards the protease, which is translocated from the top to the bottom of the serpin and as a consequence of the two opposite forces applied, the serpin pull and the water friction, is distorted and inactivated. It is crucial, though challenging, for the reliability of the simulation to take into account this environment.

It is possible to incorporate the solvent particles explicitly, as in Fig. 3.3. This means that a model of each solvent molecule is constructed and used to solvate the system. The interactions are calculated by taking into account not only the solute but also all solvent molecules located in the simulation box: the potential is then given by the sum of contributions of the constituents of the solute and all molecules of the solvent. In case of water there exist different types of explicit descriptions: they divide in simple models, flexible models and polarisable water models. In the first case the water molecules are rigid and interact via Coulomb and Lennard Jones potential. The interactions are between sites of the molecule, whose number determine the type of the model: for example, a three site model, as

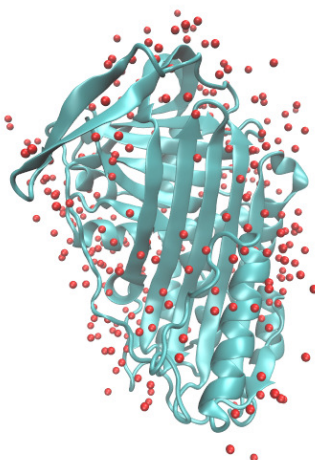


Figure 3.3: Explicit solvent, red points, around a protein.

expressed in Fig. 3.4 on the left, where the sites correspond to the atoms, a four-site model that places the negative charge on a dummy atom (labeled M in Fig. 3.4 in the middle), and improves the electrostatic distribution around the water molecule, or a five-site description that locates the negative charge on dummy atoms representing the lone pairs of the oxygen atom, Fig. 3.4 [68, 69].

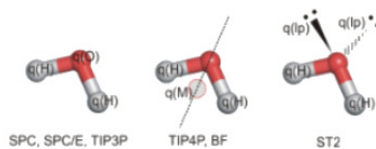


Figure 3.4: Simple explicit water models: three-site, on the left, four-site, in the middle, and five-site, on the right. Figure is from [68]

In the second explicit atom model internal conformational changes are included. Finally, the third model is for inhomogeneous systems with considerable contribution of the electrostatic field of ions and other charged groups [68]. Typically, the number of solvent particles required for such a model is of the order of thousands even for a short chain. For this reason the solvent molecules, if explicitly represented, introduce a high degree of accuracy but can on the other hand make simulations rather inefficient from the computational-cost point of view [70].

An alternative is offered by the implicit solvation, that decreases the computational effort and eliminates the need for the equilibration of water

around the solute. Another advantage related to the use of implicit solvent is the estimation of the free energy: since solvent degrees of freedom are taken into account implicitly, estimating the free energy of solvated structures is more direct than with explicit water models [71]. Indeed, in this way “the noise” given by local minima arising from small variations in solvent structure is eliminated [72].

This approach is based on replacing real water environment consisting of discrete molecules by a continuum medium with the dielectric and properties of water [70].

The total energy of a solvated molecule can be written as expressed in eq. 3.1 [70]:

$$E_{tot} = E_{vac} + \Delta G_{solv}, \quad (3.1)$$

where E_{vac} represents molecule’s potential energy in vacuum and ΔG_{solv} is defined as the free energy of transferring the molecule from vacuum into solvent. The problem is represented by the last term and the following approximation is made [70]:

$$\Delta G_{solv} = \Delta G_{pol} + \Delta G_{nonpol}, \quad (3.2)$$

where ΔG_{pol} is the free energy of first removing all charges in the vacuum and then adding them back in the presence of a continuum solvent environment and $\Delta G_{nonpolar}$ is the free energy of solvating a molecule from which all charges have been removed [70].

Different approaches have been used in order to calculate these terms, but the most often used model is the *Generalized Born model*, or GB model. In the GB model each atom of a molecule in solution is represented as a sphere of radius ρ_i with a charge q_i at its center. The interior of the atom is assumed to be filled uniformly with material of dielectric constant 1. This group of atoms is set in spherical cavities embedded in a polarizable dielectric continuum [73] with high dielectric value ϵ , 80 for water at 300 K.

For what concerns G_{pol} , under the generalized Born model it assumes the form [70]:

$$G_{pol} = \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^n \sum_{j>1}^n \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp\left(\frac{-r_{ij}^2}{4b_i b_j}\right)}}, \quad (3.3)$$

where q_i and q_j are pair of charges separated by a distance r_{ij} in a solvent of dielectric constant ϵ with a Born radii b_i and b_j , respectively. The Born radius of an atom reflects its degree of burial inside the molecule: if an ion is insolated b_i is equal to its VDW radius ρ_i , while if it is deeply buried $b_i \gg \rho_i$ [72].

Regarding the non-polar part, $\Delta G_{nonpolar}$ of eq. 3.2, this term is the sum of G_{cav} and G_{vdw} . ΔG_{vdw} is the solute-solvent van der Waals term and the cavity term, ΔG_{cav} , is the free energy required to form a cavity in the solvent large enough to accommodate the solute. A common approximation of these terms is to assume $\Delta G_{nonpolar}$ proportional to the total solvent accessible area (SASA) of the molecule [70]:

$$\Delta G_{nonpolar} \approx \sigma \times SASA, \quad (3.4)$$

with the proportionality constant derived from experimental solvation energies of small non-polar molecules [70].

In the DRP simulation we usually use the GB model, so, for a more complete treatment of the approach see the Appendix The generalized Born model.

3.3 Realistic all-atom force field

A force field is constituted by an expression for the potential energy function and all the parameters used in that function [74]. The sets of parameters are obtained by comparison with experiments and quantum mechanical calculations. There exists a class of force field that are also called *realistic force field* based on interactions between residues regardless of their contact in the native state. Among them there are force fields applied to all-atom systems, AMBER [75, 76], CHARMM [77], GROMOS [78], OPLS [79] just for citing some used for studying proteins, or to coarse grained models, for example VAMM (Virtual atom molecular mechanics) that relies with beads placed at the central C_α position [80] and MARTINI, initially thought for simulating molecular dynamics in lipids but extended later also to proteins, works by modelling on average four heavy atoms by a single grain bead [81]. In this section we will focus on two all-atom force fields, AMBER and CHARMM, particularly interesting and promising in protein simulations. The AMBER force field has been elaborated in the 80's and in the years it has been upgraded and improved. Amber's general form reads:

$$\begin{aligned} V(r_1, \dots, r_N) = & \sum_{bonds} k_b (l - l_0)^2 + \sum_{angles} k_a (\vartheta - \vartheta_0)^2 \\ & + \sum_{torsions} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] \\ & + \sum_{j=1}^{N-1} \sum_{i=j+1}^N \left[\epsilon_{i,j} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right], \end{aligned} \quad (3.5)$$

where the first, second and third terms are the contributions of the bonds, angles and bond rotations while the fourth term represents the non-bonded energy between all atom pairs, that is decomposed in van der Waals and electrostatic energies. r_{0ij} is the equilibrium distance. Amber force field describes explicitly all hydrogen atoms.

In earlier versions of AMBER charges are derived from quantum chemistry calculations at the Hartree-Fock STO-3G level via fitting of partial atomic charges to the quantum electrostatic potential. Then, thanks to the ability to use larger basis sets the charges have been fitted at the Hartree-Fock 6-31G* level. This basis set tends to overestimate bond-dipoles in respect of the gas phase values in an amount comparable to that in empirical water models SPC or TIP3. Overpolarization is expected to be a consequence of electronic polarization in liquids [82].

In DRP simulations we used an improved version of AMBER, AMBER ff99SB-ILDN, that has optimized the torsion potentials of backbone and side chains and presents an improved helix-coil balance since the previous versions have been demonstrated to over-stabilize α -helical peptide conformations [75, 76, 83]. This potential has been successfully applied to study protein folding, where other potentials have failed: an example is represented by the simulation of FiP35, a protein constituted by two β hairpins, performed by Shaw and coworkers [45]. AMBER ff99SB-ILDN shows also a well agreement with NMR data [83].

Another force field that shows reasonably good agreement with experimental results is an improved version of CHARMM force field, the CHARMM22*, used in recent works [50]. Also CHARMM22* shows improvements in the optimization of torsion parameters and in the helix-coil balance. This force field is constructed by taking into account the description of the bonds, angles and bond rotations, the Lennard-Jones and Coulomb terms and two additional contributions: the Urey-Bradley term, $\sum_{Urey-Bradley} k_u (u - u_0)^2$, that describes an interaction based on the distance, u , between atoms separated by two bonds, and an improper dihedral term, which is used to maintain planarity, $\sum_{impropers} k_\omega (\omega - \omega_0)^2$, where $\omega - \omega_0$ is the out of plane angle.

3.4 $G\bar{o}$ model

The $G\bar{o}$ model was originally pioneered by $G\bar{o}$ and coworkers [84]. Since it defines a function for the potential energy and a set of parameters it is a force field, however, it differs from the realistic force fields described above. In this model the residues, that are in contact in the native state, are defined to have a favorable interaction energy. The $G\bar{o}$ -model picture may be justified in terms of the minimal-frustration principle and funneled energy

landscape concept because the use of such a potential lowers the energy of the native conformation relative to all other conformations, resulting in an energy landscape which is smoother and less “frustrated” than the analogous energy landscape corresponding to a potential built from “generic” beads [22].

The $G\bar{o}$ -models have been used in a number of studies at various levels of resolution, from one-bead coarse grained description to even an all-atom one.

In the most cases such models are able to reproduce and explain the details of the folding of a number of proteins, since the majority of studied proteins have been demonstrated to conform to the principle of minimal frustration [21, 16, 22, 23], while they clearly fail in those systems whose folding mechanism is driven at least partially by non native interactions.

Among the different types of $G\bar{o}$ model a useful method is that proposed by Karanicolas and Brooks [85], implemented in this work, that overcomes the limitations of the previous models in which the favorable potential applied to all native contacts was identical in energy, in order to discern the degree to which topology alone could determine mechanisms of protein folding. Such models are unable to distinguish between the folding mechanisms of proteins with similar topologies, since any information about types of residues is neglected. Karanicolas and Brooks designed a model with the maximal degree of information by incorporating some effects derived from sequence [85].

The model elaborated by Karanicolas and Brooks is applied on the one-bead coarse grained description of the protein. The interaction energy of residues separated in sequence by three or more bonds and that are in contact in the native state reads [85]:

$$V_{ij} = \varepsilon_{ij} \left[13 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 18 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} + 4 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \quad (3.6)$$

where r_{ij} is the distance between residues i and j , σ_{ij} is the distance between i and j at which the interaction energy is a minimum, and $-\varepsilon_{ij}$ is the strength of the interaction at this distance. ε_{ij} depends on the type of the interaction, namely hydrogen bonds between backbone atoms and side chain-side chain interactions, an effect of sequence because in this way it is possible to take into account the type of residues.

In respect of a typical Lennard-Jones potential the formulation in Eq. 3.6 shows an increase in the curvature of the potential around the minimum and a small energy barrier, as showed by the comparison between the two potentials in fig. 3.5. The added small barrier is thought to take into account the “desolvation penalty”, which pairs of residues must pay before

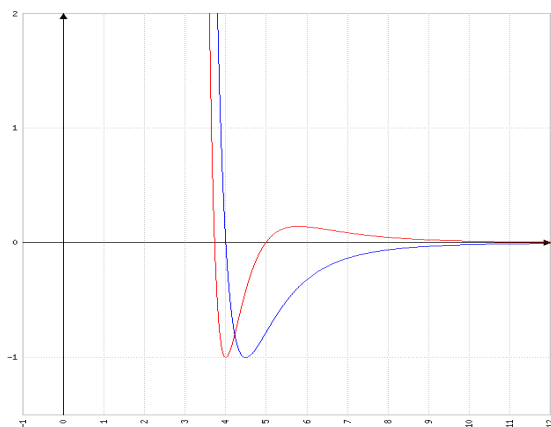


Figure 3.5: Comparison between a typical Lennard-Jones potential, blue line, and the $G\sigma$ potential, red line.

a favorable contact can be formed. As it is possible to observe in Fig. 3.5 the potential energy function of Eq. 3.6 is steeper near the minimum than that of a Lennard-Jones interaction. The reason for this choice is in the observation that, if two side chains are separated by a distance of $\sigma_{ij} = 4\text{\AA}$, it could be that the corresponding C_α are separated by a distance of, say, $\sigma_{ij} = 8\text{\AA}$. By using this value in a typical Lennard-Jones potential the result is a much broader well than that obtained with $\sigma_{ij} = 4\text{\AA}$. With the formulation of Eq. 3.6 it results a narrower potential.

Since the residues in contact in the native state are set to have a favorable interaction energy and the strength of the interaction depends on the type of contact, it is important to accurately construct a map of the native contacts defining their nature. The hydrogen bonds between $C = O$ of residue i and $N - H$ of residue j are searched with the method of Kabsh and Sander [86]: the electrostatic interaction energy between two H-bonding groups by placing partial charges on the $C, O(+q_1, -q_1)$ and $N, H(-q_2, +q_2)$ atoms is calculated. This formula reads:

$$E = f q_1 q_2 \left(\frac{1}{r_{NO}} + \frac{1}{r_{HC}} - \frac{1}{r_{HO}} - \frac{1}{r_{NC}} \right), \quad (3.7)$$

where $q_1 = 0.42e$ and $q_2 = 0.20e$, being e the unit electron charge, and $r(AB)$ is the interatomic distance from A to B, see Fig. 3.6 for a better explanation of the used distances.

r is in angstroms, $f = 332 \text{\AA} \text{ kcal}/e^2\text{mol}$ is the dimensional factor and E is in kcal/mol. The H-bond is assigned if E is less than a cutoff value: $-0.5\text{kcal}/\text{mol}$. Each pair of residues in contact in the native state through

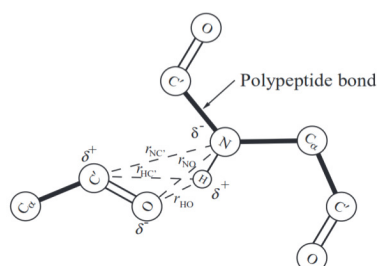


Figure 3.6: Distances used for calculating Coulomb hydrogen bond energy, from [87].

a H-bond interacts via the potential in Eq. 3.6 with ε_{ij} set to the strength of a given hydrogen bond and σ_{ij} set to the α -carbon separation distance of this pair in the native-state structure.

In cases where a hydrogen bond was assigned to a pair in which a native contact had already been defined, four additional weak native contacts were defined. If residues i and j interact via either two hydrogen bonds or a hydrogen bond and a side-chain contact, then the pairs $(i-1, j)$, $(i, j-1)$, $(i, j+1)$, $(i+1, j)$ are also defined as native contacts. In this case ε_{ij} is set to $0.25 \times E_H$, where E_H is the energy of the given hydrogen bond, σ_{ij} is the same as the previous case [85].

Side chain-side chain native contacts were assigned by collecting all pairs of residues that have at least one pair of heavy atoms in their side-chains closer than 4.5\AA . By evaluating Eq. 3.6 ε_{ij} value is set proportional to the corresponding Miyazawa-Jernigan [88] contact potential for the particular pair of residues, renormalized in order to match the hydrogen bond native contact energy scale.

For what concerns residues that are not in contact in the native state they are subject to a potential of the type of a hard-core one of the form [85]:

$$V_{ij} = \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} \right], \quad (3.8)$$

where σ_{ij} is calculated by averaging the radii of the residues involved. The radius is defined for each residue as the distance to the closest residue that is not defined as native contact. Then, ε_{ij} for all pairs of residues not in contact in the native state is set to $1.5 \times 10^{-3} \varepsilon_{res}$ and $\varepsilon_{res} = 0.0054 T_f$ kcal/mol. T_f is the temperature where the native state and the unfolded state are equally populated. For a derivation of these last terms, see Appendix $G\bar{o}$ model: the unfolding temperature and the native contact energy. It is to be observed that the treatment of water in such coarse grained description

of the proteins is included indirectly in the parameters, which can be a radius that describes the propensity of the residues to be surrounded by other residues, as in the hard-core like contribution, or the distance between residues in native contact or the MJ terms: in fact, these values describe, though indirectly, effects of hydrophobic forces that cause the residues to assume particular positions.

The potential is completed with the addition of harmonic potentials applied to each of the virtual bonds, Eq. 3.9, angles, Eq. 3.10, and dihedral terms reflecting the preferences of the backbone dihedral angles imposed by size and geometries, Eq. 3.11. Also in this last term effects of sequence are taken into account.

$$V(r_{ij}) = \frac{1}{2}k(r_{ij} - r_0)^2, \quad (3.9)$$

where r_{ij} is the distance between consecutive C_α , $k = 378 \text{ kcal/mol}\text{\AA}^2$ is the elastic constant and $r_0 = 3.8\text{\AA}$ the equilibrium distance.

$$U_{angle}(\theta_{ijk}) = -\frac{1}{\gamma} \log \left[e^{-\gamma[k_\alpha(\theta_{ijk} - \theta_\alpha)^2 + \varepsilon_\alpha]} + e^{-\gamma[k_\beta(\theta_{ijk} - \theta_\beta)^2]} \right], \quad (3.10)$$

where θ_{ijk} is the pseudo-angle formed by residues i , j and k . $\theta_\alpha = 92^\circ$ and $\theta_\beta = 130^\circ$ are the equilibrium values of the helical and the extended pseudo-angles, respectively.

$$U_{torsion}(\varphi_{ijkl}) = \sum_{n=1}^4 [1 + \cos(n\varphi - \delta_n)V_n], \quad (3.11)$$

where φ_{ijkl} is the dihedral angle between the planes identified by the position of the beads i , j , k and j , k , l . The constants δ_n and V_n depend on the type of residues identified by the label j and k .

After this description of the $G\bar{o}$ model of Karanicolas and Brooks applied on a coarse grained protein it is also important mention that there is a trend that prefers apply native topology-based potentials on high structural resolution models: the most representative one is the all-atom $G\bar{o}$ -model, which explicitly models all the heavy atoms except hydrogens and uses $G\bar{o}$ -like potentials for shaping the interactions between atoms [89].

3.5 $G\bar{o}$ model with the addition of non- $G\bar{o}$ interactions

In the previous paragraph we presented a version of the $G\bar{o}$ -model as elaborated by Karanicolas and Brooks. In this model only the native interactions

are treated as favourable attractive in the context in which the folding of proteins, at least small proteins, is thought to be energetically biased and minimal frustrated towards the native state.

However, there are systems, whose folding or conformational changes could not be explained with the use of the sole favorable native interactions.

In this work we used the model developed by Kim and Hummer [90], in which non- $G\bar{o}$ energy specified by the identity of the residues has been added. In the improved $G\bar{o}$ -model proposed by Karanicolas and Brooks the type of residues has been taken into account for what concerns the capacity of the chain to move (dihedral terms) and in defining the contribute of the native contacts. All these terms are not able to consider any favorable contribute of residues not in contact in the native state or interactions of electrostatic nature between all the residues. In the model of Kim and Hummer the non bonded part of the potential consists of the native and the non- $G\bar{o}$ contributions. The last is divided into non-native and electrostatic interactions:

$$V_{nb} = \sum_{\text{native}(i,j)} V_{G\bar{o}}(r_{ij}) + \sum_{\text{non-native}(i,j)} V_{ng}(r_{ij}) + \sum_{\text{all}(i,j)} V_{elec}(r_{ij}), \quad (3.12)$$

where $V_{G\bar{o}}(r_{ij})$ is the contribution of residues that are in contact in the native state as developed by Karanicolas and Brooks. $V_{ng}(r_{ij})$ is a modified Lennard-Jones potential and $V_{elec}(r_{ij})$ is the Debye-Hückel type expression.

The Lennard-Jones term depends on the amino acid type of the residues i and j and can be attractive or repulsive. Repulsive interactions are applied between amino acid pairs that interact less favorably with each other than with the solvent and *vice versa*. For pairs of residues that experience an interaction with a strength $\epsilon_{ij} < 0$, the non native interaction potential is given by:

$$V_{ng}(r_{ij}) = 4|\epsilon_{ij}| \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \quad (3.13)$$

where $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, being σ_i and σ_j the van der Waals radii of the residues i and j . For pairs of residues that repel each other, so that $\epsilon_{ij} > 0$, the non-native potential energy function takes the following form:

$$V_{ng}(r_{ij}) = \begin{cases} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + 2\epsilon_{ij}, & r_{ij} < \sqrt[6]{2}\sigma_{ij} \\ -4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], & r_{ij} \geq \sqrt[6]{2}\sigma_{ij} \end{cases} \quad (3.14)$$

The LJ strenghts ϵ_{ij} are defined in this way:

$$\epsilon_{ij} = \lambda (e_{ij} - e_0), \quad (3.15)$$

where $e_{ij} (<0)$ is the MJ [88] contact potential between residues i and j , while e_0 is an offset parameter that balances the preference of residue-residue interactions relative to residue-solvent interactions, in this point appears the role played by water. The parameter λ scales the strenght of the LJ interactions compared to the physical electrostatic interactions. As an illustration of the effect of this contribution, the effective interaction strength between the neutral hydrophilic residues GLU and GLN is 0.08 kcal/mol, so that effective interaction is repulsive, while for the hydrophobic residues ILE and LEU is -0.46 kcal/mol, so that the effective interaction is attractive [24].

For what concerns the electrostatic interaction between residues i and j :

$$V_{elec}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 D} \frac{\exp\left[-\frac{r_{ij}}{\xi}\right]}{r_{ij}}, \quad (3.16)$$

where q_i and q_j are the electrostatic charges of residues i and j , ξ is the Debye screening length $\simeq 10\text{\AA}$, ϵ_0 is the dielectric vacuum constant and D is the relative dielectric constant of water, set at 80.

An analysis that is based on simulations that arise from purely $G\bar{o}$ model and from models that have both the contributions from native and non native interactions could be useful in order to understand the possible role played by these interactions in determining the folding events towards the native state.

Chapter 4

Monte Carlo method applied to protein folding

4.1 The Monte Carlo approach: an introduction

4.1.1 The method

Monte Carlo (MC) approaches form a large and important class of numerical methods based on random numbers used for solving problems of systems at equilibrium or even of dynamics towards equilibrium.

Equilibrium MC [36] methods simulate the random thermal fluctuation of the system at equilibrium from state to state. Let suppose that the system is in state μ . If $R(\mu \rightarrow \nu)dt$ is the probability that the system is in state ν a time dt later and $w_\mu(t)$ is a weight that represents the probability that the system will be in state μ at time t , the evolution of $w_\mu(t)$ assumes the form [36]:

$$\frac{dw_\mu}{dt} = \sum_\nu [w_\nu(t)R(\nu \rightarrow \mu) - w_\mu(t)R(\mu \rightarrow \nu)], \quad (4.1)$$

where the first term on the right side is the rate at which the system is undergoing a transition into state μ , the second term is, on the contrary, the rate at which the system is undergoing a transition from state μ into another state. If the two terms cancel one another for all μ then the rates $\frac{dw_\mu}{dt}$ will all vanish and the weights are taken constant for the rest of the time. This is a system at equilibrium. The weights at equilibrium are called the equilibrium occupation probabilities [36]. In 1902 Gibbs showed that, for a system in thermal equilibrium with a reservoir at temperature T , the equilibrium occupation probabilities are:

$$p_\mu = Z^{-1} \exp[-\beta E_\mu(x)], \quad (4.2)$$

where E_μ is the energy of the state μ , $\beta = 1/kT$ being k the Boltzmann's constant, and Z is the partition function $Z = \sum_\mu e^{-\beta E_\mu}$. The probability distribution expressed in Eq. 4.2 is known as the Boltzmann distribution [36].

The usual aim of MC simulations is to calculate, given an observable Q , such as internal energy for example, the expectation value $\langle Q \rangle$. The expectation value for an observable for a system in equilibrium reads:

$$\langle Q \rangle = \frac{\sum_\mu Q_\mu e^{-\beta E_\mu}}{\sum_\mu e^{-\beta E_\mu}}. \quad (4.3)$$

The average is obtained over all states μ of the system. This is tractable in the very smallest of systems, in larger we can only do this average over a subset of the states. Monte Carlo technique works by choosing a subset of states at random from a probability distributions p_μ . If M states μ_1, \dots, μ_M are chosen the best estimate of Q is [36]:

$$Q_M = \frac{\sum_{i=1}^M Q_{\mu_i} p_{\mu_i}^{-1} e^{-\beta E_{\mu_i}}}{\sum_{j=1}^M p_{\mu_j}^{-1} e^{-\beta E_{\mu_j}}}. \quad (4.4)$$

Q_M is called the *estimator* of Q . The greater M is the more accurate the estimate is and, when $M \rightarrow \infty$ $Q_M = \langle Q \rangle$ [36].

The *importance sampling* is the technique for picking out the important states among the large number of possibilities and the most common form of this method is taking a sample of states in which the likelihood that any particular one appears is proportional to its Boltzmann weight [36]. In this way the estimator of Q becomes:

$$Q_M = \frac{1}{M} \sum_{i=1}^M Q_{\mu_i} \quad (4.5)$$

According to the central limit theorem the average calculated on statistically independent values Q_{μ_i} converges for increasing M to a gaussian distribution. σ_M can be used to estimate the uncertainty of the estimator [36].

$$\sigma = \sqrt{\frac{\overline{Q^2} - \overline{Q}^2}{M-1}}, \quad (4.6)$$

where \overline{Q} is the average of Q over M states, namely the estimator of Q , as defined in Eq. 4.4.

The tricky part of performing a Monte Carlo simulation is the generation of an appropriate random set of states according to the Boltzmann probability distribution. The basis of such simulations is represented by the so

called *Markov processes*. In a Markov process, given a system in one state μ , a new state of that system ν is generated. This process occurs according to a probability, called the *transition probability*, that should not vary over time and should depend only on the properties of the current states μ and ν . Moreover, the transition probability of passing from μ to ν must also satisfy:

$$\sum_{\nu} P(\mu \rightarrow \nu) = 1, \quad (4.7)$$

since the Markov process must generate some state ν when handed a system in the state μ [36].

In a MC simulation the Markov process is used repeatedly to generate a *Markov chain* of states and it is chosen so that when it is run for long enough starting from any state of the system it will eventually produce a succession of states with probabilities given by the Boltzmann distribution. In order to achieve this, two further conditions on the Markov process are needed: the condition of ergodicity and detailed balance [36]. The former is the requirement that it should be possible for the Markov process to reach any state of the system from any other state, if it runs for long enough. The latter ensures that, after the system has come to equilibrium, the generated configurations are according to the Boltzmann probability distribution, rather than any other distribution. This constraint is introduced:

$$\frac{P(\mu \rightarrow \nu)}{P(\nu \rightarrow \mu)} = \frac{p_{\nu}}{p_{\mu}} = e^{-\beta(E_{\nu} - E_{\mu})} \quad (4.8)$$

Another fundamental concept in MC methods for simulating states according to the Boltzmann distribution is the *Acceptance ratio* [36], that allows to obtain the desired set of transition probabilities from any algorithm is chosen. The transition probability is divided into two parts:

$$P(\mu \rightarrow \nu) = g(\mu \rightarrow \nu)A(\mu \rightarrow \nu), \quad (4.9)$$

being $g(\mu \rightarrow \nu)$ the selection probability, namely the probability, given an initial state μ , that the algorithm will generate a new target state ν , and $A(\mu \rightarrow \nu)$ is the acceptance ratio. The acceptance ratio says that if, in running the simulation, the algorithm generates a new state ν from a starting state μ , the state should be accepted and the system changed in the new state ν a fraction of the time $A(\mu \rightarrow \nu)$. The rest of the time the system should stay in the state μ [36].

4.1.2 The Metropolis algorithm

For what concerns the implementation of the Monte Carlo method the most famous and widely used algorithm, applied also in the work presented in this thesis, is the *Metropolis algorithm* [36]. In this algorithm the selection probabilities $g(\mu \rightarrow \nu)$, as appear in Eq. 4.9, for each of the possible states ν , are all chosen to be equal. Therefore, the constraint imposed by the detailed balance, Eq. 4.8, takes the form:

$$\frac{P(\mu \rightarrow \nu)}{P(\nu \rightarrow \mu)} = \frac{g(\mu \rightarrow \nu)A(\mu \rightarrow \nu)}{g(\nu \rightarrow \mu)A(\nu \rightarrow \mu)} = \frac{A(\mu \rightarrow \nu)}{A(\nu \rightarrow \mu)} = e^{-\beta(E_\nu - E_\mu)} \quad (4.10)$$

What remains is now correctly choose $A(\mu \rightarrow \nu)$ to satisfy Eq 4.10. The way to maximize the acceptance ratios, namely to produce the most efficient algorithm, is to give the larger of the two ratios the largest value possible, 1, and then adjust the other to satisfy the constraint. Let us suppose, of the two states μ and ν , μ has the lower energy and ν the higher: $E_\mu < E_\nu$. Then the larger of the two acceptance ratios is $A(\nu \rightarrow \mu)$, so we set that equal to 1. As a consequence of this choice, $A(\mu \rightarrow \nu)$ must then take the value $e^{-\beta(E_\nu - E_\mu)}$ [36]. The Metropolis algorithm can be summarized as follows:

$$A(\mu \rightarrow \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{if } E_\nu - E_\mu > 0 \\ 1 & \text{otherwise} \end{cases} \quad (4.11)$$

In this way, the implementation of the Metropolis algorithm can be divided into the following steps:

1. Pick a configuration μ
2. Pick a trial configuration ν
3. If the state ν has an energy lower than or equal to the present one it should always be accepted
4. If it has a higher energy it should be accepted with the probability given in Eq. 4.11.
5. Start again from point 2 with the new state (or, if rejected, the old state) and repeat for N times.

The choice of the trial configurations is essential for the success of the method. In the following the moves for obtaining a new configuration applied to the particular cases of protein thermodynamics and dynamics used in this thesis are presented.

4.2 Application to protein folding

We used the Monte Carlo method based on the Metropolis algorithm applied on a coarse grained description of the protein. The potentials implemented are according to the $G\bar{o}$ and $G\bar{o}$ -Hummer models.

To sample the ensemble of chain conformations at thermal equilibrium, we have used a combination of different types of local and global trial moves, namely:

- crankshaft moves [91] applied to one residue: a randomly selected single bead is rotated around the axis defined by its nearest neighbors. The angle of the rotation is randomly selected in the interval $\delta\varphi_{max} = \pm 30^\circ$.
- end-point moves: the last 10 residues on both terminals are rotated rigidly with respect to the rest of the chain by up to 30° around a random axis passing through the most interior bead of the end segment.
- Cartesian moves: a randomly selected bead is displaced, within a sphere with radius of 0.015 nm,
- pivot moves [92]: one amino acid is picked at random and the chain portion involving all amino acids with smaller (or alternatively larger) sequence index are rotated by up to 30° around a random axis passing through the picked amino acid. This move is a global one while the other are more local.

The trial moves were accepted or rejected according to the standard Metropolis criterion. The boldness and the relative probability of the different moves was set in order to have a global acceptance ratio close to 50%. It is clear that the global moves are crucial for reaching the equilibrium, since they involve larger conformational changes in the protein.

However, a different approach is required when, instead of the system at equilibrium, the topic is the analysis of the dynamics for reaching an equilibrium. If one is interested in the behaviour of the system at equilibrium all what is required is that the equilibrium distribution of states sampled by an algorithm is the correct Boltzmann distribution and the conditions of ergodicity and detailed balance ensure this. These conditions say nothing however about the way in which the system comes to equilibrium and it is clear that various choices of the dynamics of the algorithm will make a difference here [36].

In the particular project, the Monte Carlo method has been also applied to the study of conformational changes in serpin family, that lead the protein to reach the final state from a metastable one. In this case, for simulating the dynamics of a protein, but a more general description of the dynamics

of a polymer is also valid, we set the probability of global moves to zero and implement in the program only the local moves. In this way the obtained scheme is analogous to dynamics performed with Verdier-Stockmayer kink-jump algorithm combined with an out of plane crankshaft move (KJC) [91].

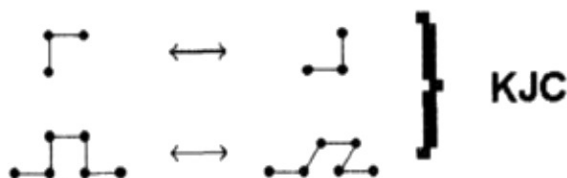


Figure 4.1: Scheme of the moves in the KJC algorithm. Figure is from [93]

The KJC algorithm, although it has been shown that it is nonergodic and thus inappropriate for studying static quantity (for this the addition of global moves, verified to be ergodic [92], is required), little is known about the size of the effect of the nonergodicity. However, KJC, thanks to the local character of the moves, is used to study dynamic properties, since it is thought to mimic the intrinsic dynamics of a polymer in solution [93]. The same local character of the moves identifies also our choices for generating trial configurations, in this way we are confident to be able to simulate properly the events towards the serpin stable state. The method that implements local moves has been also successfully applied to the study of the dynamics related to knotted proteins [24].

Chapter 5

Dominant Reaction Pathways approach

5.1 The overdamped Langevin equation

The Dominant Reaction Pathways approach (DRP) [94, 95, 96, 97] is an approximate method which can be used to simulate the reactive dynamics in systems obeying the overdamped Langevin equation.

The generalized Langevin equation describes the dynamics of a system coupled to a heat bath [96]. In order to introduce the topic a one-dimensional system composed by N atoms is proposed but all the discussion can be also extended to a three-dimensional one. If x_k is the cartesian coordinate of the k -th atom and the friction force acting on the particle is given by Stoke's law $f_f = -\gamma v$, then Langevin's equation takes the form:

$$m_k \ddot{x}_k = -\gamma_k \dot{x}_k - \nabla U(\mathbf{X}) + \xi(t), \quad (5.1)$$

where $\mathbf{X} = (x_1, x_2, \dots, x_N)$ is a vector specifying all the atomic coordinates, $U(\mathbf{X})$ is the potential energy and $\xi(t)$ is a Gaussian random force with zero average that takes into account the collisions of the particles with the molecules of the medium. $\xi(t)$ satisfies the fluctuation-dissipation relation:

$$\langle \xi_k(t) \xi_k(0) \rangle = 2D_k \delta(t) \quad (5.2)$$

The acceleration term on the left in Eq. 5.1 introduces effects that are damped on a time-scale $t \gtrsim m/\gamma_k \equiv \tau_D$, that results by taking the Fourier transform of Eq. 5.1:

$$-m\omega^2 \tilde{x} + i\omega\gamma \tilde{x} + \tilde{F}(\tilde{x}) + \tilde{\xi}(\omega) = 0 \quad (5.3)$$

and

$$\tilde{x} = \frac{\tilde{F}(\tilde{x}) + \tilde{\xi}(\omega)}{m\omega^2 - i\omega\gamma} \quad (5.4)$$

The damping time scale results from searching the values of ω for which the acceleration vanishes, in particular $|i\omega\gamma| \gg |m\omega^2|$. It results: $\gamma \gg m\omega$, hence $\tau \gg m2\pi/\gamma$. Our study is focused on the dynamics of an amino acid chain for which the damping time scale $\tau_D = m2\pi/\gamma$ is of the order of a fraction of ps, much smaller than the microscopic time scale associated with the dynamics of torsional angles, which takes place at the *ns* time scale [96, 98]. So, for $t \gg \tau_D$ Langevin equation reduces to:

$$\frac{\partial x_k}{\partial t} = -\frac{1}{k_B T} D_k \nabla U(\mathbf{X}) + \eta_k(t), \quad (5.5)$$

where $D_k = \frac{k_B T}{\gamma_k}$ is the diffusion coefficient, k_B is the Boltzmann's constant, T is the temperature of the heath-bath and $\eta_k(t) = \frac{1}{\gamma_k} \xi(t)$.

5.2 Path integral representation and its application to the DRP problem

The probability of finding the system in a conformation \mathbf{X}_f at time t_f starting from a conformation \mathbf{X}_i at t_i and evolving with the overdamped Langevin equation 5.12 is given by the Fokker-Plank equation [94, 96]. This probability can be written in terms of a path-integral representation: [94]:

$$P(\mathbf{X}_f, t_f | \mathbf{X}_i, t_i) = e^{-\frac{U(\mathbf{X}_f) - U(\mathbf{X}_i)}{2k_B T}} \int_{\mathbf{X}_i}^{\mathbf{X}_f} D\mathbf{X}(\tau) e^{-S_{eff}[\mathbf{X}(\tau)]}, \quad (5.6)$$

where

$$S_{eff}[\mathbf{X}(\tau)] = \int_{t_i}^{t_f} d\tau \left(\frac{\dot{x}_k^2(\tau)}{4D_k} + V_{eff}[\mathbf{X}(\tau)] \right), \quad (5.7)$$

is the so-called effective action. V_{eff} is the effective potential and reads:

$$V_{eff} = \frac{D_k}{4(k_B T)^2} \left[(\nabla_k U(\mathbf{X}))^2 - 2k_B T \nabla_k^2 U(\mathbf{X}) \right]. \quad (5.8)$$

The DRP approach is based on the saddle-point approximation of the path integral. The saddle-point defined as *dominant* reaction pathways are those that maximize the exponential part $e^{-S_{eff}}$ in 5.6, that is, those that minimize the effective action functional [100]. They are solutions of the classical equations of motion generated by the effective action [96]:

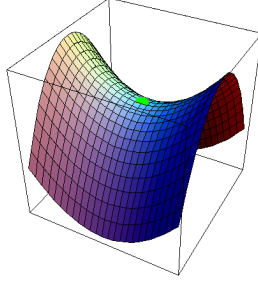


Figure 5.1: For protein folding the saddle point approximation is valid. Figure is taken from [99]

$$\ddot{\mathbf{X}} = 2D\nabla V_{eff}(\mathbf{X}) \quad (5.9)$$

In principle it is possible find the dominant paths by numerically minimizing the effective action functional [101]. In order to do this a discretization rule has to be defined. A choice that is commonly adopted in numerical simulation is the so-called Ito prescription. However, depending on the stage at which the discretization is carried out the action takes different functional forms [102]. A discretization at the level of Eq. 5.7 gives the action [102, 101]:

$$S_1[\mathbf{X}] = \Delta t \sum_{i=1}^{N_t} \left[\sum_{k=1}^N \frac{(x_k(i+1) - x_k(i))^2}{4D_k \Delta t^2} + V_{eff}[\mathbf{X}(i)] \right], \quad (5.10)$$

where X denotes a path, the index $i = 1, \dots, N_t$ runs over the time steps, the index $k = 1, \dots, N$ runs over the atoms in the protein, $V_{eff}(\mathbf{X})$ is the effective potential in discretized form [101]:

$$V_{eff}(\mathbf{X}) = \frac{1}{4(k_B T)^2} \sum_k D_k \left(|\nabla_k U(\mathbf{X})|^2 - 2k_B T \nabla_k^2 U(\mathbf{X}) \right), \quad (5.11)$$

On the other hand, the discretization could also be performed at the level of Eq. 5.5, namely [102]:

$$x_k(i+1) = x_k(i) - \frac{\Delta t D_k}{k_B T} \nabla U[\mathbf{X}(i)] + \sqrt{2D_k \Delta t} R_k(i), \quad (5.12)$$

where the index i is referred to the i -th time step, $i+1$ indicates the following time step after the i -th, Δt is an elementary time interval and R_k is a

Gaussian noise vector with zero average and variance $\langle R_k(i)R_k(0) \rangle = \delta(t)$. Since the distribution $P(R_k(i)) = e^{-R_k^2(i)/\sqrt{2\pi}}$ and the force only depends on the previous position $[x_1(i)\dots x_N(i)]$, the probability $P(x_k(i+1)|x_k(i); \Delta t)$ can be obtained by changing the variables from $R_k(i)$ to $x_k(i+1)$. For this reason, since $P[X] \equiv e^{-S_2}$, the action becomes:

$$S_2[\mathbf{X}] = \sum_{i=1}^{N_t} \sum_{k=1}^N \frac{1}{4D_k \Delta t} \cdot \left(\mathbf{x}^k(i+1) - \mathbf{x}^k(i) + \frac{\Delta t D_k}{k_B T} \nabla_k U[i] \right)^2. \quad (5.13)$$

In this equation the index $i = 1, \dots, N_t$ runs over the different time-step in the trajectory, the index $k = 1, \dots, N$ runs over the atoms in the protein, k_B is the Boltzmann's constant and D_k is the diffusion coefficient of the k-th atom.

S_1 and S_2 are two legitimate prescriptions for the discrete representation of the propagator, Eq. 5.6. It has been argued that the actions are equivalent when evaluated over diffusive, nondifferentiable trajectories as generated by Eq. 5.5 [102], while the two actions lead to different values when evaluated over differentiable paths. Moreover, it is suggested that in direct minimization study of diffusive trajectories the action S_1 should be favored over S_2 [102].

For this reason, a strategy to find the path with highest probability consists in directly numerically minimizing in each time-step the action functional S_1 , Eq. 5.10. However, in case of a protein folding transition with large number of time steps needed, it is very challenging to directly minimize the effective action in Eq. 5.10, because of the high computational effort that arises from the presence of the the numerical calculation of the laplacian of the potential energy. Indeed, the time interval Δt has to be chosen typically of the order of $1fs$. For this reason the number of degrees of freedom to be simulated and minimized is huge: $3N \times N_t$, where N is the number of atoms in the protein and N_t the total number of steps, typically at least 10^6 [94, 101, 103].

A possible solution is given by observing that the system defined by the effective action 5.7 conserves the ‘‘effective energy’’ [103]:

$$E_{eff} = \frac{1}{4D} \dot{x}^2(t) - V_{eff}[x(t)]. \quad (5.14)$$

Hence, for any fixed pair of native and denaturant configurations the dominant paths can be equivalently found by minimizing an effective Hamilton-Jacobi (HJ) action in the form [94]:

$$S_{HJ} = \sum_{i=1}^N \sum_{k=1}^N \Delta l_{i,i+1} \sqrt{\frac{1}{D_k} (E_{eff} + V_{eff}[\mathbf{X}(i)])}, \quad (5.15)$$

being $\Delta l_{i+1,i} = \sqrt{(\mathbf{X}(i+1) - \mathbf{X}(i))^2}$ the elementary displacement in configuration space. The parameter E_{eff} determines the time at which any given frame l of the path is visited [94], according to:

$$t_f - t_i = \int_{\mathbf{X}_i}^{\mathbf{X}_f} dl \sqrt{\frac{1}{2(E_{eff} + V_{eff}[\mathbf{X}(l)])}}. \quad (5.16)$$

The advantage of adopting the HJ formulation is that it is possible, in this way, to replace the time discretization with the discretization of the curvilinear abscissa l , which measures the Euclidean distance covered in configuration space during the reaction. So, the gap in the time scales vanishes and only about 10^2 frames are usually sufficient to provide a convergent representation of a trajectory [103, 101]. However, the direct minimization of the HJ using relaxation algorithms does not provide a practical strategy for investigating the folding of typical polypeptide chains. Indeed, it was observed that the folding pathways exploration is extremely slow and that, consequently, the trajectories are found to remain highly correlated to the initial trial path, for an exceedingly long time. For this reason the recent version of DRP method, implemented in the simulations used in this work, uses the S_2 functional [101, 104] by finding the dominant trajectory according to a statistical analysis based on scoring *a posteriori* the relative likelihood of each computed folding pathways sharing the same boundary conditions and obtained with a bias potential, explained in the next section. The method calculates for each step the probability, namely the weight, for each path X to be realized in the unbiased over-damped Langevin dynamics, that reads:

$$Prob[\mathbf{X}] \propto e^{-\sum_{i=1}^{N_t} \sum_{k=1}^N \frac{1}{4D_k \Delta t} \cdot \left(\mathbf{x}^{k(i+1)} - \mathbf{x}^{k(i)} + \frac{\Delta t D_k}{k_B T} \nabla_k U[i] \right)^2}, \quad (5.17)$$

the indices and constants are the same as in Eq. 5.13. The probability in Eq. 5.17 is estimated at each simulation step for a set of trajectories and at the end of the simulation the path with the maximum value, namely the minimum action, is chosen as dominant.

5.3 rMD

A set of trial all-atom trajectories connecting from a given initial configuration towards the native state is generated by using the MD velocity Verlet

algorithm and by applying the force field AMBER ff99SB-ILDN. The DRP algorithm, described in the previous paragraph, is used to identify, with the minimum Onsager-Machlup functional, the most probable path in each set of trial trajectories sharing the same boundary conditions. Moreover, in order to generate an ensemble of trial trajectories connecting the initial state to a final one, both given as input to the program based on the DRP, we used the *ratchet-and-pawl* MD (rMD) algorithm [105, 106], implemented by introducing a time-dependent bias potential $V_R(\mathbf{X}(t))$ that makes very unlikely for the system to evolve back to previously visited configurations, exerting no work on the system when it spontaneously proceeds towards the native state [101]. The rMD algorithm requires a generalized coordinate that gives a measure of the vicinity of the instant configuration to the native state. We chose, as from [105], a collective coordinate (CC) which measures the distance between the contact map in the instantaneous configuration $\mathbf{X}(t)$ and the contact map in the native configuration \mathbf{X}^{native} :

$$z[\mathbf{X}(t)] \equiv \sum_{i < j}^N \left[C_{ij}[\mathbf{X}(t)] - C_{ij}[\mathbf{X}^{native}] \right]^2, \quad (5.18)$$

where C_{ij} is a continuous representation of the contact map defined as:

$$C_{ij}[\mathbf{X}] = \left[1 - (r_{ij}/r_0)^6 \right] / \left[1 - (r_{ij}/r_0)^{10} \right], \quad (5.19)$$

where $r_0 = 7.5\text{\AA}$ is a fixed reference distance.

The biasing potential introduced is defined as:

$$V_R(\mathbf{X}, t) = \begin{cases} \frac{k}{2}(z[\mathbf{X}(t)] - z_m(t))^2, & \text{for } z[\mathbf{X}(t)] > z_m(t) \\ 0, & \text{for } z[\mathbf{X}(t)] \leq z_m(t), \end{cases} \quad (5.20)$$

where $z_m(t)$ is the minimum value assumed by the collective variable z along the trajectory, up to time t . In equation 5.20, k is the so-called ratchet constant and its value depends on the system.

In the original formulation of the rMD algorithm [105], the variable $z_m(t)$ is updated only when the system visits a configuration with $z[\mathbf{X}(t + \delta t)] < z_m(t)$. With this choice, z_m monotonically decreases during the course of the simulation. As in the work of [25], in order to escape from kinetic traps, we chose to weaken the effect of the bias by allowing the system to back-track along the direction defined by CC. We obtained this by occasionally updating z_m also when it increases with a probability given by a Metropolis acceptance/rejection algorithm. To be more precise, z_m is updated to $z'_m = z[\mathbf{X}(t + \delta t)] > z_m(t)$ if:

$$\exp\{-\beta \cdot k_a \cdot [0.3(z[\mathbf{X}(t)] - z_m(t)) + 2.0(z[\mathbf{X}(t)] - z_m(t))^3]\} > \eta, \quad (5.21)$$

where $\eta \in [0, 1]$ is a random number sampled from a uniform distribution, $\beta = 1.0$, and k_a , which depends on the simulated system, is a parameter given as input of the program.

Each trial trajectory consisted of a number of steps of rMD with a nominal integration time step that is usually $\Delta t = 1$ fs.

5.4 MD vs. rMD-DRP

Protein folding and conformational changes are rare thermally activated processes, this means that the system must overcome a free energy barrier, with a rate that depends on the temperature. In general, for rare events $k_B T / \Delta G \ll 1$, where ΔG is the barrier height.

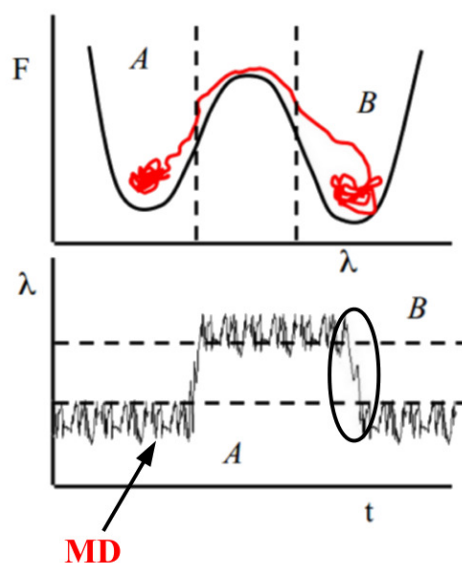


Figure 5.2: Rare events: representation of the free energy as a function of the order parameter, above, and the order parameter as a function of time. Figure is from [107].

As expressed in Fig. 5.2, as a consequence of the fact that protein folding and conformational changes are rare events the system spends a major fraction of time on the bottom of the free energy basins. This generates a decoupling of time scales, since the microscopic time scales associated to local conformation rearrangements are many orders of magnitude smaller than the

macroscopic time scale associated to the inverse reaction rate [103]. Pure Molecular Dynamics (MD) simulations waste most of the computational time in simulating local thermal fluctuations in metastable configurations [103] and, despite the increasing in computational power and software, processes characterized by slow dynamics are to date beyond the limits of recent MD simulations, as expressed in the paragraph about the state of the art of MD simulations.

On the contrary, the rMD algorithm based on DRP approach has the advantage that it does not waste the CPU time in thermal oscillations [108], allowing to investigate reactions so far out of reach. Moreover, the DRP approach has been validated by comparison with coarse grained MD simulations based on a $G\bar{\sigma}$ -model [109] and more recently with the results from the special purpose supercomputer Anton[101], giving results in very good agreement.

However, the final native structure of the protein is required, in this sense the rMD-DRP method is not able to predict the result of folding but focuses its attention on the events that lead to the final state and possibly on the reasons that cause the misfolding. Moreover, a disadvantage of the use of a bias potential in the DRP simulations is that any information about the mean first passage time in the absence of a bias is not available. The bias, indeed, has the effect of distorting the time sloping down the free energy landscape towards the native state. However, as it will be described in the Chapter dedicated to Serpins and their conformational changes, it is possible to estimate a ratio of the rates of reactions between proteins that differ in a few amino acids but that are identical in structures and explore similar configurations during the path, as it is expected that in such systems the bias would work in a similar manner.

Chapter 6

Case of study 1: Folding process of the immunity proteins IM7 and IM9

6.1 Introduction to the problem

After the description of models and methods applied to the study of protein folding, we present in this section the results provided by simulations of the folding of immunity proteins IM9 and IM7.

6.1.1 Intermediate states during folding

This chapter deals with the folding mechanism associated with two proteins: IM7 and IM9. The observations and results are taken from the manuscript: “**Folding process of the immunity proteins IM7 and IM9**”, G. Cazzoli, P. Faccioli, F. Wang, P. Wintrode, in preparation.

The importance of this study is related to the investigation of the nature of the interactions that stabilize eventual intermediate states. This analysis is to be placed in the big picture that has the purpose of understanding how proteins can achieve the final configuration from unfolded states.

As said in the chapter “Proteins”, the concepts of funneled energy landscape [21, 4] and minimal frustration [21, 16] have been introduced for clarifying this aspect, at least for small proteins.

As a consequence of these principles the proteins’ landscape is expected to be smoother than that of vitreous systems and leading to a minimum identified with the native state. Possible displayed ruggednesses in the landscape are explained by a nonuniform compensation of the entropy and energy changes upon forming native contacts [110]. In presence of large

ruggednesses that can trap the protein by hindering the process towards the native state, long-lived intermediate states that energetically compete with the native state become populated [111]. The existence of such intermediates, called topology-driven, could in principle be predicted from the native structure alone by using funneled landscape [112, 22, 113, 114], namely, in this case a $G\bar{o}$ model can be successfully applied.

However, as stressed in first chapter of this work, purely topology driven models are not unfallible in every cases and, although improved in order to take into account the maximal degree of information [22], turn out not to completely explain folding, conformational changes and presence of intermediate states. In this situation the role played by favourable but nonnative interactions is then considered [115, 116]. Among the systems, in which the folding events and the nature of the driven interactions are not clear, we find the colicin immunity binding proteins of *Escherichia coli*, a family of small four-helix proteins with sequence similarity around 50% [117]. In Fig. 6.1 two proteins of this family, IM7 and IM9, have been aligned, in order to note the shared topology. In Fig. 6.2 are indicated the secondary structures.

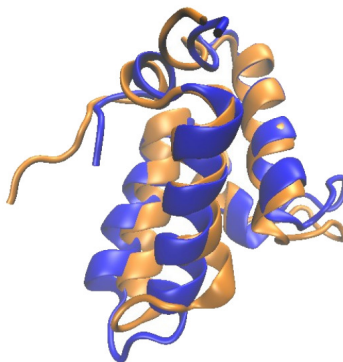


Figure 6.1: Sequence alignment of IM9 (PDB code: 1IMQ.pdb), colored in orange, and IM7 (PDB code: 1CEI.pdb), colored in blue. The alignment has been performed with SuperPose [118]

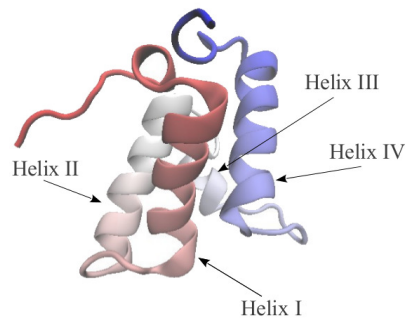


Figure 6.2: Immunity protein IM9. The 4 helices are pointed out.

6.1.2 Experiments and simulations try to characterize IM7 and IM9 folding

While, for IM7, it is well accepted that the protein folds via an intermediate state [119, 120], for IM9 there is no agreement regarding the possible presence of an intermediate during its folding. In particular, experiments observe an intermediate for IM9 only under acidic conditions [121], giving rise to two opposite positions: the existence of the intermediate in every case also at neutral conditions but too rapid to be detected, suggested by some experiments [121] and simulations [22], or the possibility of inducing the intermediate state only by lowering the pH keeping a two-state folding mechanism in the other cases [122, 123].

Another debated question is related to the nature of the interactions that stabilize the intermediate state: it has been proposed that a topology-driven approach fails to describe the behaviour of immunity protein during folding [110] and that the interactions that play a role in this step are a combination of native and nonnative interactions [117, 124, 122, 110]. However, another study performed by Karanicolas and Brooks by running molecular dynamics simulations based on a $G\bar{\sigma}$ -model [22], predicts the folding transition states of a number of proteins, including the topologically analogous proteins IM7 and IM9, leading to results in agreement with experiments and suggesting in this way the primary role of native interactions in determining most protein folding transition states [22]. These simulations based on the $G\bar{\sigma}$ potential allow to hypothesize the presence of an intermediate state in both IM7 and IM9.

Also for what concerns the structures that characterize the intermediate or, more in general, the transition state ensemble (TSE) of both the proteins IM7 and IM9, a general agreement has not yet been reached. The obtained Φ -values tell us that in both IM9 [125] and IM7 [117] the folding of Helix III

is not so important in the intermediate state (low values) than the folding of the other helices, Helix I, II and IV, whose related Φ -values are higher, although a coarse grained MD simulation based on an improved $G\bar{o}$ model [22] hypothesizes a partial folding of Helix IV.

Regarding Helix-Helix docking the high Φ -values suggest that the interface between Helix I-Helix IV is structured in both proteins[126] as well as that between Helix II-Helix IV in IM7 showed by diffusion single-pair fluorescence resonance energy transfer[127]. However there is not well agreement with simulations, since MD simulations restrained by experimental Φ -values [128] and coarse grained MD simulations based on the $G\bar{o}$ -model [22] do not find a native-like nature of the interfaces between Helix II-Helix IV and Helix I/II-IV, respectively. Moreover, for what concerns the interface between Helix I-Helix II the Φ -values reflect the formation of native contacts in this region for IM7 and IM9 but less than in the interface Helix I-IV[126]. On the contrary, a fully native like nature in this region is observed in coarse grained simulations [22].

Finally, a study [129] suggests the role of the residual structures in unfolded state in contributing to the folding mechanism of IM7.

Clarify these aspects could mean understand the limits of the predictability of folding. So, the proposed work has been structured in the following manner: firstly we performed all-atom simulations with AMBER 99SB-ILDN force field based on rMD algorithm and DRP approach in order to investigate IM7 and IM9 folding mechanism. By applying this method, we avoid the insertion of any experimental information with the only exception of the final state. Moreover, our simulation is all-atom with a realistic force field, in this way the possible nonnative interactions are taken into account.

Once considerations around the presence or absence of an intermediate in IM9, the confirm of such intermediate in IM7 and the structural characterization of this state have been derived from DRP results, a comparison with works found in literature is carried out. In this way the possible comprehension of involved interactions is expected. For completing the study the role played by starting states has been analyzed by generating different unfolded states for folding.

6.2 Methods

6.2.1 PDB files preparation

We have performed all-atom simulations with AMBER 99SB-ILDN force field based on rMD algorithm and DRP approach in implicit solvent.

The targets of rMD simulations are provided by the energy minimized structures of immunity proteins IM9 and IM7, PDB codes 1IMQ.pdb and

1CEI.pdb, respectively. In order to verify the role of particular secondary structures during folding we carried on also simulations on a mutant version of IM7: PDB code 2k0d.pdb. Energy minimization has been performed with Gromacs molecular dynamics package 4.5.5 [130, 131] in implicit solvent. The number of minimization steps was set to 50,000 steps and the force field AMBER ff99SB-ILDN was used.

Regarding the beginning structures of the simulations the denatured proteins have been obtained in two ways: one by performing a molecular dynamics simulation at high temperature, 1600K, starting from the PDB folded structures and followed by a MD simulations at T=300K in order to relax the system, the other derived by applying the program NMRPipe [132] that provides a set of random coils from a given PDB. The obtained random coils have been then energy minimized by Gromacs.

6.2.2 Parameters for rMD simulations

First of all we performed trial simulations in order to find the parameters that regulate the ratchet potential, k setted to 1×10^{-3} eV, and the rate at which configurations represented by a higher variable z than the reference z_m are accepted, k_a setted to 100.

For IM7, IM9 and mutant version of IM7 60 sets of trail trajectories have been generated, running for 15×10^4 steps of MD with a nominal integration time step of $\Delta t = 0.5$ fs, performed using a Velocity-Verlet algorithm and coupled to a Nose'-Hoover thermostat.

6.2.3 RMSD, or *Root-mean-square-deviation*

RMSD value is a measure of the difference between a given configuration, as taken from a simulation, and the reference and is calculated in Å: $RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n [(x_i - x'_i)^2 + (y_i - y'_i)^2 + (z_i - z'_i)^2]}$, being n the number of points in the system and the coordinates with and without prime index the position of the $n - th$ point in the reference and in the particular configuration, respectively. In the following analysis the native state is taken as reference.

We chose to limit the analysis only to the successful DRP folding trajectories, namely those that achieve at the end of simulation a RMSD to native < 3 Å

6.2.4 Calculation of the fraction of native contacts

In the following treatment the fraction of native contacts along the DRP trajectories has been calculated. We adopted a criterion according to which two residues with index difference ≥ 3 are said to be in contact if the distance of their C_α is less than 7.5 Å.

6.2.5 Definition of the “Kinetic free energy”

On the basis of how many times a given configuration, defined by the values of the fraction of native contacts, is visited by the DRP a representation in analogy with the free-energy landscape plots is obtained. This is achieved by computing the logarithm of this value normalized by the total number of configurations and then multiplied by $-kT$, becoming an energy in unit of kcal/mol. This quantity has been reported as a function of the FNC , total and/or restricted to a specific area. It is important to stress that the result is not a pure free-energy because of the characteristics of the DRP simulations: there is the ratchet algorithm that biased the system to reach the target and the simulations do not work at equilibrium. If a configuration is never visited, then we decided to force the analogous of the free-energy to assume a finite value, that depends on the number of data and is chosen to be higher than the other calculated values of “kinetic free energy”. This choice is only for sake of clarity and regions that reach this value must be considered to be not-allowed zones. In this sense this quantity, that in the following we will call the *kinetic free energy* could be considered a proper indicator of the regions most visited by the trajectories and consequently of the possible intermediates. This analogy of the free energy has been previously used in Ref. [25] with success.

6.3 Results

6.3.1 Characterization of the folding process in IM9 and IM7

In this first stage of investigation we study the paths starting from **unfolded thermal conformations** until native state. Our purpose is to verify the possible presence of the intermediate state in both IM7 and IM9 and, if established the existence of this intermediate, to characterize this state for what concerns structures and stabilizing interactions.

In Fig. 6.3 a representation of the kinetic free energy for the two immunity proteins as a function of the fraction of total native contacts is reported. There are three minima visible for both proteins. The first minimum at $FNC_{tot} \sim 0.2$ corresponds to the unfolded state. The second minimum, at $FNC_{tot} \sim 0.7$ for IM7 and $FNC_{tot} \sim 0.6$ for IM9, is an intermediate state, the third minimum the native state. So, our first consideration is that the intermediate state is visible in *both* the proteins, as proposed by Ref.[22, 121].

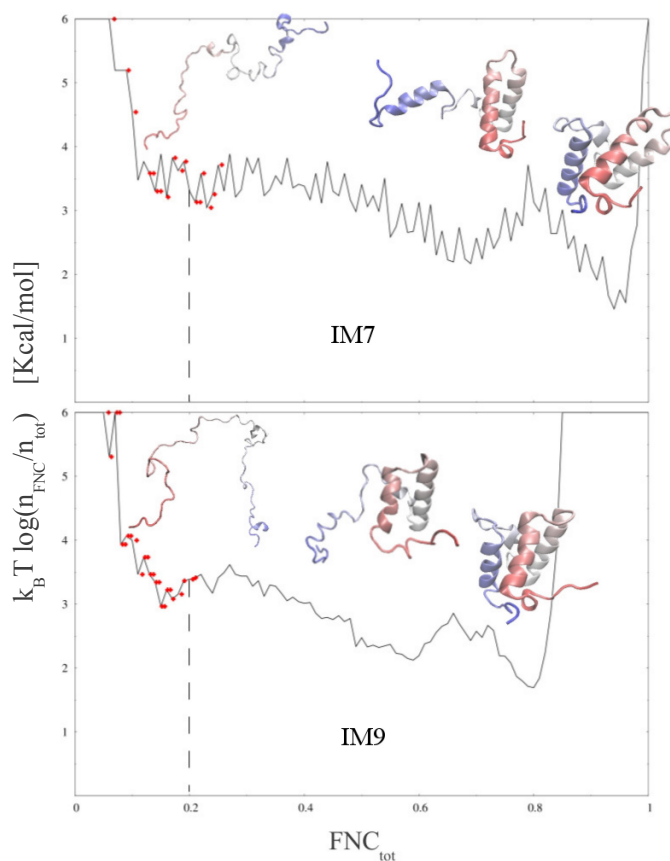


Figure 6.3: Kinetic free energy as a function of the fraction of native contacts for Im7, top, and IM9, bottom. DRP configurations representative of the unfolded, intermediate and final states are reported. A dashed line shows the region with $FNC_{tot} \sim 0.2$, a reference for correctly visualize the rMD-DRP simulation starting configurations, red points, as after the relaxation step.

In order to characterize these observed intermediate states we projected the DRP trajectories on the plane defined by the total fraction of native contacts and the fraction of native contacts restricted to the interfaces between pairs of helices. In this way configurations are defined and the kinetic free energy is calculated. With the term “*fnc restricted to the interface*” we

consider all the native contacts that stabilize the single helices and those between the two helices. The attention has been focused on the pairs: Helix I-Helix II (figures 6.4 a) and d) for IM7 and IM9 respectively), Helix I-IV (figures 6.4 b), for IM7, and e), for IM9) and Helix II-IV (figures 6.4 c), IM7, and f), IM9). For this analysis, we chose not to report the interactions that involve Helix III because this helix is composed by a lower number of residues, 7 residues, in respect of the other helices, more than 15 residues. For this reason we expect that Helix III and the related changes during folding will not have a relevant contribution to the compute of the native contacts at the interface with another, bigger, helix. In order to understand the picture and to better characterize the structures in the intermediate it is important to visualize also the folding of the single helices in respect of the fraction of native contacts, as expressed in Fig. 6.5.

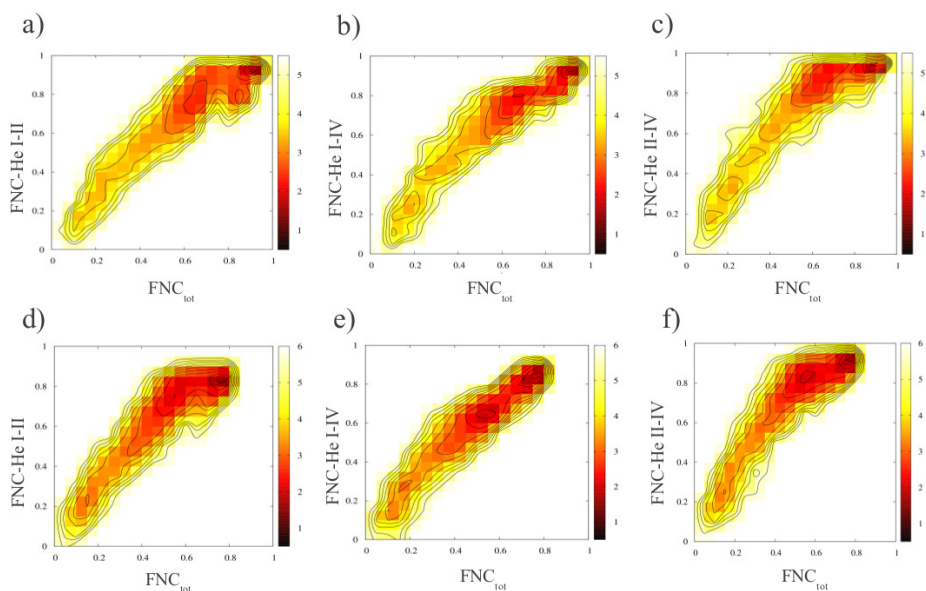


Figure 6.4: Kinetic free energy calculated on the plane formed by the fraction of total native contacts and the fraction of native contacts at the interface between Helix I- II, a) and d) for IM7 and IM9 respectively, Helix I- IV, b) for IM7 and e) for IM9, Helix II-IV, c) for IM7 and f) for IM9.

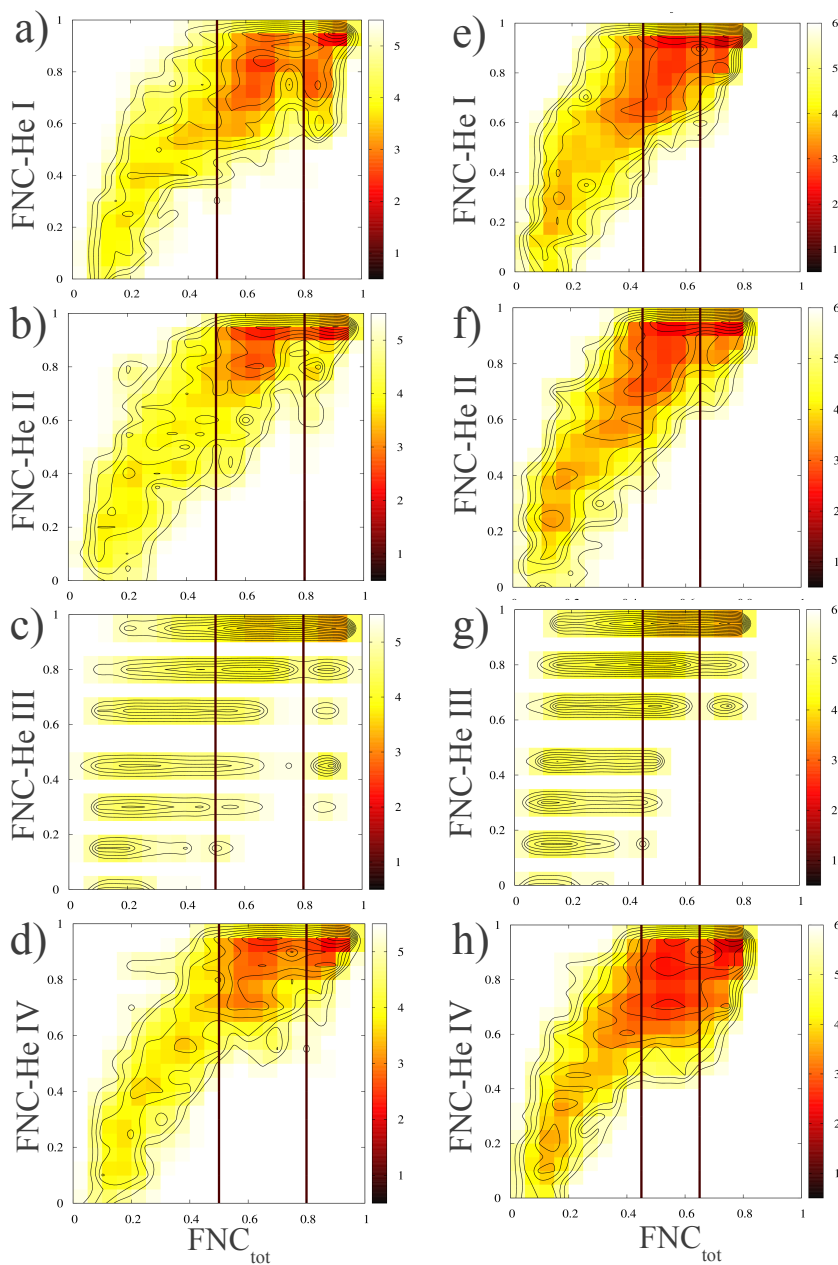


Figure 6.5: Kinetic free energy on the plane formed by the fraction of total native contacts and the fraction of native contacts of Helix I, a) and e) for IM7 and IM9 respectively, Helix II, b) for IM7 and f) for IM9, Helix III, c) for IM7 and g) for IM9, Helix IV, d) for IM7 and h) for IM9. The vertical black lines identify the range in which the intermediate state is found

From Figures 6.4 and 6.5 it is possible to hypothesize that folding process and the intermediate state have the same characteristics for both IM9 and IM7. Moreover, the results let us suppose that the intermediate state is characterized by the formed helices I and II. These features are in agreement with experimental Φ -values [125, 117] and MD simulations [22, 126, 128]. Regarding the docking of these helices between each other we observe native contacts in the region between Helix I and Helix II, since the fraction of native contacts referred to this interface goes from 0.7 to 0.9. This native-like nature is also reflected by the results of MD coarse grained $G\bar{o}$ -model simulations [22]. The Φ -values show that native contacts are formed but that they are less preserved than those in the interface between Helix I-Helix IV [126].

We note instead that the region composed by Helix I, Helix IV and interface is, at the intermediate of IM9, only partial formed, indeed the FNC is around 0.6. The same trend but with a more structured interface is observed at the intermediate of IM7, in this case FNC goes from 0.6 to 0.8. These observations and the representations of the folding of single helices in Fig. 6.5 d) and h), suggest for Helix IV a misalignment in IM7, since Fig. 6.5 d) indicates a formed helix, and a misalignment and/or a possible partial folding in the case of IM9, since 6.5 h) shows a broader distribution of conformations, in perfect agreement with the results obtained by a purely $G\bar{o}$ -model [22]. Φ -values referred to this regions indicate a structured interface [126]. However, it is likely that these values are not so sensitive to be able to take into account the distribution of FNC lower, but not too low (as the unfolded state), than the values in the native state.

A native-like nature is noted also in the case of the interface between Helix II-IV, refused by [22, 128] but observed on the contrary by FRET experiments [127].

For what concerns Helix III, the discretized region in Fig. 6.5 c) and g) is explained by considering that the helix is composed by only 7 residues. Nevertheless, this picture can give an idea of the behaviour of the helix at the intermediate: it fluctuates, because FNC restricted to the helix goes from 0.2 to 1.0, without assuming a particular value. The conclusions can be that helix III does not affect the intermediate state and is in agreement with the low Φ -values [125, 117]. Simulations based on the $G\bar{o}$ -model predict the unfolding of Helix III at the intermediate, although we observe that this Helix can be indifferently in the folded and unfolded state. These differences are only apparent because the meaning of all these results is that Helix III does not have a role in stabilizing the intermediate.

Therefore, it is possible to say that our results are in agreement with those proposed by simulations performed by Karanicolas et al. [22], who used a purely $G\bar{o}$ -model for predicting the existence of an intermediate state in *both* proteins. This state results to be qualitatively similar to that pre-

dicted by our DRP simulation with some exceptions that do not change the substance, indeed for both the studies Helix I and II are formed and native-like, Helix IV not native-like in the sense that it is misaligned or not well formed and the structure of Helix III does not have a role in stabilizing the intermediate. So it seems likely that the intermediate state is stabilized by native interactions.

6.3.2 IM7 mutant

With the aim to verify the irrelevance of Helix III in stabilizing and characterizing the intermediate, experiments have been performed on a mutant version of IM7 [133]. In this mutant residues are added in the region of Helix III. The purpose of this part was to test whether Helix III seems not to participate at the stabilization of the intermediate state because of the small size and the consequent low propensity to form native contacts [133]. The study found that the addition of residues does not affect the behaviour of Helix III. We continued this project and performed DRP simulations on this mutant. Fig. 6.6 shows kinetic free energy plotted versus the fraction of total native contacts formed.

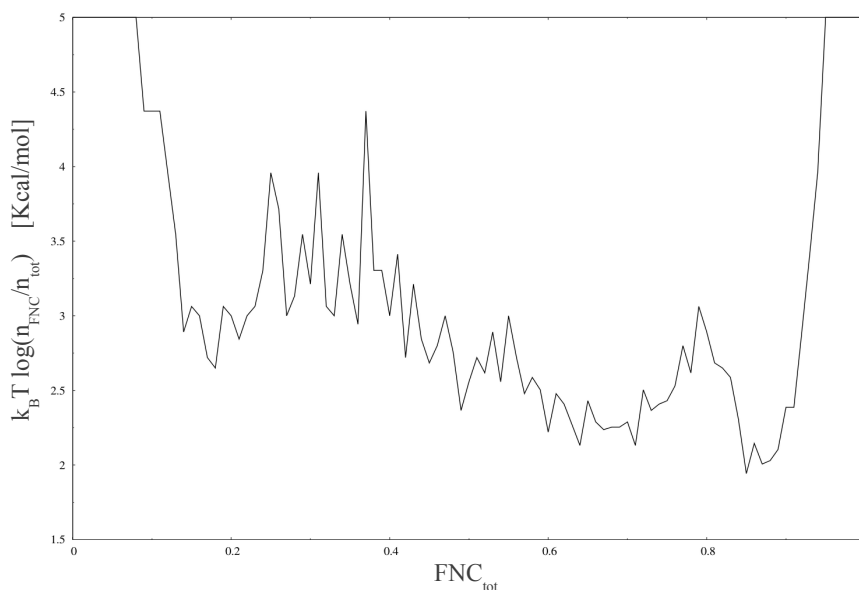


Figure 6.6: Kinetic free energy as a function of the total fraction of native contacts for the mutant protein of IM7.

As shown in Fig. 6.6, the three-state folding mechanism has been pre-

served as a proof of the explanation that Helix III does not play a role in the stabilization of the intermediate state, as expressed in Ref. [133], too. The intermediate state is characterized, also in this case, by the well structured Helices I and II, as shown in Fig. 6.8 a) and b), and a misalignment of Helix IV. Indeed from Fig. 6.8 d) it is clear that in the intermediate state, $FNC_{tot} \sim 0.7$, Helix IV is native-like but from Fig. 6.7 the interface between Helix I and IV is not well structured. It seems that Helix III is more formed than in the case of the wilde type. However, the results confirm that the state of the Helix III does not affect the stability of the intermediate state and, more in general, the characteristics that lead to the native state.

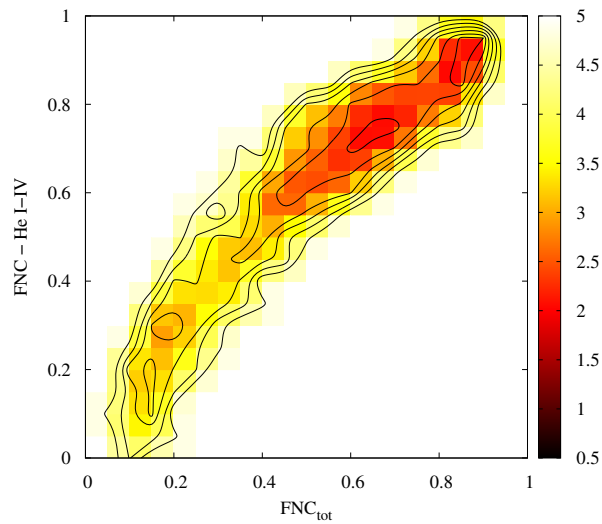


Figure 6.7: Kinetic free energy projected on the plane formed by total FNC and FNC limited to the interface between Helix I and Helix IV.

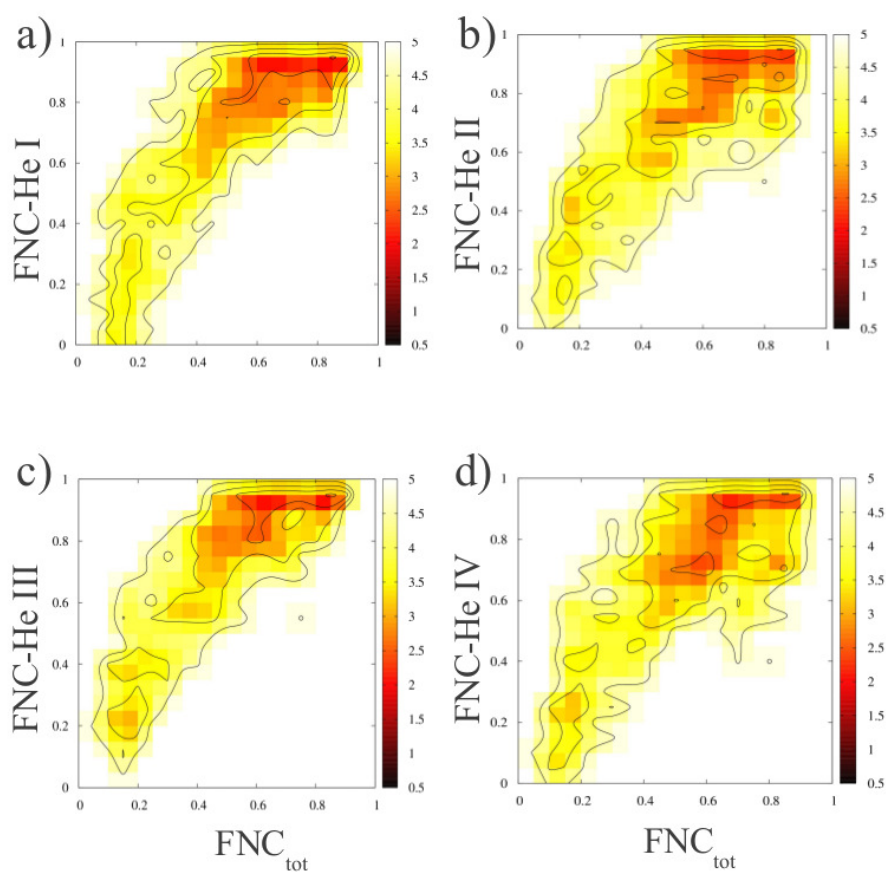


Figure 6.8: Kinetic free energy projected on the plane formed by the fraction of total native contacts and the fraction of native contacts limited to Helix I, a), Helix II, b), Helix III, c), and Helix IV, d), for the mutant IM7.

6.3.3 Characterization of the initial unfolded state

In order to completely understand the folding of the two immunity proteins it is important to study also the starting structures of the rMD simulations. In this way it is possible to note if a dependence from these configurations, as proposed by [129] exists.

Observing Fig. 6.3, obtained from thermal unfolded states, it is interesting to note that the minimum of IM9 is more pronounced than that of IM7. A possible explanation is offered by looking at the distribution of the initial configurations of the DRP folding simulations, red point in Fig. 6.3, namely the configurations from MD at high temperature and consequently relaxation. For IM9, these states are collected in the region with $FNC_{tot} < 0.2$, with a minimum at around 0.17, for IM7 are more dispersed with higher values for what concerns the fraction of total native contacts. Just after the thermal unfolding simulations and before relaxation the two proteins visit similar configurations, as reported in Fig. 6.9, for this reason we can say that the differences arise from the relaxation step.

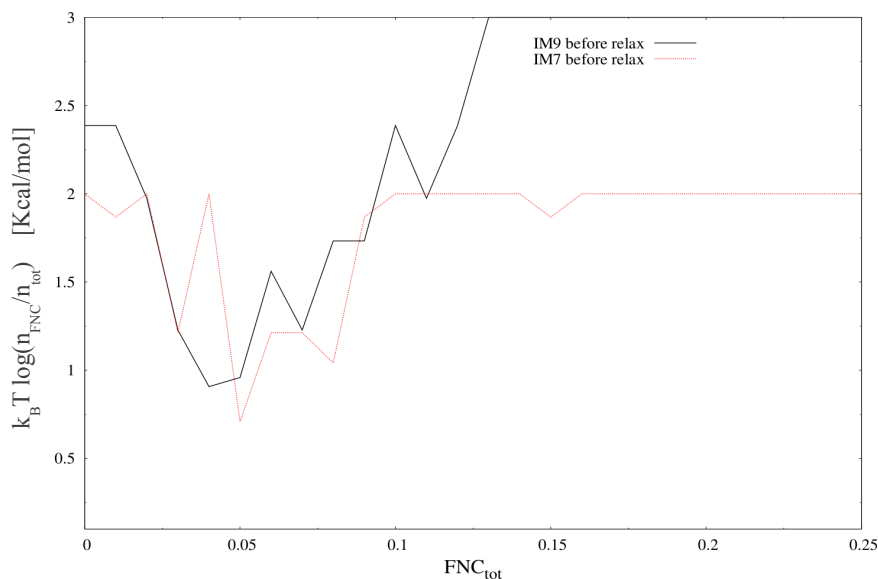


Figure 6.9: Kinetic free energy of the initial unfolded states before performing a relax simulation for IM9, solid black line, and IM7, dashed red line.

This observation suggests that protein IM7 has an higher inclination than IM9 to form native contacts. So, IM7 thermal unfolded state at the

beginning of the folding DRP simulation appears to have residual structures, in agreement with Pashley et al. [129].

But this propensity little says about possible effects on folding, that on the contrary we want to investigate. With this purpose in mind, folding DRP simulations from random coils, i. e. in total lack of residual structures, have been performed, see Fig. 6.10 and Fig. 6.11.

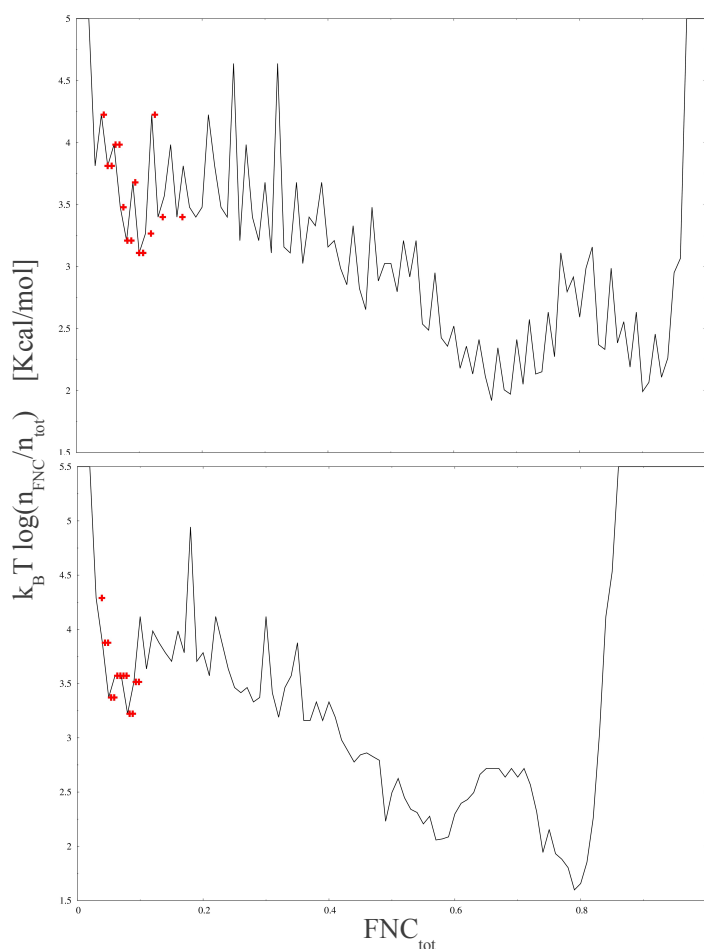


Figure 6.10: Kinetic free energy as a function of FNC_{tot} for IM7, top, and IM9, bottom. This results have been obtained from simulations starting from random coils. The configurations after the 1° rMD simulation are also reported as red symbols.

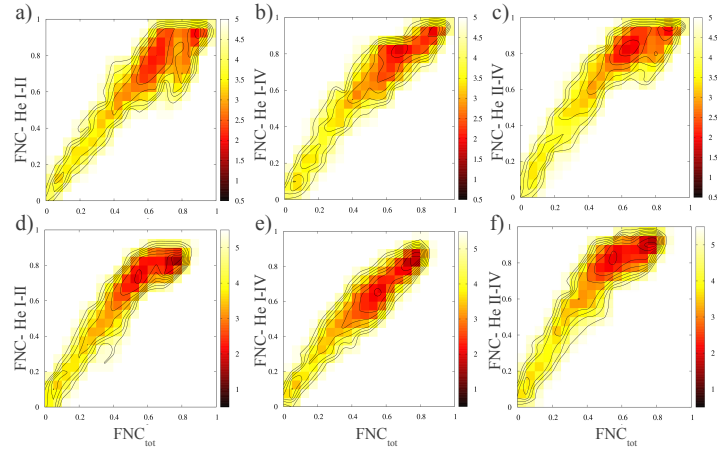
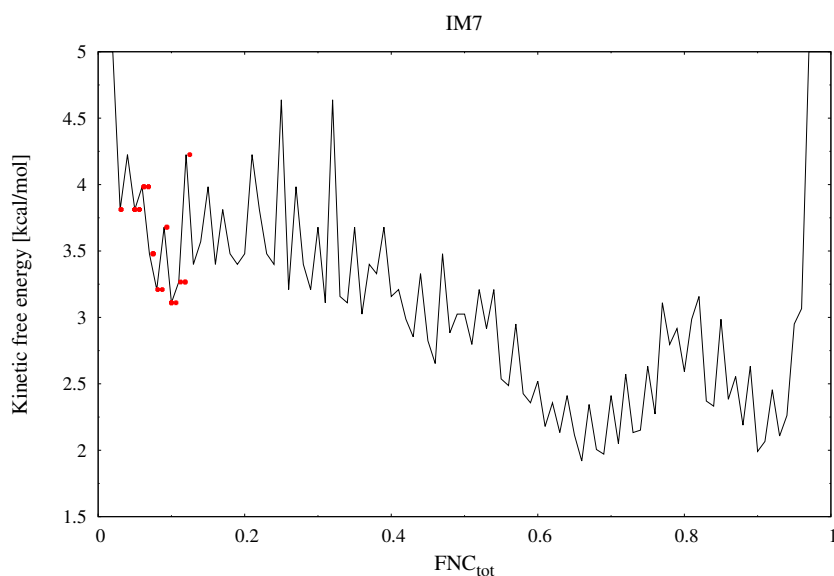
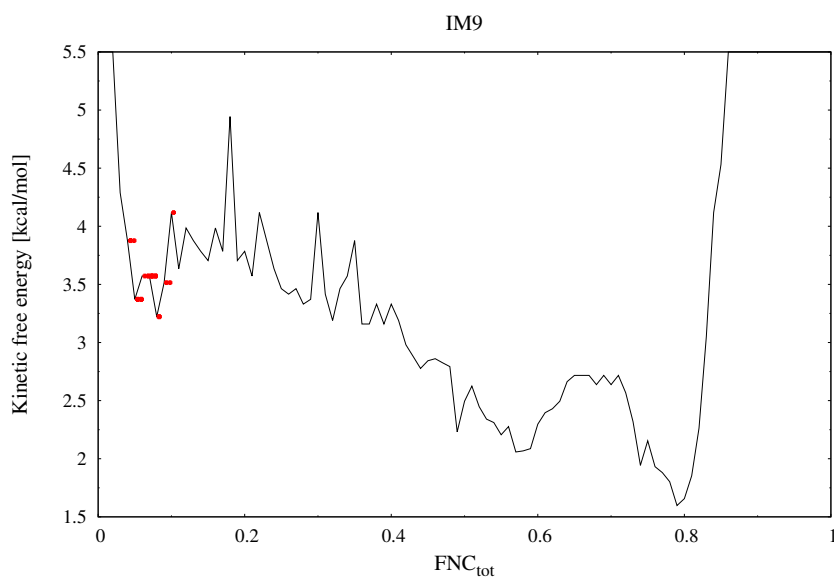


Figure 6.11: Kinetic free energy projected on the plane formed by FNC_{tot} and FNC restricted to pair of helices and their interfaces, for IM7, a) b) and c), and IM9, d) e) and f). This results has been obtained from simulations starting from random coils.

As expressed by Figs. 6.10 and 6.11, the results are well in agreement with those of Figs. 6.3 and 6.4, respectively, although at the beginning of the folding simulation the distribution of the total fraction of native contacts, around ~ 0.06 for the random coils for both IM7 and IM9, is lower than the values for the thermal unfolded states, see Fig. 6.12:



(a) Kinetic free energy for IM7, red points indicate configurations at the very beginning of the rMD-DRP simulation.



(b) Kinetic free energy for IM9, red points indicate configurations at the very beginning of the rMD-DRP simulation.

Figure 6.12: Kinetic free energy for IM7 and IM9 with starting configurations, red points, of the simulations.

It is also interesting to observe what happens after the first step of DRP simulations to the protein configurations, red symbols in Fig. 6.10: while

the FNC_{tot} of IM9 remains at this value the FNC_{tot} of IM7 increases and becomes more than 0.1, but can be also higher. This reinforces what suggested previously: IM7 has a higher propensity than IM9 to form native-like structures in the early steps of folding.

However, given the high similarity in DRP folding mechanism (intermediate state and its characterization) for different starting configurations, it is unlikely, in our opinion, that the unfolded state could contribute to determination of the folding, as proposed on the contrary by [129]. Instead, the high propensity in forming native contacts, an effect of sequence, could explain the higher stability of the intermediate in IM7 in respect of IM9. This feature can be connected with the observation reported in Ref. [22] that finds that the sum of native contact energies in Helix I-II interface is most favorable in IM7, followed by IM9.

6.4 Conclusions of this comparative study

We performed DRP simulations coupled with rMD algorithm in order to study the folding for the two proteins belonging to the Immunity family: IM7 and IM9.

The presence of the intermediate in IM9 as well as the nature of the interactions that stabilize the intermediate in both proteins and the structures are very debated, so, our work has the aim to try to give insight into these thematics, since, in addition, clarify these aspects could be helpful for understanding the limitations to the concepts of funneled energy landscape and minimal frustration, applied in the study of protein folding.

The performed simulations suggest that both IM7 and IM9 have an intermediate state during folding process and suggest, as a consequence of the high propensity to form native contacts on behalf of IM7, that the intermediate state of IM7 is more stable than that of IM9. Helices I and II are well formed and native-like while Helix IV can be formed but misaligned or partial formed. The presence of an intermediate in both proteins and the behaviour of the helices is in agreement with Karanicolas et al. [22]. However, in contrast with Karanicolas et al. we observe that Helix IV, despite not well formed or misaligned, forms native contacts in respect of Helix II, as reported in the work of [127]. Finally, Helix III, on the basis of our results, can be folded or unfolded during the intermediate stat. This observation supports the fact that it does not affect the intermediate state. An experimental validation of this statement are the low Φ values reported for this region and the investigations in manipulating the lenght of this helix [133].

Nevertheless, the differences with the work offered by Karanicolas et al. [22] do not affect the conclusions shared with our work, that the intermediate

state is observed in both proteins, that Helix I and II are formed and aligned in a native-like manner while Helix IV is only partial native-like. As a consequence, a purely $G\bar{o}$ -model can predict, for both IM7 and IM9, the existence of an intermediate state qualitatively similar to that predicted by our DRP simulation. So, it seems likely that the intermediate state is stabilized by native interactions.

Moreover, as a consequence of our characterization of the unfolded state we suggest that the initial state does not shape events that lead to the native state.

This work opens the possibility for other studies regarding the folding processes of homologous proteins in order to investigate the reliability of simulations based on pure $G\bar{o}$ model. Then, it could bring to a better understanding of on-pathway intermediate states, that can regulate and even make more efficient the folding mechanism. A fully explanation and characterization of these states could signify a breakthrough in the comprehension of possible misfolding causes in those proteins that show such a complex folding.

Chapter 7

Case of study 2: Thermal adaptation of marine ciliate pheromones

7.1 Introduction

In this Chapter the thermodynamics of two classes of pheromones is investigated.

7.1.1 Why thermal adaptation is important?

Most of the Earth's biosphere is cold, as it is possible to evince from Fig. 7.1, and exposed to temperature below 5°C throughout the year [134].

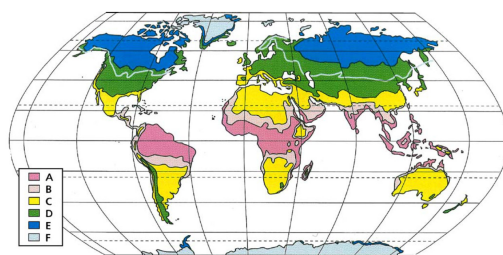


Figure 7.1: Most of the Earth is cold: A, annual minimum T above $+5^{\circ}\text{C}$, B annual minimum T above 0°C , C episodic frosts with T down -10°C , D regions with cold winters and mean annual minimum T between -10 and -40°C , E mean annual minimum temperatures below -40°C , F polar ice. Figure is reproduced with the permission from [134]

The activity of biomolecules is strongly affected by temperature vari-

ations, indeed the Kramer's formula 1.2 links the rate of occurring of a reaction with the temperature: at the decreasing of the temperature the rate becomes lower. For this reason, only cold-adapted organisms, called *psychrophilic organisms*, that develop modifications, which may be also deep and complex, result active in such environments [135]. These adaptive changes to cold environments are of keen interest, for example for the possible application in the industry of detergents, since cold-adapted enzymes could be used in washing at low temperatures with advantages such as energy-saving and applicability to synthetic fibers, or in the industry of food, for food preservation, but also in the pharmaceutical and biosensors fields, for monitoring low-temperature processes through enzymes [134].

Among the psychrophile organisms, a very interesting and useful system is offered by the marine protozoan ciliates *Euplotes nobilii* that inhabits the freezing Antarctic and Arctic waters. It has an evolutionarily closely related specie, as determined by rDNA sequence analysis [136], *Euplotes raikovi*, that, on the contrary, populate nonpolar temperature water [135]. Both families of *Euplotes* produce disulfide-rich water-borne protein pheromones for promoting their vegetative (mitotic) growth and sexual mating. The synthesized proteins from *Euplotes nobilii* have to be resistant to cold-induced denaturation and misfolding while those from *Euplotes raikovi* are mesophilic microorganisms [135].

In this way, thanks to the fact that these two pheromone families share a close evolutionary relation but a resistance to different temperatures as a consequence of the adaptation to their natural environments, these proteins become ideal organisms to investigate the fundamental physical principles which underlie thermal-adaptation in proteins [137].

This study is based on the article entitled “**Unfolding Thermodynamics of Cysteine-Rich Proteins and Molecular Thermal-Adaptation of Marine Ciliates**” published by T. Skrbic, G. Cazzolli, G. Guella and P. Faccioli on Biomolecules [137]. The paper has been written in the framework of a collaboration with professor Wüthrich and coworkers, The Scripps Research Institute, La Jolla, USA. In particular, Wüthrich's group performed the CD measurements reported in this chapter.

7.1.2 Characterization of the pheromones produced by *Euplotes nobilii* and *Euplotes raikovi*

In order to reach this aim, a set of representative members of the pheromones produced by *Euplotes nobilii*, *En*, and *Euplotes raikovi*, *Er*, has been taken into account. The pheromones and related PDB codes are collected in Table 7.1.

Er and *En* pheromones are single domain small proteins, composed by 30-40 residues for the *Er* family and by around 50 residues for the *En*

<i>Euplotes Raikovi</i>			
Name	<i>Er-1</i>	<i>Er-2</i>	<i>Er-10</i>
PDB code	1erc	1erd	1erp
<i>Euplotes Nobilii</i>			
Name	<i>En-1</i>	<i>En-2</i>	<i>En-6</i>
PDB code	2nsv	2nsw	2jms

Table 7.1: List of the *En* and *Er* pheromones investigated and corresponding PDB codes.

family. The proteins from both the families contain a compact core, with three helices bundled in up-down-up fashion and fastened together by disulfide bridges, three in *Er*'s pheromones and four in *En*'s pheromones [135]. Disulfide bridges, or CYS-CYS bond, are a type of side chain-side chain interactions between thiol groups, the functional group in Cysteine residue. The resulting bond is stronger than common side chain-side chain connections although weaker than the covalent bonds such as C-O or C-H. In a non-reducing environment at physiological temperature the disulfide bridges provide unbreakable topological constraints [138].

The high similarity in the topology of the pheromones from *En* and *Er* is visible in Fig: 7.2, where two representative pheromones, one per family, are compared. The picture underlines also the disulfide bonds, represented by colored spheres. Identical colors indicate residues that are paired together in a disulfide bond. A difference in structure between the two classes of proteins is that *En* pheromones show an N-terminal elongation of 10 up to 12 residues that does not appear in the *Er* pheromones.



(a) Ribbon representation of *Er-1*, left, and *En-1*, right (b) Cys-Cys bonds in *Er-1*, left, and *En-1*, right

Figure 7.2: Structural characterization of pheromones produced by *Euplotes nobilii* and *Euplotes raikovi*

Despite this structural similarity, *E. nobilii* and *E. raikovi* pheromones are characterized by very different unfolding/refolding thermodynamics. By comparing the temperature-dependent CD spectra of these pheromones it was found that, while the *E. nobilii* psychrophilic pheromones undergo an unfolding transition in temperature range from 55 °C to 70 °C, the mesophilic *E. raikovi* pheromones remain stable up to 100 °C [139]. In Fig. 7.3 the CD spectra of *n-1* and *Er-1*, chosen as representative for each family, are reported.

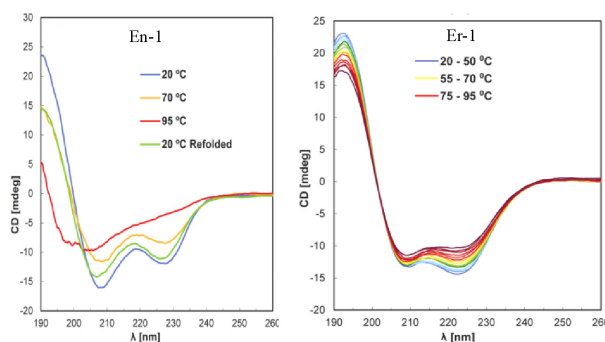


Figure 7.3: CD spectra at different temperatures for *En-1*, on the left, and *Er-1*, on the right. From this picture it is visible that *En-1* is unfolded at 95 °C while *Er-1* is stable even at this temperature. Figures are reproduced with the permission from [139].

The temperature of 95 °C has been found as the point in which the helices structures have been completely removed, while the temperature that identifies the unfolding transition is considered around 60 °C for *En-1*, as expressed in Fig. 7.4 where a sigmoidal curve for denaturation and refolding with cooling is presented. The midpoint of these sigmoidals is around 60 °C [139].

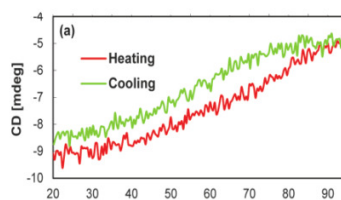


Figure 7.4: Temperature variation of the signal intensity at 220 nm during heating and cooling over the range from 20 °C to 95 °C. Figure is reproduced with the permission from [139].

These observations are valid also for the other pheromones belonging to

the two families.

7.1.3 Pheromone thermal adaptation: an open debate...

The observed difference in unfolding temperature for the two classes of pheromones is not surprising from an evolutionary point of view, because it could be expected that pheromones produced by organisms adapted to very different thermal environments may also show different thermodynamic behaviours, however, it is remarkable from a biophysical perspective. Indeed, on the basis of the energy landscape theory of protein folding the folding kinetics and thermodynamics are driven primarily by favourable interactions between residues in contact in the native state and, hence, by the native tertiary structure and topology, since the system is energetically biased towards the native state with minimal frustration. For what concerns the primary amino acid sequence and, in general, non- $G\bar{o}$ interactions, namely the interactions between residues not in contact in the native state and electrostatic interactions between all residues, are expected to perturb locally the structures but not to drive the folding process[22]. Consequently, a minimally-frustrated and native-centric picture of protein folding should predict that homologous proteins have similar unfolding temperatures.

In contrast with this observation, the thermodynamics of *E. nobilii* and *E. raikovi* pheromones seems to counteract the native-centric view leading to alternative hypotheses, where the protein primary structure and non- $G\bar{o}$ interactions are expected to have a dominant role. In line with this concept, it has been observed [135] that the physicochemical properties of the pheromones from cold-adapted *E. nobilii* are characterized by an improved backbone flexibility, due to the propensity to insert glycine residues, whose side chain is constituted only by a hydrogen allowing the residue to occupy positions not achievable for other more bulky residues, and by a reduced hydrophobicity, indeed, it has been found that in *En* pheromons there are regions with a high density of negatively charged side chains that enhance the interaction with the solvent. Similarly, the stability of mesophilic polypeptides has been suggested to be due to the presence of hydrophobic clusters [140].

7.1.4 ...and our hypothesis: the role of CYS-CYS bonds

In contrast with these positions we consider the role played by disulfide bonds in the stability of proteins by demonstrating that it is possible completely explain the thermodynamics of the two class of pheromones by taking into account only protein native structures and topology without invoking solvent-effects and non- $G\bar{o}$ interactions.

Our considerations are based on a vast research activity carried out in

order to investigate the ability of disulfide bridges in determining protein thermodynamics. An early work performed by Sheraga and coworkers [141] relies on the effect of stabilization of loops induced by disulfide bridges in case of helix-coil transition. From a comparison between the number of amide hydrogen bonds as a function of the temperature in a free chain structured in helices and in the same helix-chain but incorporated in a loop delimited by disulfide bridges it results that the helix-coil is stabilized to higher temperatures if inserted in a loop. A consideration about the role of entropy emerged: the cys-cys bonds are expected not to vanish at the increasing of temperature, at least in the range of temperatures until 100 °C since the strength of the bond is about 60kcal/mol, more than 10 fold higher than the value of $k_B T$ at 100°C. This constraint decreases the entropy gain of the unfolding transition because lowers the entropy of the unfolded form[141].

Not only the presence of disulfide bridges seems to modify the entropy of the unfolded state but also the location of these bonds is suggested to affect it: in particular Monte Carlo simulations have been showed that the longer the loop identified by cys-cys bonds the larger the variation in thermodynamics, since the entropy of the unfolded state is more reduced [142].

Moreover, the studies carried out by Camacho and Thirumalai have observed that disulfide bonds have a role during the folding mechanism in modifying also the time reaction: the bonds can stabilize intermediate states with the effect of slowing the dynamics. Camacho and Thirumalai collected the formation of cys-cys bonds following a schema, called the *proximity rule*, according to that at the early steps of folding non native cys-cys bonds form with a probability that depends on the length l of the loop, in particular, it vanishes for l small and increases for l large with a peak around 10 residues, after this value the probability decays exponentially. In a second step the non native cys-cys bonds break and native-like bonds are formed. In the third step the transition between native-like and native state is observed [143, 144].

For what concerns experiments for investigating the role played by disulfide bonds, evidences of the role of stabilizator carried out by these bonds are derived: by reducing the number of residues in a loop induced by cys-cys bonds the unfolding temperature of the protein decreases in an amount that depends on the number of residues deleted. The bigger the number of deleted residues, the higher the discrepancy in unfolding temperature [145]. Moreover, engineering experiments that add disulfide bridges show that mutant versions are more stable than the wilde type protein [146].

From this review it emerges that cys-cys bonds are expected to have a stabilization role for the native state of the protein against the denatured state. For this reason, the high amount of cystein bonds present in the treated pheromones is expected to play a role and we will try to extend

these concepts to the study of the adaptation of pheromones.

The work is structured as follows: first, we will check if we can reproduce through simulations the experimental evidences and verify whether a purely $G\bar{o}$ model can describe the different thermodynamics related to the two classes of proteins, in this way we should be able to evaluate the role played by native and non- $G\bar{o}$ interactions and possibly consider the nature of the origin of thermal adaptation. Then, once established that a $G\bar{o}$ model alone can explain the difference in unfolding temperature of the E_n and E_r pheromones, the role of disulfide bonds is investigated. In particular, we will go beyond the results to date, and we will try to understand why two classes of pheromones both cysteine-rich and with comparable average length of loops induced by disulfide bonds, around 17 residues, have so different behaviour at the increasing of temperature.

7.2 Methods

7.2.1 Monte Carlo simulations

Equilibrium Monte Carlo simulations based on the Metropolis algorithm have been performed in order to study the thermodynamics of the treated pheromones. Simulations at different temperatures in a range from 40°C until 100 °C have been carried out. The used potential follows the $G\bar{o}$ model as proposed by Karanicolas and Brooks [22]. As it is explained in the Chapter Theoretical Models, this is an improved version in which some effects of sequence in the dihedral motions and in differentiating the contribution of the type of interactions are taken into account. In order to reproduce the fact that the disulfide bridges are unbreakable in the considered range of temperatures we modeled the interaction between cysteine pairs that form disulfides bridges in the native state employing the value 15 times larger than the corresponding Miyazawa-Jernigan [88] contact potential between any pair of cysteine residues. This is justified by considering that, according to Miyazawa-Jernigan, the attractive potential between cysteine residues is $5.5RT$, a value from 3 to 4 kcal/mol depending on the temperature, multiply this values by a factor of 15 corresponds to estimate the real strength of the CYS-CYS bonds.

Since we are interested in studying the thermodynamics of the system at equilibrium and not the folding dynamics both local and global moves are employed.

For each pheromone of Table 7.1, at the particular temperature in range 40-100°C, we generated 12 independent MC trajectories, each consisting of 20,000 uncorrelated configurations. A standard autocorrelation analysis has been employed.

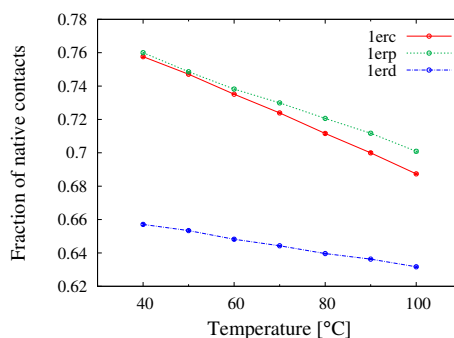
7.2.2 Input files describing starting structures

Regarding the input files, 6 starting configurations are chosen equal to the PDB native structures the other 6 are random coils.

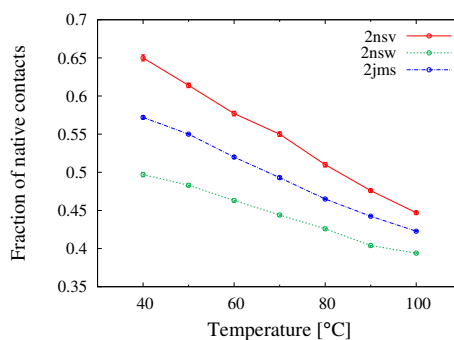
7.3 Results

7.3.1 Fraction of native contacts as a function of the temperature

The fractions of native contacts, fnc , for each temperature in the range 40-100 °C have been calculated. In Fig. 7.5 the temperature dependence of the average fraction of native contacts for *Er* and *En* pheromones is proposed.



(a) *Er* pheromones



(b) *En* pheromones

Figure 7.5: Average fraction of native contacts as a function of the temperature for the two classes of pheromones: *Er*, a), and *En*, b).

As it is possible to observe in Fig. 7.5, the results suggest that the pheromones of *Er* are more stable than those of *En* at the variation of the temperature, since the average fnc at the increasing of the temperature drops down up to 20% for the *nobilii* family while the decrease for *raikovi*

family is until 6–7%. This observation can be better evaluated by looking at the picture in Fig. 7.6, where the equilibrium distributions of the fraction of native contacts in two representative pheromones, namely $Er-1$ and $En-1$, at different temperatures are reported. In this way the different contributions at the average fraction of native contact of Fig. 7.5 can be investigated.

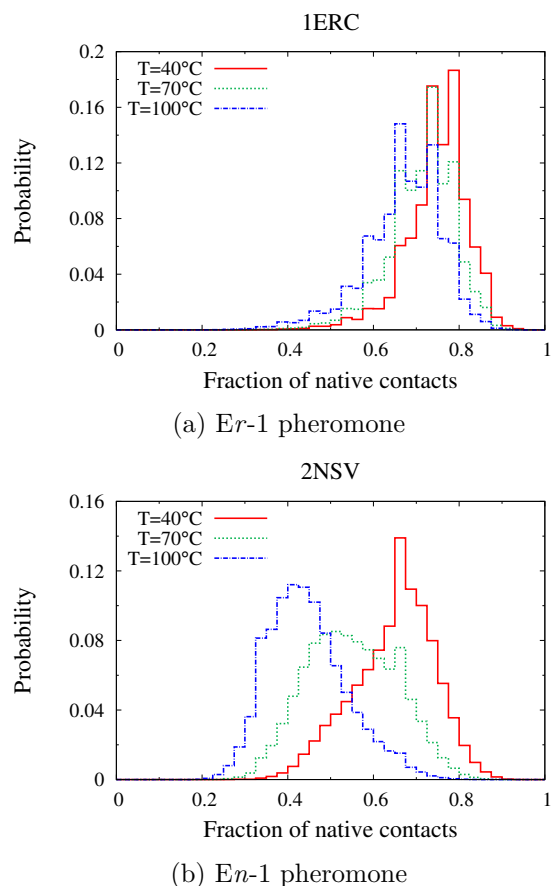


Figure 7.6: Equilibrium distributions of the fraction of native contacts for different temperatures for two representative pheromones: $Er-1$, a), and $En-1$, b).

$Er-1$ and $En-1$ are chosen as representative for the classes of *raikovi* and *nobilii* pheromones, respectively. We arbitrarily choose the $En-1$ and the $Er-1$ to represent their respective families because both their size and the number of disulfide-bridges well reproduce the average values in their families. The conclusion of the analysis would have remained unchanged if we had chosen different representatives with similar properties. From Fig. 7.6 it is possible to confirm that the distribution of $Er-1$ appears to be almost insensitive to temperature variations and the protein remains native even at

the highest temperature considered, 100°C. On the contrary, the fraction of native contacts in the *En-1* drops significantly at high temperatures.

Since the *Er* and *En* pheromones are mainly constituted by helices and the disulfide bonds do not vanish in the investigated temperature range, the values for the fraction of native contacts can be linked to the amount of helical structures and the decrease of these values can be considered as an indicator of thermal denaturation of those secondary structures. This observation allows a comparison between the simulated data and the CD experimental results. Indeed, we used the raw CD data of Geralt and al. [139] and reported in the CD spectra of Fig. 7.3 in order to quantify the fractional helicity f_H , namely the mole fraction of helical backbone within the protein, as a function of the temperature for two representative pheromones, *Er-1* and *En-1*. In Table 7.2 are reported the f_H obtained by deconvoluting CD spectra at 40 °C , 70 °C and 95 °C from the web interface DICHROWEB application [147].

Pheromone	$f_H - CD_{40^\circ C}$	$f_H - CD_{70^\circ C}$	$f_H - CD_{95^\circ C}$
<i>Er-1</i>	0.72 (0.07)	0.68 (0.08)	0.53 (0.03)
<i>En-1</i>	0.57 (0.08)	0.41 (0.14)	0.16 (0.07)

Table 7.2: Fractional helicity f_H at different temperatures for the *Er-1* and *En-1* pheromones, obtained by DICHROWEB/CONTINLL-4 analysis. The values in parenthesis are the corresponding normalized root mean square deviations.

Our results from MC simulations match even at a semi-quantitative level the distributions shown in Table 7.2. Indeed, for what concerns *Er-1*, it results from CD measurements a high and similar helix content $f_H = (65 \pm 8)\%$, in agreement with the results reported in Figs. 7.5 and 7.6 a). In case of *En-1* the CD data take into account of the decrease of the helical content at the increasing of the temperature as the simulated results in Figs. 7.5 and 7.6 b). From the comparison with CD experiments, it appears that, for both pheromones, our results underestimate the degree of unfolding at high temperature, 100 °C, while are comparable with those at low and intermediate temperature. However, although the model's parameters may be adjusted in order to reproduce quantitatively also the experimentally observed low helicity at a temperature close to 100 °C by slightly decreasing the strength of the native interactions involved in the secondary structures, this action is beyond the purpose of the work since we want to show only that the model is able to reproduce the difference between the two pheromones and not to fix this small discrepancy.

In conclusion, the above analysis suggests, in according with experimental results, that *Er* pheromones may not undergo an unfolding transition

in a range of temperature from 40 °C until 100 °C, as they remain mostly native-like in all the considered temperature range, while the fnc of En pheromones decreases as a function of the temperature allowing the hypothesis that a transition to the unfolded state can take place. In the next paragraph the unfolding temperature is searched.

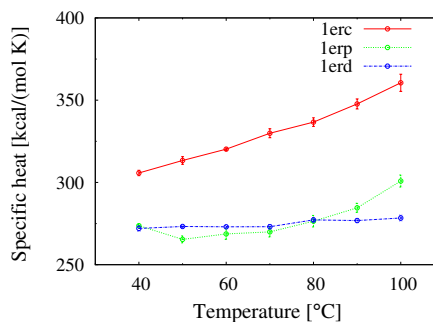
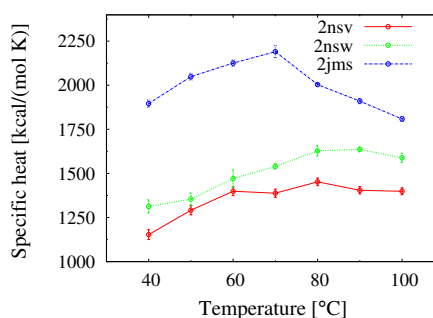
7.3.2 Specific heat and the fluctuation of the fnc

The folding-unfolding transition is similar to a first order phase transition, in the sense that, as a common point, at the transition two phases, the native and the denatured as well as, for example, the liquid and the solid of a component, coexist [148]. It is important to stress, however, that we deal with finite systems and in this sense is misleading to speak about phase transitions. From simulations we study a crossover between the native and denatured phases. In case of cystein-rich proteins, as CYS-CYS bonds are unbreakable in the considered temperature range, this crossover is much more flat than that in generic proteins without disulfide bridges. Only by taking into account these considerations it is possible to say that the goal in this section is to find the unfolding temperature, namely the temperature of the transition that should occur only in En pheromones. The calculation of the specific heat has been performed and reported in Fig. 7.7.

For Er pheromones the specific heat increases or remains constant at the increasing of temperature, but it is not possible observe clearly a peak in all the three En pheromones, as it would be hypothesized from the results presented in the previous paragraphs. However, one must observe that the specific heat is defined from energy fluctuations given, in the used model, by interactions at distance, namely hydrogen bonds and side chain-side chain interactions, that are affected by the disruption of the native contacts, and by the repulsive hardcore interactions, that, on the contrary, because the protein is continuously hold together by CYS-CYS bonds, are not suppressed if a transition takes place. In this way, while the fluctuations in energy related to interactions at distance are minimum for the denatured and native state and increase at the approaching of the transition point, the repulsive term is connected to potential energy fluctuations that are quite insensitive to the unfolding of secondary structures and generates a background contribution which makes the specific heat flat and the peak difficult to identify

On the contrary, for a globular protein free from CYS-CYS bonds, the swelling of the chain during unfolding affects the contributions from hardcore repulsive interactions and the peak in the specific heat should be sharp and visible, as in case of Fip35 reported in [101].

For this reason, in order to remove the contribution given by hardcore interactions and to be able to understand if a transition does occur, we chose a rescaled specific heat defined as follows:

(a) *Er* pheromones(b) *En* pheromonesFigure 7.7: Specific heat as a function of the temperature for *Er* pheromones, a), and *En* pheromones, b).

$$\chi(T) = \frac{\varepsilon_0^2 N^2}{k_B T^2} \overline{\Delta Q^2} = \frac{\varepsilon_0^2 N^2}{k_B T^2} \left(\langle Q^2 \rangle_T - \langle Q \rangle_T^2 \right), \quad (7.1)$$

where $\langle \rangle_T$ denotes the thermal average at the temperature T , N is the total number of native contacts, Q is the fraction of native contacts that are present in a given conformation, $\varepsilon_0 = 2$ kcal/mol is a typical contact energy and k_B is the Boltzmann constant. In this way only the contributions from the statistical fluctuations of the native interactions are taken into account.

In Fig. 7.8 the values of $\chi(T)$ as a function of the temperature for the two classes of pheromones are reported.

From the results of Fig. 7.8, *En* pheromones undergo an unfolding transition at around 60 °C, close to 80 °C for *En-2*. However, the temperatures of the program are nominal since the unfolding temperature is defined as the temperature in which the native state and the denatured state are equally populated, the other temperatures are rescaled from this, although the nominal temperatures are close to those real as the similar values suggest. But this analysis allows to understand that an unfolding transition takes place in the *nobilii* pheromone family, while the *raikovi* ones show a

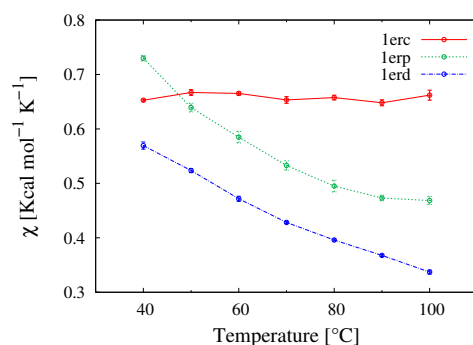
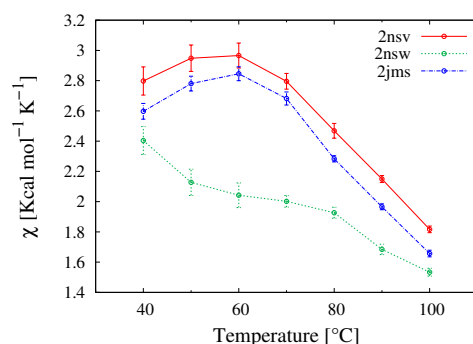
(a) *Er* pheromones(b) *En* pheromones

Figure 7.8: Fluctuations of the fraction of native contacts defined by Eq. 7.1 as a function of the temperature for *Er* pheromones, a), and *En* pheromones, b). Results are from MC simulations based on a coarse grained $G\bar{o}$ model.

different thermodynamic behaviour, in agreement with experimental results.

We are able to describe the different thermodynamics of the two types of pheromones with a coarse grained model that encodes only the information related to the protein tertiary structure and the location of the Cys-Cys bonds. In this way the dominant role played by native interactions is underlined. This finding contrasts with the explanation of the enhanced thermal stability of the *Er* pheromones based on the existence of a strongly hydrophobic cluster, as has been proposed, as well as does not support the hypothesis that *En* are less stable because of the insertion of charged residues along the chain.

Now, the problem relies on the features that really shape the different thermodynamic behaviour of the two classes of pheromones. We will focus our attention on the CYS-CYS bonds.

7.3.3 Localization of CYS-CYS bonds

As introduced, given the vast research activity on the thermodynamics of cystein rich-proteins, we focused our attention, for a possible explanation of the big differences in the thermodynamic behaviour, in the different topological constraints imposed by the specific location of the disulfide bonds.

First of all, we investigated the localization of generic tertiary contacts, that include not only disulfide bridges but all the side chain-side chain interactions and hydrogen-bonds, by calculating the Contact Order (CO) of the native structures of all the representative pheromones in accordance with [149]:

$$CO = \frac{1}{LN} \sum_{i < j} \Delta S_{ij}, \quad (7.2)$$

where N is the total number of contacts, L is the total number of amino acids in the protein, and ΔS_{ij} the number of residues in the chain sequence that separate residues i and j . In Table 7.3 the CO values of the *E. nobilii* and *E. raikovi* pheromones are reported.

<i>Euplotes Raikovi</i>			
Name	Er-1	Er-2	Er-10
CO_{total}	0.19	0.19	0.19
<i>Euplotes Nobilii</i>			
Name	En-1	En-2	En-6
CO_{total}	0.22	0.19	0.21

Table 7.3: The contact order calculated using all native contacts are reported for the investigated pheromones.

As it is possible to observe in Tab. 7.3, It appears that CO values are very similar in all the pehromones. As a consequence, the localization of generic tertiary contacts cannot be at the origin of the observed large difference in the unfolding temperature between the *En* and *Er* pheromones..

But, instead of taking into account all the contacts, we could restrict the calculation of the CO only to the disulfide bridges adjusting Eq. 7.2:

$$CO_{Cys-Cys} = \frac{1}{L \cdot N_{Cys}} \sum_{i=1, i < j}^{N_{Cys}} \Delta S_{ij}, \quad (7.3)$$

where now the indexes i and j run over Cys residues only and N_{Cys} indicates the total number of disulfide bridges. The results for $CO_{Cys-Cys}$

are reported in Table 7.4.

<i>Euplotes Raikovi</i>			
Name	<i>Er-1</i>	<i>Er-2</i>	<i>Er-10</i>
$CO_{Cys-Cys}$	0.46	0.43	0.48
<i>Euplotes Nobilii</i>			
Name	<i>En-1</i>	<i>En-2</i>	<i>En-6</i>
$CO_{Cys-Cys}$	0.34	0.32	0.32

Table 7.4: The contact order calculated only for Cys-Cys native contact are reported for the investigated pheromones.

What results from observation of Table 7.4 is that the values of $CO_{Cys-Cys}$ for *E. raikovi* pheromones are close to 0.45, while for *E. nobilii* pheromones around 0.33. This difference implies that the disulfide bonds in the mesophilic pheromones are systematically less local than in the psychrophilic pheromones. Thanks to this result a picture is delineated: non locality of the CYS-CYS bonds in *Er* pheromones implies a much broader action on the chain with a reducing of the conformational entropy gain in unfolding the secondary structures.

7.3.4 The mutant *En* pheromone

In order to verify the connection between CO and thermal stability in the *Euplotes nobilii* and *raikovi* pheromones, we designed a mutant of the *En-1* protein, in which the disulphide bonds are rearranged in order to reach values of the contact order of cysteine-bonds more typical of those of the family of *E. raikovi* pheromones.

In the mutant version of the *En-1* two of the Cys-Cys bonds have been translated from positions 11–37 to 15–37 and from 30–52 to 27–52, and Cys residues in the positions 23 and 33 have been replaced by Gly residues. The primary sequences of the wild-type and mutant *En-1* chains are the following:

En-1-wildtype:

NPEDWFTPDT **C**AYG **D**SNTAWTT **C**TTP **G**QT **C**YT **C**CSSCFDVV
GEQACQMSAQC

En-1-mutant:

NPEDWFTPDT **G**AYG **C**SNTAWTT **G**TTP **C**QT **G**YT **G**CSSCFDVV
GEQACQMSAQC

In this way, the disulfide bonds reproduce the same network of disulfide

bonds of the *Er-1* and the $CO_{Cys-cys}$ calculated for the new obtained protein has an increased value. All these conditions allow us to expect that this mutant should present a higher thermal stability, typical of *E. raikovi* pheromones. The results from the coarse grained MC based on the $G\bar{o}$ model are reported in Fig. 7.9, where the average fluctuation of the fraction of native contacts of the mutant protein is compared with that of *E. nobilii* pheromones:

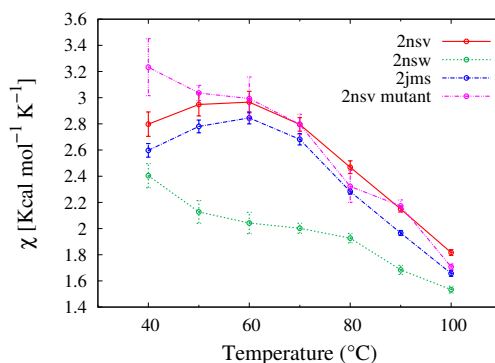


Figure 7.9: Fluctuations of the fraction of native contacts as a function of the temperature for *En* pheromones and the mutant version of *En-1*.

As shown in Fig. 7.9, the behaviour of $\chi(T)$ of the mutant protein loses the characteristics of the *En* pheromone family. Indeed, the evidence of a net peak vanishes and the thermodynamics becomes qualitatively similar to that of the *Eraikovi* pheromones. It is important to stress that in the mutant version the additional N-terminal part constituted by 13 residues is preserved, in this way it is unlikely that it has a role in determining *nobilii* instability, as proposed on the contrary by Geralt et al. [139].

7.4 Conclusions of the study

Monte Carlo coarse grained simulations based on the $G\bar{o}$ model developed by Karanicolas and Brooks [22] indicate that the different thermodynamics experimentally observed in *Euplotes nobilii* and *Euplotes raikovi* pheromones, proteins produced by organisms adapted to cold and temperate water, respectively, are due to the location of constraints determined by the disulfide bonds. This conclusion challenges the original explanation, according to which charged residues and more flexibility induced by the propensity to have included in the chain glycines are the cause of the lower stability of *E. nobilii* pheromones while the presence of a strong hydrophobic core produces the enhanced thermal resistance of *E. raikovi* pheromones. Indeed, in the performed simulations the charge of particles is not taken into account and

the role in flexibility determined by the side chains is underestimated, since it is considered only in the torsional angles but does not take large part in repulsion between adjacent regions of the chains.

This work introduces a new point of view for explaining cold-adaptation and new simulations and experiments have to be done in order to completely validated the assumptions about the connection between CO related to CYS-CYS bonds proposed in this chapter. Very interesting could be carry on CD measurements on the proposed mutant of *En-1*, in order to check whether the thermodynamics is in agreement with that predicted by MC simulations. Moreover, coarse grained MC simulations based on the $G\bar{o}$ model and CD measurements on a mutant version of the *Er-1* arranged in order to show a $CO_{Cys-Cys}$ similar to that of *E. nobilii* pheromone family could be also useful for the same reason. Simulations with a more realistic potential are also to be performed.

For what concerns the particular study of *Euplotes* pheromones another possibility for future work could be investigate also their folding, that has been suggested to occur involving more than two states [139]. Analyzing this complexity could be interesting in order to characterize the influence of cystein bonds on the events and time-scale of protein folding.

Chapter 8

Case of study 3: Atomic-level characterization of conformational changes in Serpins family

8.1 Serpins: a general introduction

The results provided by simulations about serpin proteins are explained.

8.1.1 Serine protease inhibitor

SERPIN is an acronym, coined in 1985 by Carrell [150], that means **SER**ine **Pro**tease **INH**ibitor. It describes a family of proteins involved in the regulation of numerous physiological processes such as blood coagulation, fibrinolysis, inflammation, angiogenesis and even the development of synaptic plasticity [151, 10], through the inhibition of enzymes whose activity is based on the presence, at the active site, of the residues serine and cysteine [152].

However, the importance of serpins arises also from activities that extend beyond their function to inhibit proteases. For instance, they may also regulate blood pressure and hormone transport [151].

Serpins can be found in all superkingdoms (Eukarya, Bacteria and Archaea) as well as certain viruses [153].

8.1.2 Serpins: Structure and plasticity

The proteins belonging to the serpins' superfamily share a precise structural architecture but only a modest sequence identity [10]. As Huntington reports

[10], during a Serpin Symposium [153], it was said that “if you’ve seen one serpin structure, you’ve seen them all”. This sentence highlights the similarity of serpins for what concerns architecture, that comprises nine α helices (usually labeled by letters from A until I) and three β sheets (A to C). The average size of these proteins is about 350-400 amino acids and the molecular weight is 40-50 kDa [154].

A picture of PAI-1, a serpin treated in this work, is reported in Fig 8.1. One hallmark of serpins is represented by the so called reactive center loop (**RCL**), in red in the figure 8.1, a motif constituted by about 20 residues, very flexible and containing a sequence complementary to the active site of its target protease [10].

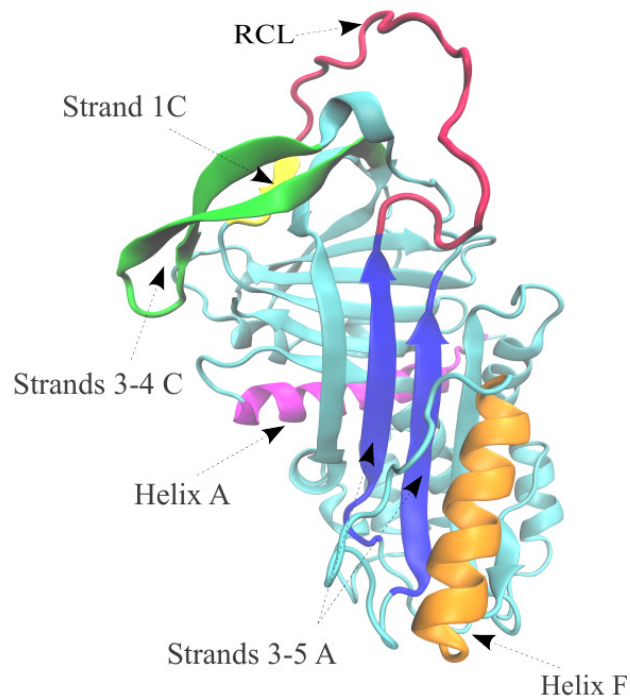


Figure 8.1: The native structure of Serpin Plasminogen activator inhibitor-1. Important secondary structures are reported.

Whereas on one hand the overall structural organization is highly conserved in all the proteins belonging to serpin family, x-ray crystallographic measurements revealed structures that show significant conformational plasticity with respect to the RCL and β -sheet A (two of the five strands of β -sheet A are reported in blue in Fig 8.1). In particular, as expressed by Fig 8.2, an increasing incorporation of the RCL in sheet A is observed [10].

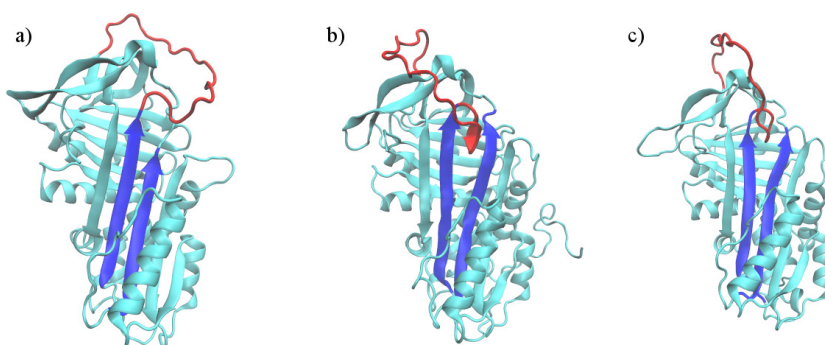


Figure 8.2: The crystallographically defined structures of serpins Plasminogen activator inhibitor-1 (PDB code: 1QLP), a), antithrombin (1T1F), b), and murine α_1 -antichymotrypsin (1YXA), c), in their native state. The RCL is colored in red, the strands 3 and 5 of β -sheet A in blue.

The plasticity shown in Fig. 8.2 is driven by the metastability of serpins in their native state: in fact, whereas proteins usually fold to their thermodynamically stable state, native serpins are kinetically trapped in a high-energy state [10], that is essential for their inhibition activity, in fact, only after the total insertion of the RCL in sheet A the serpin is in its more stable state, although not native.

8.1.3 The serpin mechanism

Serpin are not the only inhibitor of proteases, but there exist other non-serpin families, for example the Kunitz family, a representative is the bovine pancreatic trypsin inhibitor, or BPTI, that are able to perform this function. All these non serpin inhibitors use the “standard mechanism” for inhibiting the enzymes, namely a completely reversible process in which the RCL is rigid since it is held in place by disulfide bonds and fits as a key in the lock of the active site of the protease [10]. The standard mechanism is schematized in Fig. 8.3, where also the comparison with the completely different inhibitory process carried out by serpin is represented.

On the contrary indeed, the mechanism of inhibition of serpins has been

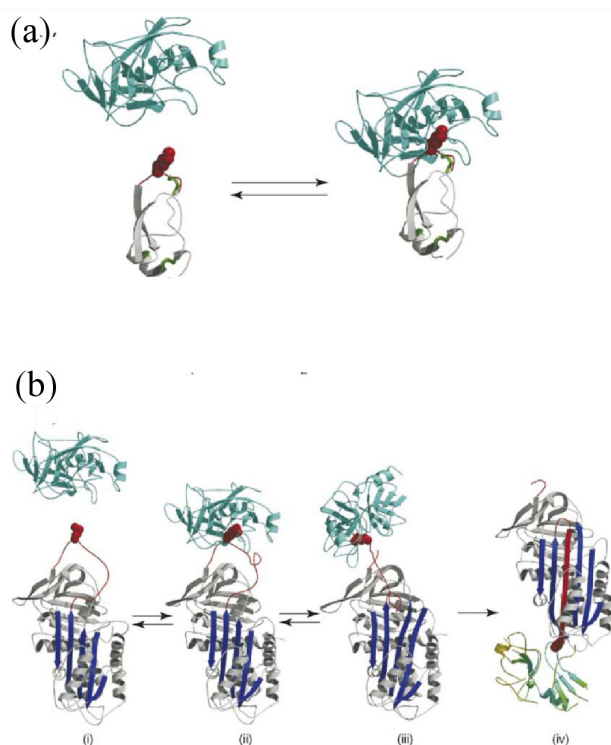


Figure 8.3: a) Standard inhibitory mechanism for a representative member of non serpin protease inhibitor, the bovine pancreatic trypsin inhibitor (colored in grey). Trypsin is colored in cyan and the RCL is in red. b) Serpin's inhibitory process. The representative serpin chosen in this raffigation is α_1 -antitrypsin. Figure is reproduced with the permission from [10].

described to be a suicide-like mechanism: the protease recognizes the RCL as a substrate loop and the Michaelis complex forms reversibly, panels I and II in Fig. 8.3 The following step is the formation of an ester bond between the active site of the enzyme, a serine or a cystein residue, and the P1 residue on the RCL (with non-primed numbers towards the N terminus, and primus numbers towards the C terminus of the serpin). As a consequence, the peptide bond that links residues P1 and P1' on the RCL is cleaved [10], panel III in Fig. 8.3. This mechanism lows the energy barrier that traps the serpin in its metastable, native state and the RCL begins to insert into the β -sheet A, which opens in oder to allow the insertion. The RCL transports the covalently bound protease with it [154]. The conformational change that undergoes the serpin is irreversible and at the end the RCL is completely incorporated in the β -sheet A, reaching the final, stable state, panel IV in Fig. 8.3. The protease is displaced from the top to the bottom of the serpin of $\sim 70\text{\AA}$ and experiences two opposite forces acting on it: the viscous drag

over the whole enzyme and the force of serpin by pulling on the protease localized in the active site [10, 154, 155, 156]. Because of these forces the active-site loop of protease is plucked out [10]: the protease is now distorted and inactivated.

The energy required for carrying out this inhibition is thought to come from the increase insertion of the RCL that leads to a more stable state [154, 157].

The inhibitory function carried out by serpins has, in respect with the other non serpin inhibitors of protease, the important advantage of irreversibility, that is crucial especially in process that are based on proteolytic cascades that lead to large amount of terminal proteases from small quantities of preceding proteases [10]. An example is provided by haemostasis, namely blood coagulation, in which a small quantity of factor IXa amplifies the formation of factor Xa that on its side amplifies the formation of terminal thrombin, the protease adhibited to coagulation. The correct amount of thrombin enables appropriate clot formation while a higher quantity circulating in blood causes thrombosis, for this reason the inhibition of factor IXa is important in order to prevent clot dissemination and thrombosis [10].

8.1.4 Serpins' diseases and misfolding

The price to be paid for ability and efficiency of the unique inhibitory mechanism carried out by serpins is the criticality due to complexity and mobility connected to this process [10]. Indeed, mutations can lead to misfolded configurations that are inactive since the metastable state is bypassed [158] or in which the energy barrier related to loop insertion is altered by speeding up or slowing down the rate at which it occurs, dramatically altering the functionality of the protease inhibitor. Any modification in the mechanism that involves the translocation of the RCL causes a failure in correct functionality by altering the equilibrium in which protease and serpin act. For example, in Cambridge II, a variant of antithrombin, a mutation renders the insertion of the RCL slower than deacylation and increases the risk of venous thrombosis, a similar loss of function could occur in another variant of antithrombin that, on the contrary, speeds up the rate at which the RCL inserts without protease attack and leads to venous thrombosis in periods of increased temperature, as happens during infection [10]. But a variation in RCL insertion rate is observed also, in absence of mutations in residue sequence, in the serpin Plasminogen activator inhibitor-1 (PAI-1), by the action of the binding vitronectin: vitronectin is essential for stabilizing the metastable state of PAI-1, a particular labile serpin that, in contrast with other representatives belonging to the family, spontaneously tends to a final stable state characterized by total insertion of uncleaved RCL, this state is also called *latent state*. However, the action of vitronectin by increasing

half-life of PAI-1 is not without effects since it results that high level of active PAI-1 are associated with cardiovascular diseases [159].

These observations highlight the fact that the mechanism is delicate and sensible to every mutation in the type of amino acids and that every change in the events that lead to the inactive state could have big effects.

Mutations do not only causes diseases directly correlated with the involved serpin, but renders serpins prone to form toxic intracellular aggregates that can be the cause of cell death and tissue destruction sharing similarities with prion, Huntigton's and Alzheimer's diseases. These aggregates of serpins are the cause of the so-called serpinopathies and include cirrhosis and emphisema, from antitrypsin aggregates, dementia, from neuroserpin polymerization, and thromboembolic disease associated with antithrombin polymerization [153].

8.1.5 Characterization of the serpin conformational change from experiments and simulations

In the last section we have discussed how the characterization of the role of mutations on the rate of insertion of RCL could have a large impact in medicine research field , or in which measure environment in the sense of ligands can affect the life time of serpins. Going more in detail the understanding of the set of movements and adjustments inside the serpins that lead to the complex motion of the RCL could be also helpful in determininig the more sensitive regions in influencing the dynamics.

However, this fully comprehension is far from being reached. The first solved crystal structure of the complex between α_1 -antitrypsin (serpin) and trypsin (protease) have been published only in 2000, given the high degree of flexibility and disorder in the configuration [10]. Other crystal structures followed this important result and gave insight into the first phases of the reaction concerning the Michaelis complex, allowing to hypothesize that dramatic conformational changes occur in serpin protein [10].

Experiments, such as mutations in residue chain or hydrogen deuterium exchange coupled with mass spectroscopy, have been performed in order to clarify all the aspects related to the insertion of the RCL and the consequent movements in secondary structures. What emerges is that particular secondary structures are expected to play a precise role during the insertion: in particular the attention has been focused on Helix F, colored in orange in Fig. 8.1. It has been shown that by increasing the interactions between this helix and β sheet A the inhibitory activity decreases, although the serpin-protease association rate was immutated [160]. These results coupled with the demonstration that a rapid deuterium uptake, expecially in the top of helix F, occurs suggest a movement or unfolding of Helix F should take place in order to allow the passage of the RCL [161, 34]. Moreover, helix

A, magenta in Fig. 8.1, with an unfolding is hypothesized to perturb the interactions in the surrounding regions leading to opening the gap between sheet A [161, 34]. In particular, this helix is indicated as the possible regulator in serpins like plasminogen activator inhibitor-1 [34]. As Helix A, also Helix D, in violet in Figure 8.4, is indicated by some studies as flexible [34]. However, for this helix not agreement is reached since it has been observed on the contrary by other studies that a low amount of H-D exchange occurs, although conformational changes without solvent exposure have been proposed [161].

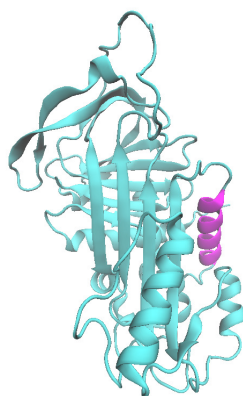


Figure 8.4: Particular of Helix D, colored in violet, in serpin, cyan.

Experiments can provide a picture about conformational changes in serpins that is, however, only hypothesized. The network of interactions is complex and it is difficult to derive the effect of particular mutations, moreover, the hydrogen-deuterium exchange technique is affected by the presence of solvent, and it could happen that flexible parts that are not in contact with the solvent do not show a high level of deuterium [161]. A help could in principle come from simulations. Unfortunately, difficulties and constraints in studying serpins are not only in experiments but can be found also in simulations, where, due to the size, around 370 residues, and the involved time scales, of the order of seconds but in some cases even more, hours or weeks, investigate serpins reaction is a very challenging task. To date, a simulation is not available about conformational changes occurring in serpins.

8.2 Our work overcomes the limits in serpin reaction investigation.

8.2.1 The challenge: is it possible simulate conformational changes in serpins?

Our work have had the aim of performing detailed simulations of what happens inside serpins during the translocation of the RCL, although the size of involved proteins and time scale of the reactions are beyond the limitations of recent simulation techniques, see Chapter Experiments and simulations in protein folding for a detailed review.

We chose to investigate the dynamics of these serpins: α_1 antitrypsin, that in the following will be called α_1 -AT, and is a prototypical serpin that inhibits a lot of proteases and can undergoes the translocation of RCL only after the cleavage by the protease, and plasminogen activator inhibitor-1, PAI-1, in the wilde type form and in two mutant versions. PAI-1 negatively regulates blood clot clearance (fibrinolysis) by mechanically inhibiting important serine proteases, including tissue type plasminogen activator and urokinase type plasminogen activator. This serpin is unique among the other representatives of the family since, as said before, can spontaneously deactivate by inserting its intact, uncleaved RCL into sheet A, resulting in the thermodynamically stable, but inactive, latent conformation [162, 163], represented in Fig. 8.5. The two mutants show only a two- and four-site mutations in respect of the wilde type but act by altering the half time of the serpin that becomes at 37°C 9-fold [164] shorter and 72-fold greater [165], respectively.

Therefore, the diversity of these serpins renders them ideal candidates for performing a study about this family since from comparison between them it should be possible give insight into the network of interactions and the sequence of events that lead to the final state. It could be possible give answer to questions such as: why, although both α_1 -AT and PAI-1 share a similar three dimensional structures and equal lenght of the RCL, the insertion of this has different degree of spontaneity? In which manner point mutations can alter the time scale of reaction in PAI-1? Finding explanation for such issues would clearly have repercussions for the whole research relative to serpins but also a breakthrough for what concerns diseases related to these serpins, in particular PAI-1, whose activity, as said, can be regulated by directly regulating the transition to the latent state with binding to the cell adhesion factor vitronectin. However, there is necessity for looking for other ways for regulating PAI-1, since high levels of active PAI-1 are associated both with poor prognoses for some cancers, presumably due to interactions with vitronectin, and with cardiovascular disease. Our study has also the more practical purpose to lead to a scheme in order

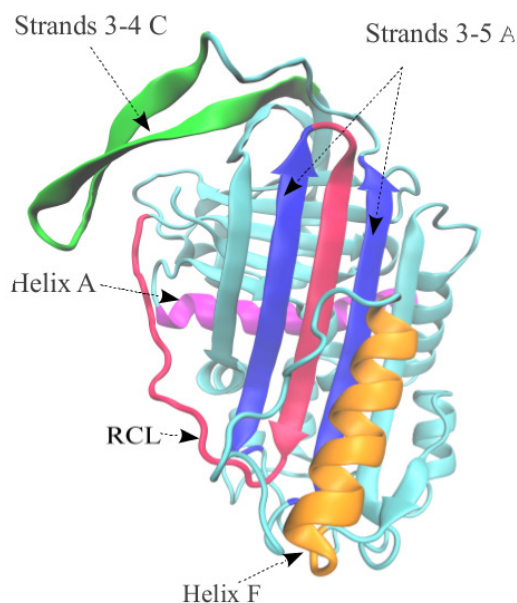


Figure 8.5: Structure of the latent state of serpin Plasminogen activator inhibitor-1.

to identify ligands whose binding can properly regulate the transition to latency.

The work that will be presented is structured as follows. The pdb file of α_1 -AT has been manipulated in order to mimic the action of the protease by cleaving the RCL. Then, preliminary coarse grained MC simulations of the reaction of the obtained system have been performed based on a $G\bar{o}$ model. Since the high amount of experimental data achievable in literature about this serpin [34, 161, 166, 167], it is expected to be a proper reference for testing MC simulations compared with experimental results. After verifying that the method is able to reproduce the dynamics of α_1 -AT and analysing the reaction in detail in order to obtain a general picture and an indication about the structures, whose behaviour has to be focused, we will proceed beyond and try to apply the technique to uncleaved α_1 -AT, that with intact RCL should remain in the metastable state stable for weeks, and PAI-1, that, on the contrary and despite the intact RCL, should undergoes spontaneously the transition with a half time of 1 hour at 37 °C [164]. As it will be explained, MC simulations do not demonstrate to properly describe this different degree of spontaneity. Then, the investigation proceeded by

carrying out rMD-DRP simulations. For a constraint in the program cleaved α_1 -AT has not been studied but the work focused on intact α_1 -AT, PAI-1 wilde type and the two mutants of PAI-1 that show different time scales. All DRP performed simulations and related analyses are reported on the basis of the work “**Atomic-level characterization of the serpin latency transition**”, written by G. Cazzoli, F. Wang, S. a Beccara, A. Gershenson, P. Faccioli and P. L. Wintrode, submitted.

8.3 Methods

8.3.1 MC simulations

The first step of our investigation about the dynamics of serpins has been to perform one-bead coarse grained Monte Carlo simulations based on the Metropolis algorithm. The applied potentials have been modeled by using $G\bar{o}$ interactions, as proposed by Karanicolas and Brooks [22], and $G\bar{o}$ with the addition of non- $G\bar{o}$ interactions, as developed by Kim and Hummer [90]. In this way the role palyed by sequence effect and non-native interactions is investigated. Moreover, in carrying on these simulations we want to verify if in this case a coarse grained model is enough for describing the system.

We are not interested in studying the properties of the system at equilibrium, that is when it has reached the final stable state, but we are interested in investigating the dynamics that lead the serpin from the metastable state to the end of the process. For this reason only moves with a local character are employed, that means that global pivot has been neglected.

The MC simulations have been performed at the nominal temperature $T=300K$. Since, at the beginning, we could not achieve the final state and, on the contrary, the serpin remained trapped in an intermediate state, we had to rescale the contribution given by the native interactions by multiplying this value by a factor of 2.5. This factor normally is setted to 1.0 but can be changed in principle for matching the unfolding temperature provided by the simulation with the real temperature of the transition. By increasing this factor the result is that we lower the temperature and decrease thermal fluctuations: with this choice the protein reach the stable final state.

MC simulations of the reaction in serpins require as input two pdb files: the metastable starting configuration and the final state.

The input configurations are:

- cleaved α_1 -AT: 372 residues, the metastable and ending state are obtained by manipulating 1qlp.pdb and 1ezx.pdb, respetively, for obtaining the cleavage of bond between reisdues P1-P1' without removing any residue.

- unclaved α_1 -AT: 372 residues, the metastable state is given by 1qlp.pdb while the ending state is described by 1iz2.pdb. The sequence of residues in 1iz2.pdb has been modified in order to match that of 1qlp.pdb.
- intact PAI-1: 379 residues, the metastable state is from 1oc0.pdb and the latent state from 1dvn.pdb. The missing residues in PDB 1oc0.pdb were modeled by using Mod-Loop package [168]. Moreover, some mutations in the metastable state's PDB have been introduced in order to match the sequence of residues in the latent form's PDB.

8.3.2 rMD-DRP simulations

We perform all-atom rMD simulations based on the DRP approach with the implementation of the force field AMBER ff99SB-ILDN in implicit solvent. For a limitation of the program we cannot simulate the cleaved α_1 -AT but only the uncleaved α_1 -AT and PAI-1 in the wilde type form and in two mutant versions.

The program that implements the method requires two input files, namely the starting and the ending configurations:

- PAI-1 wilde type (WT): The metastable and latent structures were provided by PDB data bank, codes 1oc0.pdb and 1dvn.pdb. The pdb's have been manipulated in the same way that permits to obtain the input file for the MC simulation, but in this case a refinement with Rosetta [169] has been added in order to remove all the steric clashes between side chains that a not well optimized pdb can show. Then, an energy minimization in implicit solvent by using Gromacs molecular dynamics package 4.5.5 [130, 131] was performed in both PDBs. The number of minimization steps was set to 50,000 steps, the force field AMBER ff99SB-ILDN was used.
- PAI-1 mutants: The mutations for the two PAI-1 mutants were made directly on the wilde type PDB, by changing the type of the residues. Then, the preparation was the same than the input files of the wilde type form: the coordinates of the just introduced residues were provided by Mod-Loop, refined by Rosetta and the new PDBs were energy minimized by Gromacs. The mutants are: PAI-1 quadruple mutant N150H/K154T/Q319L/M354I, according to [165] and called in the following PAI-1 stab since as effect of these mutations half life is 72-fold greater than wilde type [165], and PAI-1 double mutant G38S/Q322H, as proposed in [164], in which a different residue numeration according to α_1 -antitrypsin template numbering scheme has been used. This last mutant will be called PAI-1 destab since has a half-time 9-fold shorter than wilde type [164].

- uncleaved α_1 -AT: as in the previous simulations, the metastable state is described by 1qlp.pdb while the final state by 1iz2.pdb. The sequence of 1iz2.pdb has been modified in order to match 1qlp.pdb. Then follow the refinement with Rosetta and the energy minimization with Gromacs.

After some tests, the parameters that regulate the strength of the ratchet potential, k , and the rate at which configurations represented by a variable z higher than the reference z_m are accepted, k_a , have been set at $0.01eV$ and 0.01 , respectively. For each studied protein 12 trail trajectories have been generated, running for 5×10^4 steps of MD with a nominal integration time step of $\Delta t = 1$ fs, performed using a Velocity-Verlet algorithm, coupled to a Nose-Hoover thermostat. Other sets of trail trajectories have been simulated for each systems and the results have been compared between each others. They always result in perfect agreement.

8.3.3 RMSD

In the following analysis the final stable state of serpin has been chosen as reference for calculating RMSD values.

8.3.4 Euclidean distance

The Euclidean distance is the distance in the space of the coordinates between two configurations, namely, given two configurations X_i and X_{i+1} the Euclidean distance between them is calculated in the following way: $\sqrt{(X_{i+1} - X_i)^2}$ [108]. In the following results it will be calculated in Å.

8.4 Results

8.4.1 MC results and analysis

For cleaved α_1 -AT we obtained 6 coarse grained successful trajectories among 12 simulated trajectories with the use of the pure $G\bar{o}$ model. The result is schematized in Fig. 8.6:

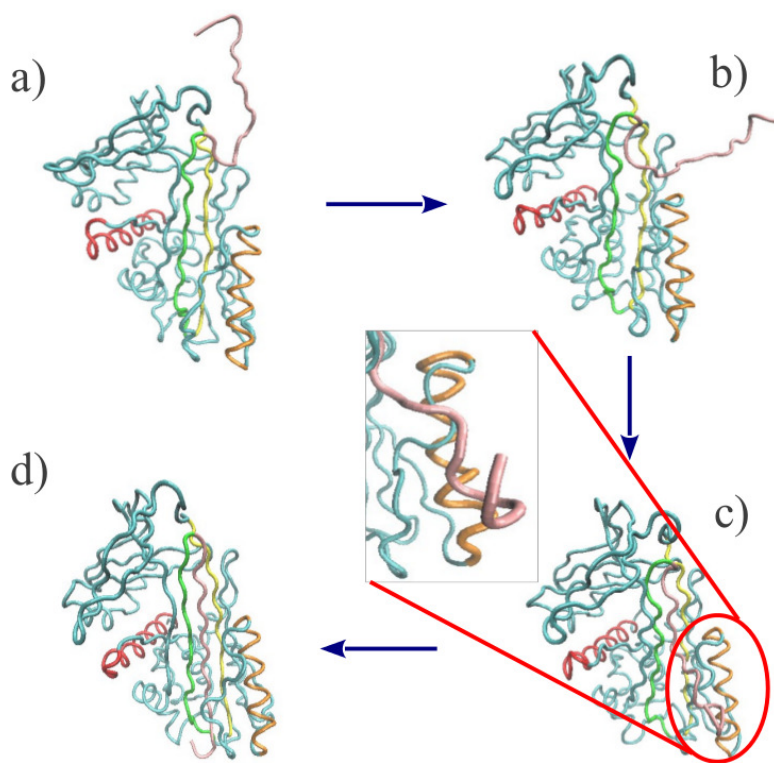
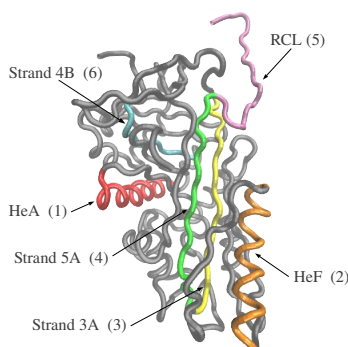
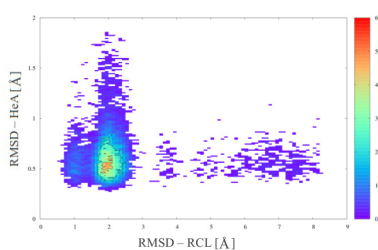


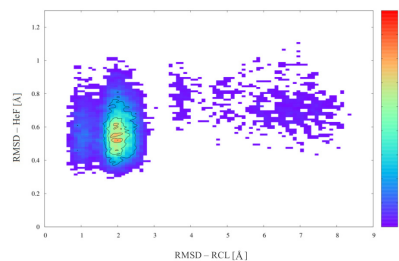
Figure 8.6: Snapshots from MC simulations of α_1 -AT. a) starting configuration, b) RCL begins to insert, c) Insertion of RCL goes further until the barrier represented by Helix F, colored in orange. The particular of the obstacles caused by Helix F is enlarged on the left. d) Complete insertion of RCL.

RMSD values for Helix A and Helix F have been calculated in respect of the ending structure. In Fig. 8.7 the number of times a state, defined by the values assumed by the RMSD of the helix and the RCL, is visited during the simulation is reported.

From the observation of Fig. 8.7 it is possible to note that the system reaches the final and thermodynamically stable state, indicated by a higher density of points in the range $3-1\text{\AA}$ defined by the values of RMSD-RCL. For what concerns Helix A, picture b) in Fig. 8.7 suggests that this secondary structure does not vary during the transition. It fluctuates indeed around the value 0.7\AA , standard deviation of 0.2\AA , for values of RMSD-RCL above 3\AA , and 0.3 in the region below this value. The average is obtained by calculating a weighted average of the values in the different regions, where the weights are defined by the density of states in each point. Standard deviation, in order to evaluate if the observed changes are not only thermal vibrations, is

(a) Metastable state of cleaved α_1 -AT

(b) Densities of states defined by the RMSD values of Helix A and RCL.



(c) Densities of states defined by the RMSD values of Helix F and RCL.

Figure 8.7: Results obtained from coarse grained MC simulations on α_1 -AT. The system reaches the final state.

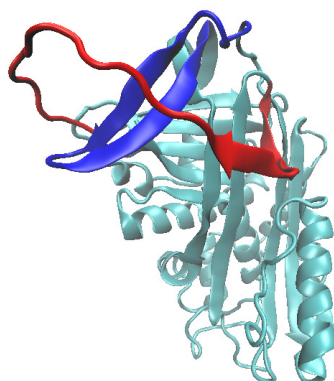
also calculated. In case of Helix F, Fig. 8.7 c) shows a different behaviour than Helix A that could be explained with a variation in the structure of Helix F during the transition: it is possible to note indeed that the RMSD value of Helix F varies from 0.7 Å, standard deviation of 0.2 Å, in the region where RMSD-RCL values are above 5 Å, to 0.8 Å, standard deviation of 0.1 Å, for RMSD -RCL between 3-5Å, and then finally it is 0.6 Å with a standard deviation of 0.1 Å at the end of the transition. There is an overlap between the values of the first two regions but this analysis suggests us that a change in the helix F-conformation occurs: it seems indeed that helix F, starting from a configuration comparable with the final one, assumes another configuration more distant from that and then returns in a conformation more similar than that reported in pdb.

What it is possible to conclude is that for the cleaved α_1 -AT we are able to simulate the reaction occurring in serpins in a one-bead description and with a pure $G\bar{o}$ model. The experimentally observed flexibility in Helix F is stressed also in our simulations, underlining the fact that the suggested

translocation or unfolding could likely occur in Helix F, allowing the RCL to overcome this obstacle on its path towards the final state. After the RCL passed a return of Helix F in the previous position is registered. On the contrary, Helix A does not seem to show a flexibility during the insertion of the RCL. The possible explanations for the observed difference with the experiments may be related to the fact that in the simulation we do not take into account the protease, that occur on the contrary in experiments [161], and that could cause more changes inside serpin in allowing the passage of the complex. However, another reason could be identified also in the used model: then, we tried to perform the same simulation with the addition of non- $G\bar{o}$ interactions. The same results emerges without any new information. However, the fact that the opening of the sheet A and the total insertion of RCL occurs also in the simulated case without a traslocation or an unfolding of Helix A suggests us that it is unlike that the role of Helix A is that of perturbing the surrounding region and allowing the opening of the sheet.

For a complete picture we try to simulate with MC method based on $G\bar{o}$ model also uncleaved α_1 -AT and PAI-1. The result is that the spontaneity of the process of PAI-1 is not reproduced at all (the simulation never reaches the latent state) while it seems that uncleaved α_1 -AT posses a high degree of spontaneity since all the probed trajectories achieve the final state. Including non- $G\bar{o}$ interactions through the $G\bar{o}$ -Hummer model only leads to slowing down the dynamics. This observation is in contrast with experiments, for this reason it is likely that a coarse grained topology-driven model, although developed in an improved manner thanks to the models of Karanicolas and Brooks [22] and Kim and Hummer [90] is not able to take into account all the clashes that is very probable that occur given the insertion of the RCL inside two strands of sheet A. It is also probable that this effect could be lower in case of cleaved α_1 -AT, since there is only a part of RCL that moves towards the bottom of the serpin. This part is shorter and certainly more flexible than an intact RCL that, for inserting, has to overcome the so called “gate region” and passes very close to Helix A, see Fig. 8.8 a) for a better explanation. The fact that Helix A in PAI-1 is longer than that in α_1 -AT, 4 residues more, and that according to native-centric model it is not flexible, could be the reason why PAI-1 never reaches the latent state. Fig. 8.8 shows an anticipation of the coming results of DRP simulations. As it will be clear, they represent a proper description of the reaction and are useful now for clarifying the statement about Helix A: in fact, b) represents the situation of Helix A at the beginning of the insertion and c) the movement of the helix in order to allowing the passage of the RCL.

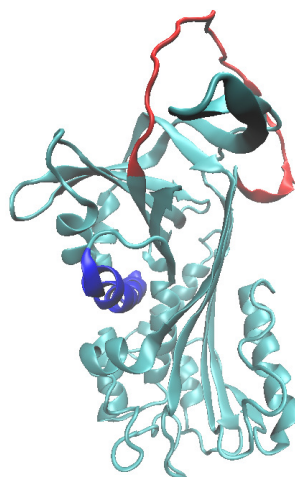
Coarse grained MC simulations based either on only native interactions or on native and non native interactions can describe the reaction for a system with a lower degree of clashes and higher level of spontaneity, that



(a) The RCL, colored in red, passes the gate region, blue.



(b) Helix A, in blue, as it appears at the beginning of the insertion of the RCL, red.



(c) In order to allowing the passage of intact RCL Helix A changes conformation.

Figure 8.8: Particulars of the insertion of the RCL. These results are from DRP simulations of PAI-1 explained in a detailed manner in the following.

is what can be found in cleaved α_1 -AT, although not all the experimental evidences are reproduced, as the case of the motion of Helix A, but this represents also the limit of this method. However, this result is important because it provides another proof of the flexibility of Helix F and opens a debate around the role of Helix A. In order to overcome the difficulties of a coarse grained description we performed simulation at the all-atom detail with a realistic potential and based on the rMD-DRP approach.

8.4.2 Complete all-atom DRP simulations of the latency transition in serpins.

In this paragraph the first atomistic simulations of the active-to-latent transition occurring in serpins are presented.

First of all, the reactions for PAI-1 wild type and mutants are analyzed. DRP trajectories for all the three serpins show a complete transition that leads the intact RCL from the top to the bottom of the protein, as expressed in Fig. 8.9, where the RMSD values in respect of the latent state as a function of the Euclidean distance are reported.

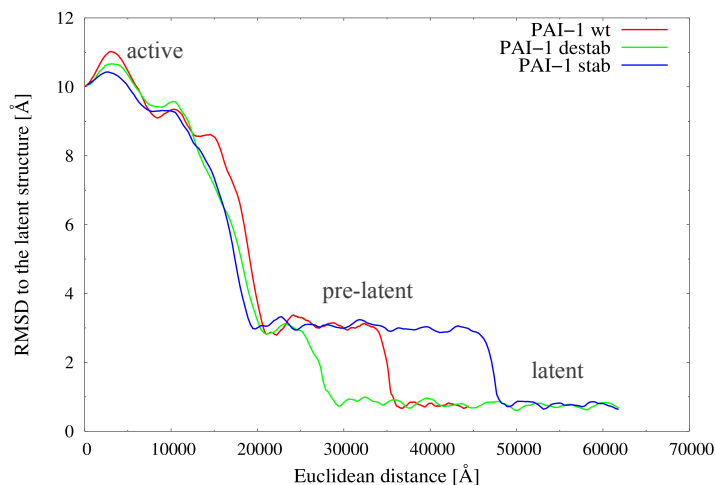


Figure 8.9: RMSD to the latent state as a function of the Euclidean distance, for all three PAI-1 variants.

The plots show a similar behaviour for the three serpins: this suggests us that the conformational changes in PAI-1 and mutants are the same. We chose to represent the RMSD plot as a function of the euclidean distance since it is the natural unit provided by the program. However, it is possible to connect the distance covered by a configuration in the space coordinates with the time. To be more precise, the program prints a configuration every 300 Å covered (this value is setted before running the program), while the energy is reported every 25 integration time steps (value setted by the user). In this way it is possible to connect the euclidean values with the time and a linear function is found. We emphasize that the accelerated MD scheme which is used to construct the variational set of trial paths in the DRP algorithm distorts all time scales. Hence the nominal time units have not a direct physical interpretation. Clearly, the underlying assumption is

that the distortion of time scales is the same for all the chains, given the high similarity in sequence, structures and in the events that lead to the latent state. As a consequence of this consideration and since the angular coefficient has a value very close for all the three serpins, every observation about the behavior vs. euclidean distance can be transposed to the time. From the analysis of Fig. 8.9, it results that the reactions occur with different time-scales, which are qualitatively in agreement with the experimental results: $\tau_{destab} < \tau_{WT} < \tau_{stab}$.

DRP trajectories of Fig. 8.9 show that the active to latent transition proceeds through a pre-latent intermediate, revealed by a clear plateau in all three curves. How long PAI-1 stays in this intermediate, that is a measure of the stability of this state, decides the time scale of all the reaction. Fig. 8.10 schematizes the principal configurations visited by PAI-1 WT during the reaction. These snapshots, given the previous considerations, are the same for the mutants, the difference is in the time the serpin resides in the intermediate.

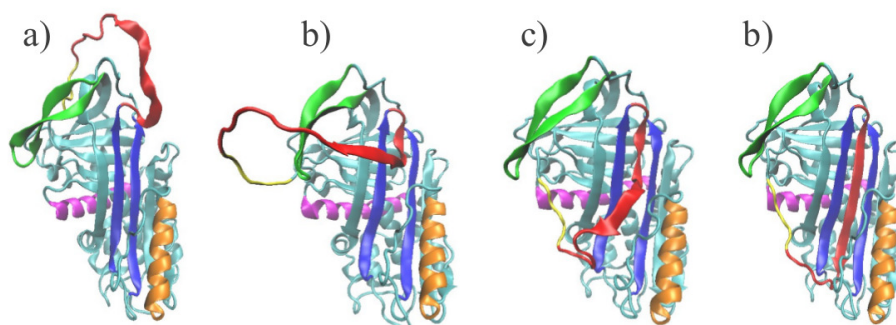


Figure 8.10: The PAI-1 transition from DRP simulations: a) from the metastable state the RCL begins insertion, b), then it is evidence of a pre-latent intermediate presented in c) until the latent state, d), characterized by fully insertion of the RCL.

In Fig. 8.10 the RCL is colored in red, Helix F in orange, the two principal strands in sheet A in blue and Helix A in violet. Strand 1C is also reported, colored in yellow, since it is involved in the huge modification inside serpin: as a consequence of the insertion of the RCL it has been detached from sheet C, the sheet behind sheet A. The particular of Fig. 8.10 c) represents the intermediate state. We found that in this state the RCL is partially inserted into sheet A up to P12 or P11. A similar pre-latent state has been previously proposed based on the results of fluorescence and antibody binding experiments [162, 170]. In order to find a support for our observations in the experimental results, our collaborator prof. P.

Wintrode tried a docking simulation with a molecule, AZ3976, that has been verified to accelerate the active to latent transition [170]. It has been suggested that this inhibitory function of AZ3976 in respect of PAI-1 is carried out by binding to a pre-latent state in a pocket between Helices F and E [170] and preventing in this way the return to the active conformation. In this way the pre latent state is seen as in equilibrium with the active state and the molecule has the role to change the equilibrium towards the latent state. Wintrode's simulations support this mechanism: indeed, by taking the DRP generated pre-latent structure it is shown that AZ3976 places in the experimentally identified pocket. On the contrary, AZ3976 could not be docked to active PAI-1. Moreover, MD simulations performed with Gromacs package and starting from the DRP simulated pre latent intermediate shows that serpin is not trapped in this state but moves towards the active state. All these results from DRP and pure MD simulations and computational docking are in agreement with experimental data strongly and consequently support the existence, structural configuration and implications of the pre-latent state. By controlling the stability of the pre-latent state, with binding of other ligands, it is possible in principle regulate the activity of PAI-1.

On the contrary, there is no evidence of such intermediate state in un-cleaved α_1 -AT, as reported in Fig. 8.11. Since the rMD algorithm acts on the dynamics of the protein by biasing the system towards the latent state, we are able to see the complete reaction, although this reaction should not be spontaneous.

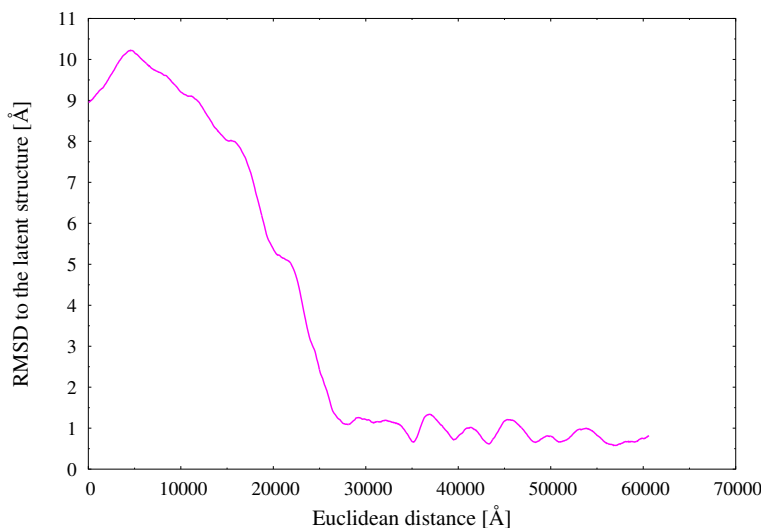


Figure 8.11: DRP simulation of latency transition for α_1 -AT.

A derivation of a time scale comparable with that of PAI-1 is not possible,

because the different number of residues and dynamics let us to expect that the bias will distort the time scale in a different manner than PAI-1. However, for comparing the results between the two types of serpins and check whether our simulations take into account the different degree of spontaneity of the reaction, very helpful is the plot in Fig. 8.12, where the potential energy as a function of the Euclidean distance covered by the C_α atoms along the reaction pathway is represented.

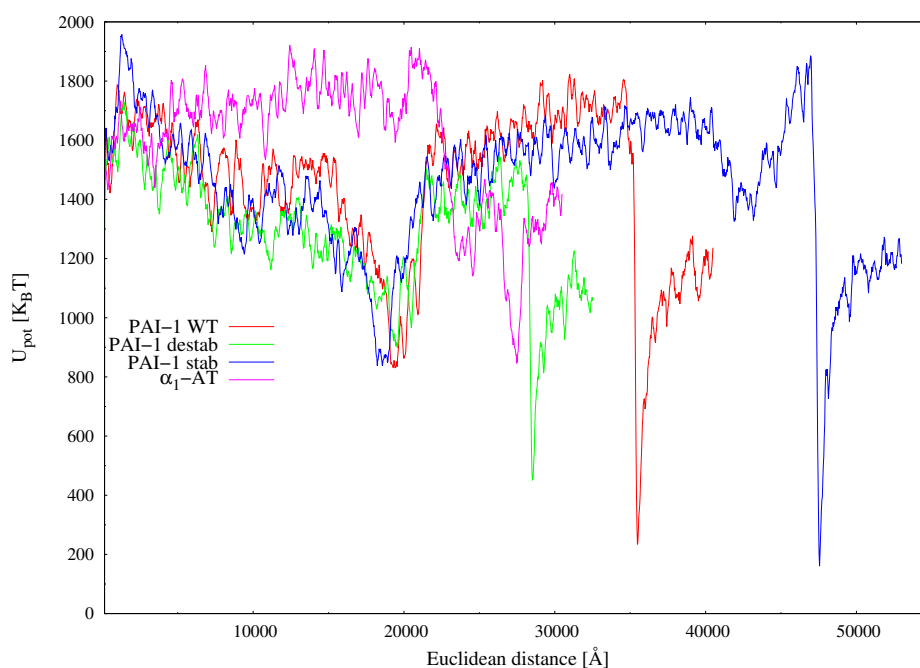


Figure 8.12: The potential energy along the dominant paths as a function of the Euclidean distance, for the three PAI-1 variants and for α_1 -AT.

From a theoretical point of view, characterizing the kinetics involves computing both energy and entropy changes. However, given the high degree of structural similarity and that the motion of insertion of RCL occurs in every cases, it is expected that entropy variations should not affect the observed kinetic differences that it is likely to be driven by potential energy. By focusing on potential energy variations we note that, by observing Fig. 8.12, the PAI-1 and α_1 -AT DRP energy profiles along the reaction pathway reflect the fact that the two serpins possess different degree of stability in the active state. Indeed, for α_1 -AT, in violet, the overall energy change remains unfavorable for almost the entire reaction and the energy landscape does not display a deep minimum compatible with an on-pathway pre-latent intermediate. On the contrary, the plots for all the three PAI-1 mutants show a clear intermediate state, reached at around 20000 Å, as suggested also by

Fig. 8.9. Then, the system has to overcome a barrier in order to reach the latent state, identified by a suddenly drop in energy until the stable state.

8.4.3 Kramers-Arrhenius analysis of reaction kinetics

If the high level of stability in active state of α_1 -AT can be explained by only observing the unfavorable energy profile, for what concerns PAI-1 WT and mutants the energy profile are very similar except the moment in which they reach the latent state. In the following we will try to derive a semi-quantitative calculation about the reaction rates in order to compare our results, obtained from simulations carried on at the nominal temperature of 300K, with those from experiments, performed at 310K. Since the difference in temperature is thought to modify the rates but the real temperature at which the simulation works is not known, present atomistic models are indeed known to yield unfolding temperatures which are only accurate within about a 5% tolerance, so that the nominal temperature of the DRP simulation may not correspond to the physical temperature, we will try to derive, on the basis of our results, the real temperature and the ratio of rates of the mutant versions of PAI-1 in respect of WT at the experimental temperature.

As suggested by other works [170, 162] and on the basis of our results we can consider the pre-latent and active state in equilibrium and the transition from the pre-latent to the latent state rate-limiting. Under these considerations the reaction kinetics could be treated as two-state (from the pre latent to the latent state) and Kramers-Arrhenius theory may be used to relate energy barriers and rates.

The Kramers' formula for rate when the treated system is subjected to a bias potential reads:

$$k^R(T) = k_0 \cdot e^{-\frac{(\Delta G^{PL-V})}{KT}}, \quad (8.1)$$

where R is the index that indicates that the rate is referred to a biased system, the ratchet-and-pawl algorithm in the particular case, V is introduced in order to modeled the effect of the bias that accelerates the dynamics and is expected to be invariant for all the three serpins, k_0 is a prefactor, ΔG is the free-energy barrier between the states pre-latent (identified by P), and latent (L).

If the indices *stab* and *WT* identify the PAI-1 stab and PAI-1 WT, respectively, then the ratio between the rates is:

$$\frac{k_{stab}^R(T)}{k_{WT}^R(T)} = \frac{k_0 \cdot e^{-\frac{\Delta G_{stab}^{PL}}{KT}} e^{\frac{V}{KT}}}{k_0 \cdot e^{-\frac{\Delta G_{WT}^{PL}}{KT}} e^{\frac{V}{KT}}}. \quad (8.2)$$

In this way it is possible to observe that the ratio between the rates of the biased system is the same as the ratio between rates of a non-biased system. In the following we will neglect the index R. If T_1 indicates the experimental temperature of 310K and T the real temperature of the simulation, then Eq. 8.2 can be rewritten:

$$\frac{k_{stab}(T) k_{stab}(T_1) k_{WT}(T_1)}{k_{WT}(T) k_{stab}(T_1) k_{WT}(T_1)} = \frac{k_{stab}(T_1) k_{stab}(T) k_{WT}(T_1)}{k_{WT}(T_1) k_{stab}(T_1) k_{WT}(T)}. \quad (8.3)$$

And Eq. 8.3 becomes:

$$\frac{k_{stab}(T)}{k_{WT}(T)} = \frac{k_{stab}(T_1)}{k_{WT}(T_1)} \frac{e^{-\frac{\Delta U_{stab}^{PL}}{KT} - \frac{NKT}{2KT} + \Delta S_{stab}^{PL}} e^{-\frac{\Delta U_{WT}^{PL}}{KT_1} - \frac{NT_1}{2KT_1} + \Delta S_{WT}^{PL}}}{e^{-\frac{\Delta U_{stab}^{PL}}{KT_1} - \frac{NKT_1}{2KT_1} + \Delta S_{stab}^{PL}} e^{-\frac{\Delta U_{WT}^{PL}}{KT} - \frac{NKT}{2KT} + \Delta S_{WT}^{PL}}}, \quad (8.4)$$

where ΔU^{PL} and ΔS^{PL} are the energy and the entropy, respectively, N is the number of the degrees of freedom of the system. From Eq. 8.4 it is possible to obtain:

$$\begin{aligned} \frac{k_{stab}(T)}{k_{WT}(T)} &= \frac{k_{stab}(T_1)}{k_{WT}(T_1)} e^{\alpha(T) \frac{\Delta U_{stab}^{PL}}{KT_0}} e^{-\alpha(T) \frac{\Delta U_{WT}^{PL}}{KT_0}} \\ &= \frac{k_{stab}(T_1)}{k_{WT}(T_1)} e^{\alpha(T) \left[\frac{\Delta U_{stab}^{PL} - \Delta U_{WT}^{PL}}{KT_0} \right]} \end{aligned} \quad (8.5)$$

where $\alpha(T) = T_0 \frac{-T_1 + T}{T T_1}$, where T_0 is the nominal temperature of the simulation, 300K. If we replace *stab* with *destab* we obtain the same expression also for PAI-1 destabilized. The term on the left side of Eq. 8.5 can be estimated directly from the results of the DRP simulations, by measuring the waiting time in the pre-latent state. This was done by using the evolution of the RMDS to latent for the different PAI-1 variants shown in Fig. 8.9. We obtain, for both mutants:

$$\frac{k_{stab}(T)}{k_{WT}(T)} \simeq \frac{1}{2}, \quad \frac{k_{destab}(T)}{k_{WT}(T)} \simeq 3 \quad (8.6)$$

The potential energy differences on the right-side of Equation 8.5 can be obtained from the DRP energy profiles plotted in Fig. 8.12. For this purpose it is necessary to identify the top of the energy barrier. We chose to consider as transition point the last relative maximum in the transition region immediately before the energy profile begins its monotonic decrease

to the global minimum corresponding to the latent state. Once the DRP ratio of rates and the energy barriers have been estimated, we are able to extract the physical temperature T of the DRP simulation using as input the experimental values for the ratios of reaction rates. In this way, we obtain: for PAI-1 destabil a temperature of $T \simeq 312K$ while for PAI-1 stab $T \simeq 314K$, suggesting the analysis based on the present simple model is quite robust. These values for the physical temperature are about 5% away from the nominal temperature, hence within the expected accuracy of the atomistic model. We used the average of the two temperatures, $\bar{T} \simeq 313K$, for estimating both ratios at 310K, leading to the results shown in Table 8.1.

PAI-1 type	half-life	Exp. k(mut)/k(WT)	DRP k(mut)/k(WT)
Wilde Type	2.0 h [165], 55 m [164]	-	-
PAI-1 stab	145 h [165]	0.014	0.004 ± 0.03
PAI-1 destabil	6.1 m [164]	9.0	15 ± 7

Table 8.1: The experimental half-times measured at 310K for the three PAI-1 serpins are shown. Then, the ratios of the rates of the relative kinetics of the active to latent transition for the PAI-1 mutants in respect of that of the wilde type are reported as from experiments, indicated in Table by Exp., and DRP simulations, DRP in Table.

The errors reported in table 8.1 for the DRP predictions have been obtained taking into account for the uncertainty on the physical temperature parameter T . An additional source of uncertainty, which has not been included in the error estimate, comes from the different prescriptions to identify the transition state in the pre-latent to latent transition, for example by averaging all the peaks in the region before energy drops until latent state. Given these facts and the admitted simplicity of the kinetic model adopted, our predictions should only be regarded as an order-of-magnitude estimate of the increase or decrease of life-times as a function of temperature with respect to the wild-type PAI-1 serpin.

8.4.4 Characterization at the atomistic detail of the conformational changes in serpins during latency transition.

At this point we know that we are able to properly simulate the latency transition, by reproducing the different rates for mutants PAI-1 and α_1 -AT and observing only for PAI-1 an intermediate, fact supported by experimental results. In this paragraph we will try to describe the events during the

latency transition.

At the beginning of the reaction we observe that the full opening of strands 3A and 5A, in blue in Fig. 8.13, occurs. Simultaneously, also the detachment of s1C, in yellow, from sheet C, behind sheet A, is observed.

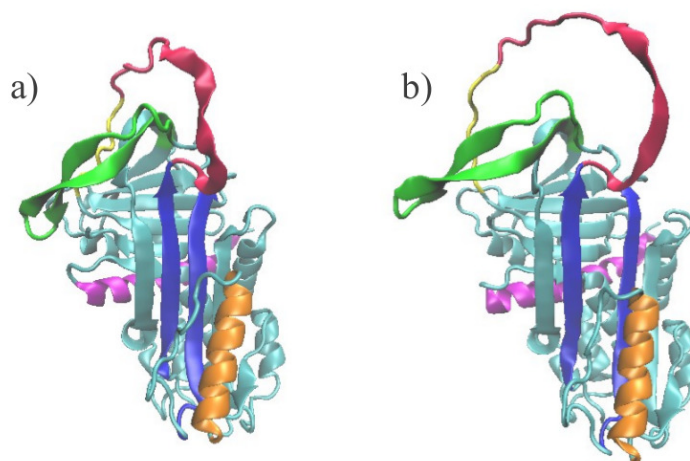


Figure 8.13: DRP snapshots of PAI-1 WT. In panel a) the reaction is at its very beginning and sheet A, blue, is close. Then, the RCL, red, has not yet started inserting that full sliding of sheet A, blue, is observed. At the same time stand 1C is detached from its secondary structure.

We would have expected, on the basis of serpin structures with partially inserted RCLs, as can be observed in Fig. 8.2, that an initial separation at the top of sheet A, followed by further “unzipping” of strands 3 and 5 A might occur. However, the explanation that the native state with partial insertion of RCL is stabilized in a different way than the latency transition occurs can be supported by experimental results that, on the contrary, seem to verify our DRP observations. Indeed, it has been suggested that, thanks to X-ray crystal structures and mutagenesis data, vitronectin increases the half life of PAI-1 by binding on it and sterically hindering the sliding of strands 1A and 2A into the gap between helices D and E which would also allow strand 3A to slide, fully opening sheet A [171]. Vitronectin should thus decrease the probability of the latency transition by blocking s1C detachment and the active to pre-latent transition.

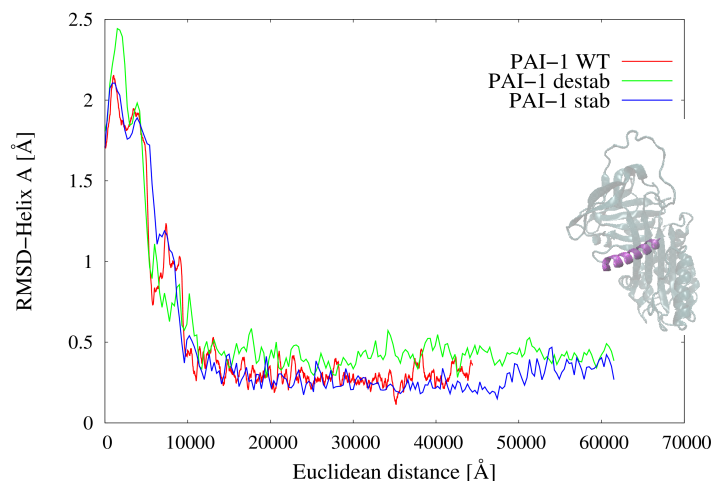
In the early stages of the transition an interesting role is played by Helix A, as witnessed by Fig. 8.14, where a normalized RMSD is reported as a function of Euclidean distance for PAI-1 mutants and α_1 -AT. If Helix A is of the same residue-length in the three version of PAI-1, in case of α_1 -AT is 4 residues shorter. So, in order to compare rmsd values related to different-size systems we use a normalized version of RMSD, as reported in

[172] and based on the observation that similar rmsd values have a different significance if referred to proteins of variuos-lenght chain. The formula that we used for a chain of N residues reads:

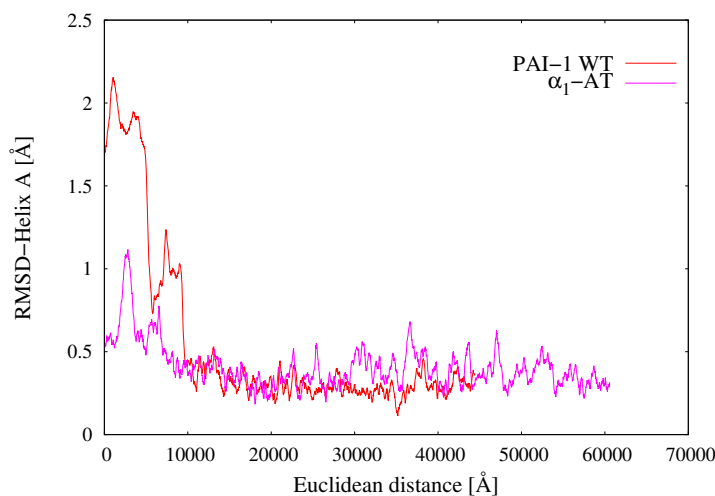
$$RMSD_{norm} = \frac{RMSD}{1 + \ln\sqrt{\frac{N}{L}}}, \quad (8.7)$$

where L is the number of residues chosen as reference. In the work of Carugo et al. [172] L has been taken as 100, because this is the mean number of amino acids per domain in a protein, in the analysis reported below we study the beahaviour of the RMSD of a part of the protein, the alpha-helices, so we take $L = 24$, that is the size of the helix in PAI-1 family.

As it is possible to observe in Fig. 8.14, an increase in flexibility in Helix A seems to be related to a high level in spontaneity. Indeed, at the beginning of the reaction, until 10000 Å it is suggested that Helix A undergoes a large conformational change, showed by RMSD variation from higher values, around 2Å, to values lower than 0.5Å. This change could be an unfolding or a partial motion and can be clerly evaluated among PAI-1 proteins, expecially in the case of PAI-1 destabil. In case of α_1 -AT the small variation in RMSD along the path let us hypothesize also a smaller conformational change of this region. Because sheet A opens in the very first steps of simulations and, according to MC simulations that do not reproduce the movements of Helix A but describe the opening of the sheet, it is unlikely, in our opinion, that Helix A could have a role in allowing strands A to slide, as on the contrary proposed in [161]. However, the reported results are in agreement with Tsutsui et al. [34], that think that Helix A could concurr in regulating PAI-1 activity. We think, on the basis of DRP simulations, that Helix A is a barrier for intact RCL in its path towards the latent state and that a huge change in conformation is needed in order to allowing RCL to overcome the helix. PAI-1, that can reach the latent state in a spontaneous way, has from evolution a high degree of flexibility in this part while α_1 -AT shows a more rigidity. Indeed, since this serpin in real world is stable in its active form for weeks and RCL inserts in sheet A only after protease cleavage Helix A is not interested in any changes because strand 1C and c terminal of RCL remains in their place. This analysis suggest another possibility for regulate PAI-1: by acting on Helix A.



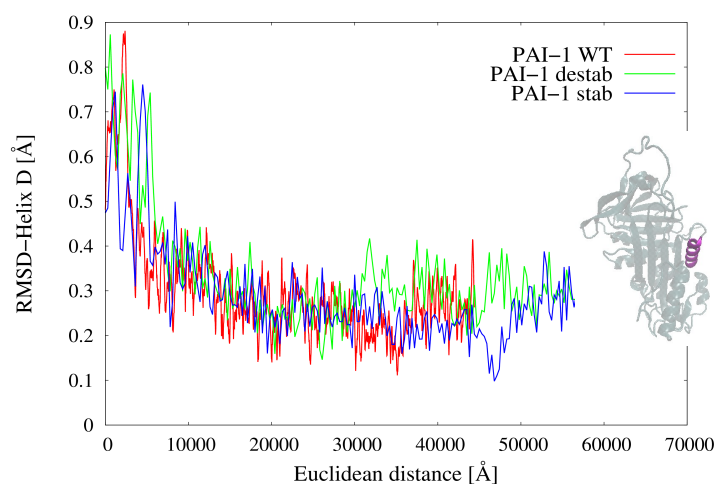
(a) RMSD to latent state restricted to Helix A as a function of the Euclidean distance for the PAI-1 mutants. On the right side of the panel the location of Helix A, in violet, inside serpin.



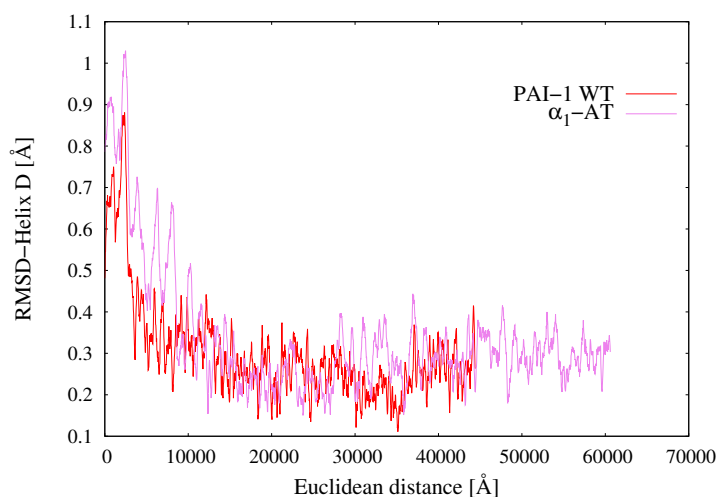
(b) RMSD to latent state restricted to Helix A as a function of the Euclidean distance for PAI-1 WT and α_1 -AT.

Figure 8.14: Study of the behaviour of normalized RMSD to latent state for Helix A during the simulation. Comparisons between PAI-1 mutants and α_1 -AT have been carried out.

Another interesting region, in the early stages of reaction, is represented by Helix D, as reported from H-D exchanges studies [34]. In Fig. 8.15 the results as obtained from DRP simulations are shown. Also in this case normalized RMSD values are calculated.



(a) RMSD to latent state restricted to Helix D as a function of the Euclidean distance for the PAI-1 mutants. On the right side of the panel the location of Helix D, in violet, inside serpin.



(b) RMSD to latent state restricted to Helix D as a function of the Euclidean distance for PAI-1 WT and α_1 -AT.

Figure 8.15: Study of the behaviour of normalized RMSD to latent state for Helix D during the simulation. Comparisons between PAI-1 mutants and α_1 -AT have been carried out.

As it is possible to note by observing Fig. 8.15, in all serpins a variation of the RMSD values as a function of the Euclidean Distance in the region until 10000 Å is noticed. It is interesting to consider that this conformational change occurring in Helix D seems to be higher in α_1 -AT. Given the complexity of the network of interactions it is difficult understand the role of Helix D but we think it is connected with a decrease in spontaneity, likely in making more unfavorable the opening of sheet A.

Then, on the basis of DRP simulations, the RCL passes the gate region, green, as reported in Fig. 8.16, captured in Fig. 8.16.

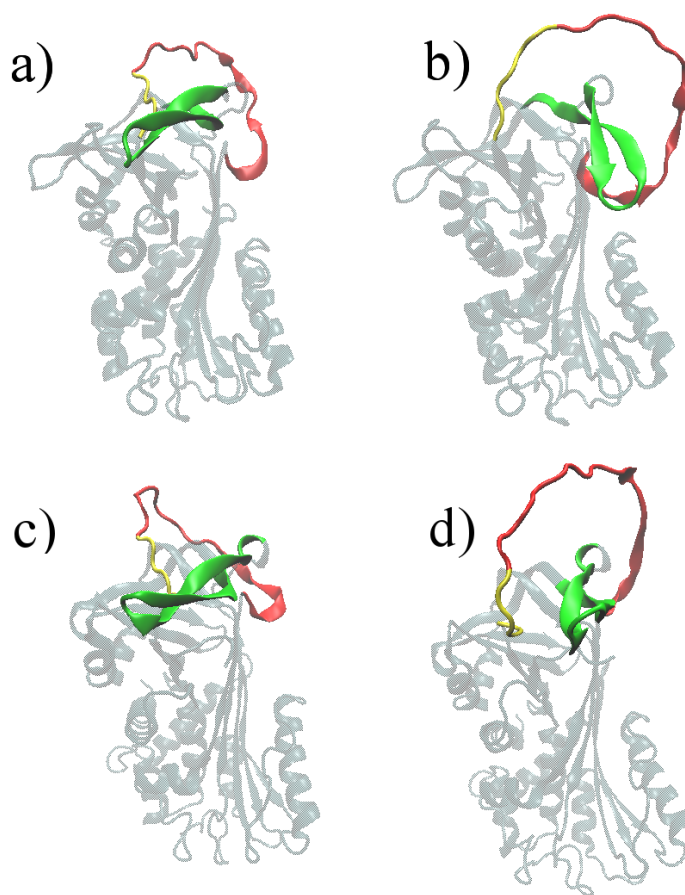


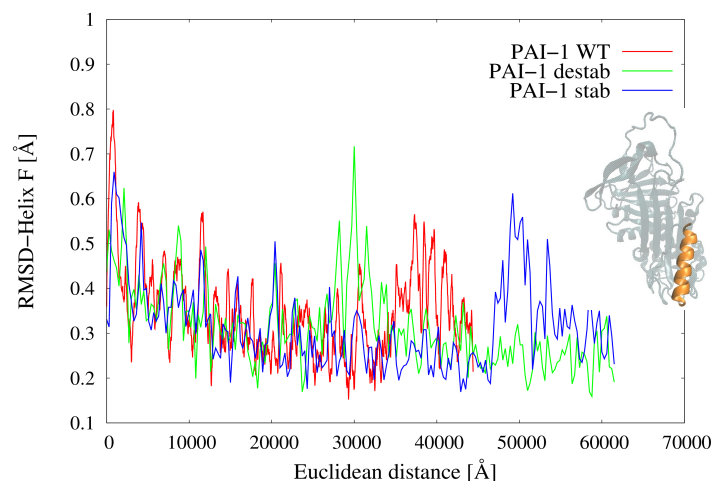
Figure 8.16: a) and c) panels report PAI-1 and α_1 -AT at the beginning of the reaction. b) and d) constitute a comparison between the moments of the overcoming of RCL in PAI-1 and α_1 -AT respectively.

Intact RCL, red, passes gate region, green, as reported in Fig. 8.16, in two distinct manner in respect of the type of serpin: in PAI-1 DRP

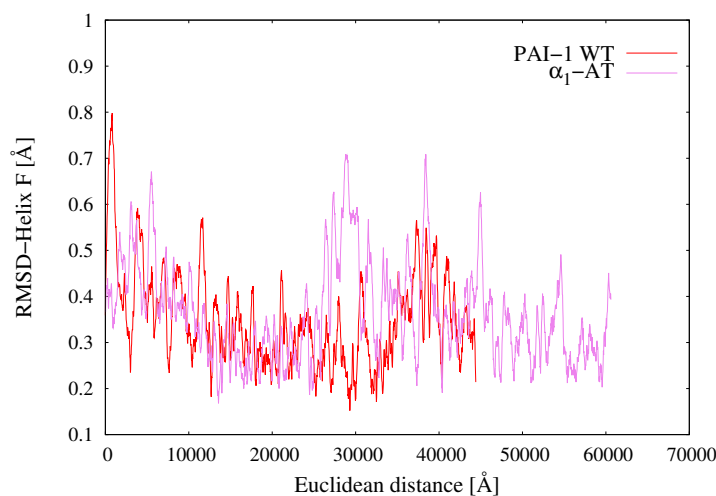
simulations suggest a high flexibility of gate region that undergoes a large bending, as shown in panel b). In contrast, for α_1 -AT it seems that a rigid push out from its normal position by the RCL is carried on. This observation is supported by previous hydrogen/deuterium exchange measurements that confirm the notion that the gate is more flexible in PAI-1 than in α_1 -AT [173, 174].

After this passage, the reaction proceeds downhill until the reaction coordinate, Euclidean Distance, reaches approximately 20000 Å, as from Fig. 8.9 and Fig. 8.12. At this point the RCL is partially inserted and the pre-latent intermediate is reached. The final stage of the transition involves displacing helix F and inserting the final residues of the RCL (P6-P4) into sheet A. Helix F returns to its position after the passage of RCL, see Fig. 8.17:

From Fig. 8.17 it is noteworthy that RMSD of Helix F undergoes an increase and successive decrease during the path. This displacement is in agreement, for what concerns PAI-1, with the time-scale of the mutants: this observation underlines the fact that Helix F is the last barrier for reaching the latent state. These processes are accompanied by a significant drop in energy, indicating that relatively subtle structural changes relieve a large amount of the conformational strain. Our simulations give therefore insight for what concerns the role of Helix F in serpins.



(a) RMSD to latent state restricted to Helix F as a function of the Euclidean distance for the PAI-1 mutants. On the right side of the panel the location of Helix F, in orange, inside serpin.



(b) RMSD to latent state restricted to Helix F as a function of the Euclidean distance for PAI-1 WT and α_1 -AT.

Figure 8.17: Study of the behaviour of normalized RMSD to latent state for Helix F during the simulation. Comparisons between PAI-1 mutants and α_1 -AT have been carried out.

8.4.5 Energy analysis of PAI-1 WT latency transition

Having characterized the reaction mechanism in a detailed manner and validated the DRP simulations against the experimental results, we can reliably identify the specific interactions that shape the underlying energy landscape and drive the latency transition for what regards PAI-1 WT. This analysis have been carried out by prof. Wintrode's group and is based on the decomposition of the potential energy into Coulomb and Van-der-Waals terms caclulated on the paths proveded by DRP. It emerges that the initial transition from active to pre-latent is driven primarily by the formation of favorable Coulombic interactions, amounting to a total favorable Coulomb of $\sim -11k_B T$, that is opposed on the contrary by van der Waals interactions, $E_{VW} \sim +5k_B T$). Interestingly, nearly 40% of the favorable Coulomb energy, $\sim -4.5k_B T$, is contributed by only a restricted set of residues, listed in Fig. 8.18, which are distributed throughout the structure.

These observed energy changes are almost quantitatively reversed in going from the pre-latent to the highest energy conformation, namely the maximum in energy profile before energy drops untile latent state, and reversed again in moving from this conformation to the latent state. In all these three stages of the reaction, nearly 40% of the associated energy changes can be attributed to the same subset of residues that are found to dominate the native to pre-latent transition. Most of these residues have previously been shown to play a significant role in function and/or disease, Fig. 8.18, supporting their importance and consequently DRP simulations.

PAI-1 Residue (Location)	Energetic Contribution A to PL ¹	Energetic Contribution PL to HE ¹	Effects of Mutations ²
Val17 (hA)	favorable	unfavorable	Some mutations reduce activity
Ser27 (turn hA to s6B)	favorable	unfavorable	Reported mutations are non-perturbing.
Val32 (s6B)	favorable	NS	No reported mutations. This residue is in the shutter domain. While no mutations are reported at this residue, mutations in the shutter domain are associated with serpin polymerization and disease.
Tyr37 (hB)	unfavorable	NS	Mutations can be deleterious.
Phe64 (loop hC to hD)	favorable	unfavorable	No reported mutations
Asp89 (N-term to s2A)	unfavorable	NS	No reported mutations.
Arg101 (C-term to s2A)	favorable	NS	Mutations affect vitronectin binding. Arg 101 is part of the flexible joint in the region that binds vitronectin and 98% of PAI-1 sequences have an arginine in this position.
Gln107 (loop s2A to hE)	NS	favorable	No reported mutations.
Leu152 (loop from hF to s3A)	unfavorable	NS	No reported mutations. While no mutations are reported at Leu152, this loop is important for RCL insertion, and mutations in this loop often have functional consequences.
Asn172 (N-term to s3A)	unfavorable	NS	No reported mutations
Lys176 (loop from s3A to sheet C)	favorable	unfavorable	Deleterious, effects on stability.
Arg186 (s4C)	favorable	unfavorable	Deleterious, effects on half-life. Arg186 is located in the gate.
Phe208 (s3C)	unfavorable	favorable	No reported mutations. Phe208 is adjacent to the Asn209 glycosylation site in PAI-1 and glycosylation prolongs the half-life.
Glu225 (s2B)	unfavorable	NS	Deleterious, effects on half-life.
Gly264 (loop hH to s2C)	unfavorable	NS	No reported mutations. This residue is adjacent to the glycosylation site at Asn265 and glycosylation prolongs the half-life.

PAI-1 Residue (Location)	Energetic Contribution A to PL ¹	Energetic Contribution PL to HE ¹	Effects of Mutations ²
Thr267 (loop hG to s2C)	favorable	unfavorable	No reported mutations.
Pro270 (s2C)	favorable	unfavorable	Mutations can be deleterious, effects on half-life.
Lys277 (loop s2C to s6A)	favorable	NS	Deleterious, effects on folding.
Arg287 (hl)	favorable	NS	Deleterious, effects on folding, stability and activity.
Gln312 (loop hl to s5A)	unfavorable	favorable	No published mutations.
Gln319 (loop hl to s5A)	unfavorable	NS	Deleterious, effects on half-life.
Lys323 (s5A)	favorable	unfavorable	Deleterious, effects on half-life and protease inhibition
Ala335 (P12 in RCL)	unfavorable	NS	Deleterious, effects on activity. This residue is in the hinge region and is highly conserved in serpins.
Arg346 (P1 in RCL)	favorable	NS	This is the P1 residue in the RCL, i.e., the residue N-terminal to the protease cleavage site. This residue is therefore important for protease specificity.
Arg356 (loop s1C to s4B)	favorable	unfavorable	Deleterious, effects on folding and half-life. This residue is in the distal hinge region which is important for conformational changes.
Thr369 (loop s4B to s5B)	favorable	unfavorable	Deleterious, effects on folding and half-life.
Pro379 (C-terminus)	favorable	NS	Deleterious, effects on activity and folding. This Pro residue is very conserved.

Figure 8.18: Mutating residues identified as energetically important for the active (A) to pre-latent (PL) or PL to high energy state (HE). NS = not significant.

¹ Energies were determined from molecular dynamics simulations beginning from frame 1 (active) or frame 650 (pre-latent) extracted from the DRP simulation. Favorable and unfavorable changes in energetic contributions were identified using an energy cutoff of $-40k_B T$ and $+40k_B T$, respectively.

² The effects of mutations were primarily determined from the PAI-1 literature and then corroborated by reported mutations for other serpins. Corresponding mutations in other serpins were found by structurally aligning the first frame used for the DRP simulations with the following X-ray crystal structures from the protein databank (PDB): 1qlp for A1AT, 1qmn for a1-antichymotrypsin, 1e05 for antithrombin III, 2ceo for thyroxine-binding globulin.

8.5 Conclusions of this study

In order to study the complex dynamics occurring in serpins we applied first coarse grained Monte Carlo simulations with a $G\bar{\sigma}$ potential. After verified that this type of investigation cannot generate a fully corrected description of the mechanism, although able to reproduce the insertion of cleaved RCL in α_1 -AT and to provide a general picture of this process in which Helix A does not seem to play the suggested role, another simulation technique has been used. We applied the DRP approach to perform simulations based on rMD algorithm in order to characterize the reaction mechanism and the kinetics in three versions of PAI-1 and α_1 -AT with intact RCL. rMD-DRP method demonstrates to be not only able to reproduce properly all the events and time-scales during RCL insertion in perfect agreement with experiments, but allows also to note the presence of an intermediate state in PAI-1 and to clarify the role of Helix A and F. All these results allowed us to rationalize the existing experimental data into a unifying picture and to identify the interactions that play a key role in this and similar transitions in other serpins. These results should help identify ligands that modify the kinetics of the latency transition thus regulating PAI-1, in addition to the observations about the two helices. We emphasize that these are the most complex, slow and big-size protein conformational transitions simulated in atomistic detail to date.

Having assessed the accuracy and computational efficiency of the rMD-DRP the attention can be focused into other important issues, such as the study of folding of serpins, in order to understand the network of interactions that leads to the metastable state or, on the contrary, that constrasts this process. It has been observed, indeed, that point mutations in residue-chain can cause the bypassing of the metastable conformation in favour of a stable but inactive product, with dramatic effects due to the failure in carrying on protein functions [158]. Simulations on wilde type and mutants can give insight into this thematic. In addition, changes in the program are under attention with the purpose to allow also investigation of cleaved RCL.

Then, it is also possible to expect that this same approach could now be applied to study even larger systems and/or much slower reactions, thereby removing the gap between protein processes which are computationally accessible and those which are biologically relevant.

Chapter 9

General conclusions

This work treated topics related to dynamics and thermodynamics of peculiar proteins that, for reasons such as large size, slow reactions, different behaviour despite the high level of structural similarity, represent a challenging task to be studied.

We started from a general topic that deals with the characterization of three-state folding process via an on-pathway intermediate, studied by simulating folding for two homologous proteins IM7 and IM9. The apparently different process of folding of these proteins has been previously indicated to be shaped until native state by non native interactions, challenging the concepts of *funneled energy landscape* and *minimal frustration* that are thought to describe the folding of the most of proteins. Then, the study analyzes the set of changes that proteins develop in order to adapt to cold environment, investigated by performing a comparison of the unfolding thermodynamics of two classes of pheromones living in cold and temperate waters, respectively, with the aim of testing the hypothesized role played by non native interactions. Finally the investigation deals with a more particular argument related to detailed characterization of conformational changes occurring in serpin family. In particular, the attention focused on types of serpins, very similar in structure but different in sequence. This difference can be of a few points, in case of PAI-1 wild type and the two mutants PAI-1 stab and PAI-1 destab, which show a half-time 72-fold greater and 9-fold smaller in respect of wild type, respectively, or less close as in the case α_1 -antitrypsin compared with PAI-1.

With the help of computer simulations based on Monte Carlo algorithm, Molecular Dynamics as well as a more advanced simulation technique developed by Dr. Pietro Faccioli's group at University of Trento and represented by rMD-DRP simulations we tried to approach these thematic.

It results that a native-centric view based on the concepts of minimal frustration and funneled energy landscape alone could be able to treat most

of the proposed problems, even a complete and proper description of three-state folding, showed to occur in both IM7 and IM9 that are only different in the degree of stability of this intermediate, and arrangements in proteins in cold environment, identified in the topology and in particular, on the basis of our results, by the location of CYS-CYS bonds along the chain. However, it emerges also that this view is challenged in extreme cases in which the funneled energy landscape of protein is highly distorted, as happens in serpins, whose native state is a metastable state and the stable state is reached only after a dramatic change in conformation. Nevertheless, Monte Carlo simulations based on native-centric model could outline a coarse picture that allows us to define the problem, that is solved in a detailed manner by rMD-DRP simulations with a realistic potential, providing a deep investigation of a reaction that normally occurs in hours or more, well beyond the possibilities of simulation methods to date. For what concerns the serpin project, a more particular consequence has been also found. From simulations and analysis of movements of secondary structures inside serpins and residues that principally take part to the conformational change significant implications in medicine research, in particular in drug design, have been obtained by providing a scheme to identify ligands whose binding can modify the reaction of PAI-1.

These achieved results are a breakthrough in some questions related to protein folding, but, despite this, a lot of work remains to do with the improving and upgrading of techniques for simulating dynamics of more and more big systems, also in complex. This could have implications in understanding general aspects of protein folding but could also allow to apply the research to topics such as folding of serpins, that is very sensitive to mutations, conformational changes in other proteins, for example a study of the dynamics in individual domains of the 70 kDA heat shock protein HSP70 is planned, but also and especially aggregation of proteins.

Appendix A

The generalized Born model

The equation for calculating G_{pol} [70, 175, 176]:

$$G_{pol} = \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^n \sum_{j>1}^n \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp\left(\frac{-r_{ij}^2}{4b_i b_j}\right)}} \quad (\text{A.1})$$

is based on the Born equation for the free energy of the solvation of a gaseous ion [70, 175]:

$$\Delta G_{born} = -\frac{1}{2} \left(1 - \frac{1}{\epsilon}\right) \frac{q^2}{b}, \quad (\text{A.2})$$

where q is the charge of the ion, b the radius of the ion and ϵ the dielectric of the medium.

This expressions is derived by considering the well known potential of a point charge [176]:

$$V(r) = \frac{e}{4\pi\epsilon_0\epsilon r}. \quad (\text{A.3})$$

This can be applied to an ion of radius b . The appropriate potential when a charge dq is brought from infinity to the surface of the ion at $r = b$ is [176]:

$$V(b) = \frac{e}{4\pi\epsilon_0\epsilon b}. \quad (\text{A.4})$$

Then:

$$dG_{charging} = V de = \frac{e}{4\pi\epsilon_0\epsilon b} de, \quad (\text{A.5})$$

that has to be integrated from 0 charge to the final charge of the ion, q :

$$\Delta G_{charging} = \frac{1}{4\pi\epsilon_0\epsilon b} \int_0^q e \, de = \frac{q^2}{2\epsilon b} \frac{1}{4\pi\epsilon_0}. \quad (\text{A.6})$$

By setting, for sake of clarity, the term $\frac{1}{4\pi\epsilon_0}$ equal to 1 and by calculating the change in energy of a charge upon transfer from a medium with a low dielectric, ϵ_1 , to a medium with a high dielectric, ϵ_2 , we obtain [176]:

$$\Delta G_{born} = -\frac{q^2}{2b\epsilon_1} + \frac{q^2}{2b\epsilon_2} = -\frac{q^2}{2b} \left[\frac{1}{\epsilon_1} - \frac{1}{\epsilon_2} \right], \quad (\text{A.7})$$

in vacuum $\epsilon_1 = 1$ and Eq. A.2 is obtained.

And now, the next step consists in generalizing the Born equation. If we consider a system of particles with radii b_i and charge q_i , G_{pol} is given by the sum of Coulomb energy and the Born free energy of solvation [70, 175, 177]:

$$G_{pol} = \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{\epsilon r_{ij}} - \frac{1}{2} \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \frac{q_i^2}{b_i}, \quad (\text{A.8})$$

where r_{ij} represents the distance between the atoms i and j . The first term on the right can be rewritten in the following way [177]:

$$\sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{\epsilon r_{ij}} = \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}} - \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}}. \quad (\text{A.9})$$

By inserting eq. A.9 in eq. A.8 we obtain:

$$G_{pol} = \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}} - \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}} - \frac{1}{2} \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \frac{q_i^2}{b_i} \quad (\text{A.10})$$

And, since $\Delta G_{pol} = G_{pol} - G_{pol}(\epsilon = 1)$, the expression becomes [177]:

$$\Delta G_{pol} = - \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}} - \frac{1}{2} \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \frac{q_i^2}{b_i} \quad (\text{A.11})$$

Still et al. [178] combined the two terms into a single expression:

$$\Delta G_{pol} = -\frac{1}{2} \left(1 - \frac{1}{\epsilon} \right) \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{f(r_{ij}, b)}, \quad (\text{A.12})$$

where:

$$f(r_{ij}, b) = \sqrt{\left(r_{ij}^2 + b_i b_j \exp\left(-\frac{r_{ij}^2}{4b_i b_j}\right)\right)}, \quad (\text{A.13})$$

is an effective distance between the atoms. If the Born radii decrease, namely the atoms are less buried, the electrostatic interactions are screened because surrounded by high dielectric medium and the effective distance increases.

An important thing in using GB-methods is to correctly estimate the Born radius for a given atom i inside a molecule. Our implementation uses the Onufriev-Bashford-Cane model for calculating the Born radius [72]:

$$b_i^{-1} = \tilde{\rho}_i^{-1} - \rho_i^{-1} \tanh\left(\alpha\Psi - \beta\Psi^2 + \gamma\Psi^3\right), \quad (\text{A.14})$$

where:

ρ_i is the VDW radius

$$\tilde{\rho}_i = \rho_i - 0.09\text{\AA}$$

$$\Psi = I\tilde{\rho}_i$$

$$I = \frac{1}{4\pi} \int_{VDW} \theta(|\vec{r}| - \tilde{\rho}_i) \frac{1}{r^4} d^3\vec{r}$$

α, β, γ are adjustable dimensionless parameters to be optimized, a choice could be: $\alpha = 1.0, \beta = 0.8, \gamma = 4.85$

Appendix B

$G\bar{o}$ model: the unfolding temperature and the native contact energy

If we assume that the protein has two dominant states, the native state and the unfolded state, there is a temperature, T_f , at which both states are equally populated [85]. In this situation the free energy of the two states has to be equal, as expressed in Fig. B.1.

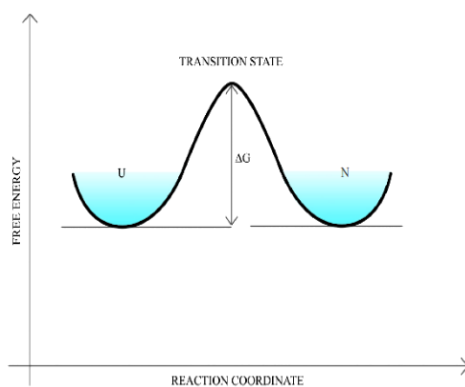


Figure B.1: 2D representation of the energy landscape of a two-state protein folding at the temperature of unfolding with the equal probability to find the protein in the unfolded or folded state.

The free energy is equal in both the states:

$$E_n - T_f S_n = E_u - T_f S_u, \quad (\text{B.1})$$

where E is the internal energy, S represents the entropy and the letters n

and u identify the native state and the unfolded state, respectively. In the $G\bar{o}$ model assumption we can indicate that the difference in energy between the two states is due to the number of native contacts formed, namely fully formed in the native state and none in the unfolded state [85].

$$E_n - E_u = T_f (S_n - S_u) = - \sum_{k=1}^Q \varepsilon_k, \quad (\text{B.2})$$

where k is index over the Q native contacts. If N is the number of residues and $\omega = k_B \log(n_n/n_u)$ the entropy difference between the native and unfolded state of a residue, being n_n the density of states in the native ensemble and n_u the density of states in the unfolded ensemble, then from Eq. B.2 follows:

$$- \sum_{k=1}^Q \varepsilon_k = T_f N \omega. \quad (\text{B.3})$$

From calculations of T_f for a wide range of residues with a fixed value of $\sum_{k=1}^Q \varepsilon_k$, it results that $\omega = 0.0054 \text{kcal/molK}$. Finally, $\varepsilon_{res} = \sum_{k=1}^Q \varepsilon_k / N = 0.0054 \times T_f$ [85].

Bibliography

- [1] F.H. Portugal and J.S. Cohen. *A century of DNA*. MIT Press, Cambridge and London, 1979.
- [2] K. Huang. *Lectures On Statistical Physics And Protein Folding*. World Scientific Publishing, Singapore, 2005.
- [3] Nature Education. Gene expression. <http://www.nature.com/scitable/topicpage/gene-expression-14121669>.
- [4] P. Echenique. Introduction to protein folding for physicists. *Contemporary Physics*, 48:81–108, 2007.
- [5] R. L. Baldwin. Weak interactions in protein folding: Hydrophobic free energy, van der waals interactions, peptide hydrogen bonds, and peptide solvation. In *Protein Folding Handbook*, pages 127–162. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008.
- [6] M. Klapper. On the nature of the protein interior. *Biochimica Et Biophysica Acta - Protein Structure*, 229:557–566, 1971.
- [7] J. Chen and W. E. Stites. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry*, 40:15280–15289, 2001.
- [8] J. A. Schellman. The thermodynamics of urea solutions and the heat of formation of the peptide hydrogen bond. *Comptes rendus des travaux du laboratoire Carlsberg*, 29:223–229, 1955.
- [9] S. Lifson, A. T. Hagler, and T. Dauber. Consistent force field studies of hydrogen-bonded crystals. 1. carboxyl acids, amides, and the C=O—H hydrogen bonds. *Nature Structural and Molecular Biology*, 101:5111–5121, 1979.
- [10] J. A. Huntington. Shape-shifting serpins—advantages of a mobile mechanism. *Trends in biochemical sciences*, 31:427–35, 2006.

-
- [11] C. M. Dobson. Protein-misfolding diseases: Getting out of shape. *Nature*, 418:729–730, 2002.
- [12] J. W. Kelly. Towards an understanding of amyloidogenesis. *Nature Structural and Molecular Biology*, 9:323–5, 2002.
- [13] J. C. Kendrew. Myoglobin and the structure of proteins. *Science*, 139:1259–66, 1963.
- [14] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
- [15] C. M. Dobson. The nature and significance of protein folding. In *Mechanisms of Protein Folding*, pages 1–33. Oxford University Press, Oxford, 2000.
- [16] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, Pathways, and the Energy Landscape of Protein Folding. *Proteins: Structure, Function and Genetics*, 21:167–195, 1995.
- [17] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal’s paradox. *Proceedings of the National Academy of Sciences USA*, 89:20–22, 1992.
- [18] C. Levinthal. How to Fold Graciously. pages 22–24, Illinois, 1969. University of Illinois Press.
- [19] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48:545–600, 1997.
- [20] L. Shu-Qun, J. Xing-Lai, T. Yan, T. De-Yong, Z. Ke-Qin, and F. Yun-Xin. Protein folding, binding and energy landscape: A synthesis. In *Protein Engineering*, pages 207–252. InTech, Rijeka, Croatia and Shanghai, 2009.
- [21] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14:70–5, 2004.
- [22] J. Karanicolas and C. L. Brooks. Improved go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *Journal of Molecular Biology*, 334:309–325, 2003.
- [23] F. Chiti, P. M. White, M. Bucciantini, F. Magherini, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Structural and Molecular Biology*, 6:1005–1009, 1999.
-

-
- [24] T. Škrbić, C. Micheletti, and P. Faccioli. The role of non-native interactions in the folding of knotted proteins. *PLoS computational biology*, 8:e1002504, 2012.
- [25] S. a Beccara, T. Škrbić, R. Covino, C. Micheletti, and P. Faccioli. Folding pathways of a knotted protein with a realistic atomistic force field. *PLoS Computational Biology*, 9:31003002, 2013.
- [26] S. J. Hagen. Solvent viscosity and friction in protein folding dynamics. *Current protein & peptide science*, 11:385–95, 2010.
- [27] S. Al-Karadaghi. Experimental methods in structural biology. <http://www.proteinstructures.com/Experimental/experimental-methods.html>.
- [28] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry, Fifth Edition*. W.H. Freeman, New York, 2002.
- [29] W. M. Elshemey, A. A. Elfiky, and W. A. Gawad. Correlation to protein conformation of Wide-angle X-ray Scatter parameters. *The protein journal*, 29:545–50, 2010.
- [30] S. B. Ozkan, I. Bahar, and K. A. Dill. Transition states and the meaning of Phi-values in protein folding kinetics. *Nature Structural and Molecular Biology*, 8:16–18, 2001.
- [31] S. M. Kelly and N. C. Price. The use of circular dichroism in the investigation of protein structure and function. *Current protein & peptide science*, 1:349–84, 2000.
- [32] Y. Hamuro, S. J. Coales, M. R. Southern, J. F. Nemeth-Cawley, D. D. Stranz, and P. R. Griffin. Rapid Analysis of Protein Structure and Dynamics by Hydrogen/Deuterium Exchange Mass Spectroscopy. *Journal of Biomolecular Techniques*, 14:171–182, 2003.
- [33] D. Weis. Hydrogen-deuterium exchange mass spectrometry. <http://mvsc.ku.edu/content/hydrogen-deuterium-exchange-mass-spectrometry>.
- [34] Y. Tsutsui, A. Sarkar, and P. L. Winthrode. Probing Serpin Conformational Change Using Mass Spectrometry and Related Methods. *Methods in Enzymology*, 501:325–350, 2011.
- [35] K Truong and M Ikura. The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. *Current opinion in structural biology*, 11:573–8, 2001.
- [36] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, USA, 1999.
-

- [37] L. Javidpour. Computer Simulations of Protein Folding. *Computing in Science and Engineering*, 14:97–103, 2012.
- [38] W. F. Van Gunsteren, M. Karplus. Protein dynamics in solution and in a crystalline environment: a molecular dynamics study. *Biochemistry*, 21:2259–2274, 1982.
- [39] M. Vendruscolo, C. M. Dobson. Protein dynamics: Moore’s law in molecular biology. *Current biology*, 21:R68–70, 2011.
- [40] R. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–52, 2012.
- [41] T. J. Lane, D. Shukla, K. A. Beauchamp, V. S. Pande. To milliseconds and beyond: challenges in the simulation of protein folding. *Current opinion in structural biology*, 23:58–65, 2013.
- [42] V. A. Voelz, G. R. Bowman, K. Beauchamp, V. S. Pande. Molecular Simulation of ab initio Protein Folding for a Millisecond Folder NTL9(1-39). *Journal of the American Chemical Society*, 132:1526–1528, 2010.
- [43] Pande Lab. Folding@home. <http://folding.stanford.edu/home/>.
- [44] V. A. Voelz, M. Jäger, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, and V. S. Weiss, S. annd Pande. Slow unfolded-state structuring in acyl-coa binding protein folding revealed by simulation and experiment. *Journal of the American Chemical Society*, 134:12565–12577, 2012.
- [45] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341–6, 2010.
- [46] K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–20, 2011.
- [47] S. Piana, K. Lindorff-Larsen, D. E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences USA*, 109:17845–50, 2012.
- [48] R. B. Best. Atomistic molecular simulations of protein folding. *Current opinion in structural biology*, 22:52–61, 2012.
- [49] R. B. Best. A "slow" protein folds quickly in the end. *Proceedings of the National Academy of Sciences USA*, 110:5744–5, 2013.

-
- [50] S. Piana, K. Lindorff-Larsen, D. E. Shaw. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences USA*, 2013.
- [51] D. Hamelberg, J. Mongan, J. A. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, 120:11919–29, 2004.
- [52] B. J. Grant, A. A. Gorfe, J. A. McCammon. Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS computational biology*, 5:e1000325, 2009.
- [53] B. J. Grant, A. A. Gorfe, J. A. McCammon. Large conformational changes in proteins: signaling and other functions. *Current opinion in structural biology*, 20:142–7, 2010.
- [54] Y. Shan *et al.* A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences USA*, 106:139–44, 2009.
- [55] S. Yang, N. K. Banavali, B. Roux. Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proceedings of the National Academy of Sciences USA*, 106:3776–81, 2009.
- [56] L. R. Masterson, A. Mascioni, N. J. Traaseth, S. S. Taylor, G. Veglia. Allosteric cooperativity in protein kinase A. *Proceedings of the National Academy of Science USA*, 2007, 2008.
- [57] L. Skjaerven, B. Grant, A. Muga, K. Teigen, J. A. Mccammon, A. Martinez. Conformational Sampling and Nucleotide-Dependent Transitions of the GroEL Subunit Probed by Unbiased Molecular Dynamics Simulations. *PLoS Computational Biology*, 7:e1002004, 2011.
- [58] C. F. Abrams, E. Vanden-Eijnden. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National Academy of Sciences USA*, 107:4961–6, 2010.
- [59] D. M. Rosenbaum *et al.* Structure and Function of an Irreversible Agonist-b2 Adrenoceptor complex. *Nature*, 469:236–240, 2011.
- [60] C. Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Current opinion in structural biology*, 18:10–5, 2008.
- [61] NIH. Theoretical and computational biophysics group, poster gallery 2006. <http://www.ks.uiuc.edu/Overview/gallery/posters/>.
-

- [62] N. V. Dokholyan. Studies of folding and misfolding using simplified models. *Current opinion in structural biology*, 16:79–85, 2006.
- [63] S. Matysiak and C. Clementi. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *Journal of molecular biology*, 363:297–308, 2006.
- [64] V. Tozzini. Lesson 1. http://homepage.sns.it/tozzini/public_files/StructuralComputationalBiology08/Lezione1.pdf.
- [65] V. Tozzini. Coarse-grained models for proteins. *Current opinion in structural biology*, 15:144–50, 2005.
- [66] T. Bereau and M. Deserno. Generic coarse-grained model for protein folding and aggregation. *The Journal of chemical physics*, 130:235106, 2009.
- [67] S. Takada. Coarse-grained molecular simulations of large biomolecules. *Current opinion in structural biology*, 22:130–7, 2012.
- [68] C. Hetenyi. Lesson 6. http://xray.bmc.uu.se/csaba/lecture_notes_6.pdf.
- [69] J. Zavadlav. *All Atom and Coarse Grained DNA Simulation Studies*. University of Ljubljana, Faculty of Mathematics and Physics, 2012.
- [70] A. Onufriev. Chapter 7 implicit solvent models in molecular dynamics simulations: A brief overview. volume 4 of *Annual Reports in Computational Chemistry*, pages 125 – 137. Elsevier, Amsterdam, 2008.
- [71] T. T. Pham, U. D. Schiller, J. R. Prakash, and B. Dünweg. Implicit and explicit solvent models for the simulation of a single polymer chain in solution: Lattice boltzmann versus brownian dynamics. *The Journal of Chemical Physics*, 131:164114, 2009.
- [72] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55:383–94, 2004.
- [73] B. Jayaram, D. Sprous, and D. L. Beveridge. Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field. *The Journal of Physical Chemistry B*, 102:9571–9576, 1998.
- [74] A. D. Mackerell. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, 25:1584–604, 2004.

- [75] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, and S. Brook. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins*, 725:712–725, 2006.
- [76] J. Wang, P. Cieplak, and P. A. Kollman. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *The Journal of Computational Chemistry*, 21:1049–1074, 2000.
- [77] B. R. Brooks and *et al.* CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30:1545–1614, 2009.
- [78] W. Scott and *et al.* The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A*, 103:3596–3607, 1999.
- [79] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110:1657–1666, 1988.
- [80] A. Korkut and W. A. Hendrickson. A force field for virtual atom molecular mechanics of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106:15667–72, 2009.
- [81] S. Prof. Marrink. Martini, coarse grain force field for biomolecular simulations. <http://md.chem.rug.nl/cgmartini/> .
- [82] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Advances in protein chemistry*, 66:27–85, 2003.
- [83] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78:1950–8, 2010.
- [84] H. Taketomi, Y. Ueda, and N. Go. Studies on protein folding, unfolding and fluctuations by computer simulation. *International Journal of Peptide and Protein Research*, 7:445–459, 1975.
- [85] J. Karanicolas, C. L. Brooks. The origins of asymmetry in the folding transition states of protein l and protein g. *Protein Science*, 11:2351–2361, 2002.
- [86] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

-
- [87] C. A. Andersen and B. Rost. Section IV: Secondary Structure Assignment. In *Structural Bioinformatics, 2nd Edition*, pages 459–487. Wiley-Blackwell, New York, 2009.
- [88] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, 256:623–44, 1996.
- [89] L. Wu, J. Zhang, M. Qin, F. Liu, and W. Wang. Folding of proteins with an all-atom Go-model. *The Journal of chemical physics*, 128:235103, 2008.
- [90] Y. C. Kim and G. Hummer. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *Journal of molecular biology*, 375:1416–33, 2008.
- [91] K. Kremer and K. Binde. Monte carlo simulation of lattice models for macromolecules. *Computer Physics Reports*, 9:259 – 310, 1988.
- [92] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient monte carlo method for the self-avoiding walk. *Journal of Statistical Physics*, 50:109–186, 1988.
- [93] Stephen R. Quake. Fast Monte Carlo algorithms for knotted polymers. *Physical Review E*, 52:1176–1180, 1995.
- [94] P. Faccioli, M. Sega, F. Pederiva, and H. Orland. Dominant pathways in protein folding. *Physical Review Letters*, 97:108101, 2006.
- [95] M. Sega, P. Faccioli, F. Pederiva, G. Garberoglio, and H. Orland. Quantitative protein dynamics from dominant folding pathways. *Physical Review Letters*, 99:118102, 2007.
- [96] E. Autieri, P. Faccioli, M. Sega, F. Pederiva, and H. Orland. Dominant reaction pathways in high-dimensional systems. *The Journal of Chemical Physics*, 130:064106, 2009.
- [97] G. Mazzola, S. a Beccara, P. Faccioli, and H. Orland. Fluctuations in the ensemble of reaction pathways. *The Journal of Chemical Physics*, 134:164109, 2011.
- [98] E. Pitard and H. Orland. Dynamics of the swelling or collapse of a homopolymer. *EPL (Europhysics Letters)*, 41:467, 1998.
- [99] D. Q. Nykamp. A saddle point of a function of two variables. http://mathinsight.org/applet/saddle_point_two_variables.
-

- [100] S. a Beccara, G. Garberoglio, P. Faccioli, and F. Pederiva. Ab-initio Dynamics of Rare Thermally Activated Reactions. *The Journal of Chemical Physics*, 132:111102, 111106, 2010.
- [101] S. a Beccara, T. Škrbić, R. Covino, and P. Faccioli. Dominant folding pathways of a WW domain. *Proceedings of the National Academy of Sciences USA*, 109:2330–5, 2012.
- [102] A. B. Adib. Stochastic actions for diffusive dynamics: reweighting, sampling, and minimization. *The journal of physical chemistry. B*, 112:5910–6, 2008.
- [103] P. Faccioli. Investigating biological matter with theoretical nuclear physics methods. *Journal of Physics: Conference Series*, 336:012030, 2011.
- [104] P. Eastman, N. Grønbech Jensen, and S. Doniach. Simulation of protein folding by reaction path annealing. *The Journal of Chemical Physics*, 114:3823, 2001.
- [105] C. Camilloni, R. A Broglia, and G. Tiana. Hierarchy of folding and unfolding events of protein g, ci2, and acbp from explicit-solvent simulations. *The Journal of Chemical Physics*, 134:045105, 2011.
- [106] E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology*, 288:441–459, 1999.
- [107] P. Bolhuis. Molecular rare events simulations. <http://molsim.chem.uva.nl/han/2008/Han-sur-Lesse-Bolhuis-2008a.pdf>.
- [108] S. a Beccara, P. Faccioli, M. Sega, F. Pederiva, G. Garberoglio, and H. Orland. Dominant folding pathways of a peptide chain from ab initio quantum-mechanical simulations. *The Journal of Chemical Physics*, 134:–, 2011.
- [109] P. Faccioli. Characterization of protein folding by dominant reaction pathways. *The Journal of Physical Chemistry B*, 112:13756–13764, 2008. PMID: 18855433.
- [110] L. Sutto, J. Lätzer, J. A. Hegler, D. U. Ferreira, and P. G. Wolynes. Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proceedings of the National Academy of Sciences USA*, 104:19825–30, 2007.
- [111] J. D. Bryngelson and P. G. Wolynes. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry*, 93:6902–15, 1989.

- [112] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal of molecular biology*, 298:937–53, 2000.
- [113] C. J. Wilson, P. Das, C. Clementi, K. S. Matthews, and P. Wittung-Stafshede. The experimental folding landscape of monomeric lactose repressor, a large two-domain protein, involves two kinetic intermediates. *Proceedings of the National Academy of Sciences USA*, 102:14563–8, 2005.
- [114] P. Das, C. J. Wilson, G. Fossati, P. Wittung-Stafshede, K. S. Matthews, and C. Clementi. Characterization of the folding landscape of monomeric lactose repressor: quantitative comparison of theory and experiment. *Proceedings of the National Academy of Sciences USA*, 102:14569–74, 2005.
- [115] Z. Zhang and H. S. Chan. Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proceedings of the National Academy of Sciences USA*, 107:2920–5, 2010.
- [116] M. Kouza, C. Hu, M. S. Li, and A. Kolinski. A structure-based model fails to probe the mechanical unfolding pathways of the titin I27 domain. *The Journal of Chemical Physics*, 139:1–31, 2013.
- [117] C. T. Friel, D. A. Smith, M. Vendruscolo, J. Gsponer, and S. E. Radford. The mechanism of folding of Im7 reveals competition between functional and kinetic evolutionary constraints. *Nature Structural and Molecular Biology*, 16:318–24, 2009.
- [118] R. Maiti, G. H. Van Domselaar, H. Zhang, and D. S. Wishart. SuperPose: a simple server for sophisticated structural superposition. *Nucleic acids research*, 32:W590–4, 2004.
- [119] N. Ferguson, A. P. Capaldi, R. James, C. Kleanthous, and S. E. Radford. Rapid folding with and without populated intermediates in the homologous four-helix proteins im7 and im9. *Journal of Molecular Biology*, 286:1597 – 1608, 1999.
- [120] A. P. Capaldi, M. C. Shastry, C. Kleanthous, H. Roder, S. E. Radford, and C. Kleanthous. Ultrarapid mixing experiments reveal that im7 folds via an on-pathway intermediate. *Nature Structural and Molecular Biology*, 8:68–72, 2001.

- [121] S. A. Gorski, A. P. Capaldi, C. Kleanthous, and S. E. Radford. Acidic conditions stabilise intermediates populated during the folding of Im7 and Im9. *Journal of Molecular Biology*, 312:849–63, 2001.
- [122] A. M. Figueiredo, G. R. Moore, and S. Whittaker. Understanding how small helical proteins fold: conformational dynamics of Im proteins relevant to their folding landscapes. *Biochemical Society transactions*, 40:424–8, 2012.
- [123] S. Whittaker, N. J. Clayden, and G. R. Moore. Nmr characterisation of the relationship between frustration and the excited state of im7. *Journal of Molecular Biology*, 414:511 – 529, 2011.
- [124] A. P. Capaldi, C. Kleanthous, and S. E. Radford. Im7 folding mechanism: misfolding on a path to the native state. *Nature Structural and Molecular Biology*, 9:209–16, 2002.
- [125] C. T. Friel, A. P. Capaldi, and S. E. Radford. Structural Analysis of the Rate-limiting Transition States in the Folding of Im7 and Im9: Similarities and Differences in the Folding of Homologous Proteins. *Journal of Molecular Biology*, 326:293–305, 2003.
- [126] E. Paci, C. T. Friel, K. Lindorff-larsen, S. E. Radford, M. Karplus, and M. Vendruscolo. Comparison of the Transition State Ensembles for Folding of Im7 and Im9 Determined Using All-Atom Molecular Dynamics Simulations With Value Restraints. 525:513–525, 2004.
- [127] S. D. Pugh, C. Gell, D. A. Smith, S. E. Radford, and D. J. Brockwell. Single-molecule studies of the Im7 folding landscape. *Journal of Molecular Biology*, 398:132–45, 2010.
- [128] J. Gsponer, H. Hopearuoho, S. Whittaker, G. R. Spence, G. R. Moore, E. Paci, S. E. Radford, and M. Vendruscolo. Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proceedings of the National Academy of Sciences USA*, 103:99–104, 2006.
- [129] C. L. Pashley, G. J. Morgan, A. P. Kalverda, G. S. Thompson, C. Kleanthous, and S. E. Radford. Conformational properties of the unfolded state of Im7 in nondenaturing conditions. *Journal of Molecular Biology*, 416:300–18, 2012.
- [130] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26:1701–1718, 2005.

- [131] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4:435–447, 2008.
- [132] F Delaglio, S Grzesiek, G W Vuister, G Zhu, J Pfeifer, and a Bax. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of biomolecular NMR*, 6:277–93, 1995.
- [133] S. E. Knowling, A. M. Figueiredo, S. Whittaker, G. R. Moore, and S. E. Radford. Amino Acid Insertion Reveals a Necessary Three-Helical Intermediate in the Folding Pathway of the Colicin E7 Immunity Protein Im7. *Journal of Molecular Biology*, 392:1074–1086, 2009.
- [134] R. Margesin, G. Neuner, and K. B. Storey. Cold-loving microbes, plants, and animals—fundamental and applied aspects. *Naturwissenschaften*, 94:77–99, 2006.
- [135] C. Alimenti, A. Vallesi, B. Pedrini, K. Wüthrich, and P. Luporini. Molecular cold-adaptation: comparative analysis of two homologous families of psychrophilic and mesophilic signal proteins of the protozoan ciliate, *Euplotes*. *IUBMB life*, 61:838–45, 2009.
- [136] A. Vallesi, G. Di Giuseppe, F. Dini, and P. Luporini. Pheromone evolution in the protozoan ciliate, *Euplotes*: the ability to synthesize diffusible forms is ancestral and secondarily lost. *Molecular phylogenetics and evolution*, 47:439–42, 2008.
- [137] G. Cazzolli, T. Škrbić, G. Guella, and P. Faccioli. Unfolding thermodynamics of cysteine-rich proteins and molecular thermal-adaptation of marine ciliates. *Biomolecules*, 3:967–985, 2013.
- [138] P. J. Hogg. Disulfide bonds as switches for protein function. *Advances in Protein Chemistry*, 28:210–214, 2003.
- [139] M. Geralt, C. Alimenti, A. Vallesi, P. Luporini, and K. Wüthrich. Thermodynamic stability of psychrophilic and mesophilic pheromones of the protozoan ciliate *Euplotes*. *Biology*, 2:142–150, 2013.
- [140] R. A. Goldstein. Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: Insights from the quasi-chemical approximation. *Protein Science*, 16:1887–1895, 2007.
- [141] D. C. Poland and Harold A Scheraga. Statistical Mechanics of Non-covalent Bonds in Polyamino Acids. VIII. Covalent Loops in Proteins. *Biopolymers*, 3:379–399, 1965.

-
- [142] V. I. Abkevich and E. I. Shakhnovich. What can Disulfide Bonds Tell Us about Protein Energetics, Function and Folding: Simulations and Bioninformatics Analysis. *Journal of Molecular Biology*, 300:975–988, 2000.
- [143] C. J. Camacho and D. Thirumalai. Theoretical predictions of folding pathways by using the proximity rule, with applications to bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences USA*, 92:1277–1281, 1995.
- [144] C. J. Camacho and D. Thirumalai. Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins*, 22:27–40, 1995.
- [145] T. E. Creighton, E. Kalef, and R. Arnon. Immunochemical analysis of the conformational properties of intermediates trapped in the folding and unfolding of bovine pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 123:129–147, 1978.
- [146] J. Clarke and A. R. Fersht. Engineered disulfide bonds as probes of the folding pathway of barnase: Increasing the stability of proteins against the rate of denaturation. *Biochemistry*, 32:4322–4329, 1993.
- [147] L. Whitmore and B. A. Wallace. Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers*, 89:392–400, 2008.
- [148] N. Go. Theoretical studies of protein folding. *Annual review of biophysics and bioengineering*, 12:183–210, 1983.
- [149] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of molecular biology*, 277:985–94, 1998.
- [150] R. W. Carrell and J. Travis. α 1-antitrypsin and the serpins: variation and countervariation. *Trends in Biochemical Sciences*, 10:20 – 24, 1985.
- [151] D. van Gent, P. Sharp, K. Morgan, and N. Kalsheker. Serpins: structure, function and molecular evolution. *The International Journal of Biochemistry & Cell Biology*, 35:1536–1547, 2003.
- [152] X. Zheng, P. L. Wintrode, and M. R. Chance. Complementary structural mass spectrometry techniques reveal local dynamics in functionally important regions of a metastable serpin. *Structure*, 16:38 – 51, 2008.

- [153] J. C. Whisstock, S. P. Bottomley, P. I. Bird, R. N. Pike, and P. Coughlin. Serpins 2005 - fun between the beta-sheets. Meeting report based upon presentations made at the 4th International Symposium on Serpin Structure, Function and Biology (Cairns, Australia). *The FEBS journal*, 272:4868–73, 2005.
- [154] M. S. Khan, P. Singh, A. Azhar, A. Naseem, Q. Rashid, M. A. Kabir, and M. Jainpur. Serpin Inhibition Mechanism: A Delicate Balance between Native Metastable State and Polymerization. *Journal of amino acids*, 2011:606797, 2011.
- [155] L. Liu, N. Mushero, L. Hedstrom, and A. Gershenson. Short-lived protease-serpin complexes: Partial disruption of the rat trypsin active site. *Protein Science*, pages 2403–2411, 2007.
- [156] J. Shin and M. Yu. Viscous drag as the source of active site perturbation during protease translocation: insights into how inhibitory processes are controlled by serpin metastability. *Journal of molecular biology*, 359:378–89, 2006.
- [157] G. A. Silverman and *et al.* The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *The Journal of biological chemistry*, 276:33293–6, 2001.
- [158] J. Huntington. The ins and outs of serpin polymerization. http://mbg.au.dk/fileadmin/site_files/mb/aktuelt/arrangementer/kjeldgaard-lectures/2011/Jim_Huntington.pdf.
- [159] B. Van de Craen, P. J. Declerck, and A. Gils. The Biochemistry, Physiology and Pathological roles of PAI-1 and the requirements for PAI-1 inhibition in vivo. *Thrombosis Research*, 130:576–585, 2012.
- [160] C. Lee, S. H. Park, M. Y. Lee, and M. H. Yu. Regulation of protein function by native metastability. *Proceedings of the National Academy of Sciences USA*, 97:7727–31, 2000.
- [161] J. Baek, W. S. Yang, C. Lee, and M. Yu. Functional unfolding of alpha1-antitrypsin probed by hydrogen-deuterium exchange coupled with mass spectrometry. *Molecular & cellular proteomics : MCP*, 8:1072–81, 2009.
- [162] D. M. Dupont and *et al.* Evidence for a pre-latent form of the serpin plasminogen activator inhibitor-1 with a detached beta-strand 1C. *The Journal of biological chemistry*, 281:36071–81, 2006.

- [163] J. K. Jensen, L. C. Thompson, J. C. Bucci, P. Nissen, P. G. W. Gettins, C. B. Peterson, P. A. Andreasen, and J. P. Morth. Crystal structure of plasminogen activator inhibitor-1 in an active conformation with normal thermodynamic stability. *Journal of Biological Chemistry*, 286:29709–29717, 2011.
- [164] M. Hansen, M. N. Busse, and P. A. Andreasen. Importance of the amino-acid composition of the shutter region of plasminogen activator inhibitor-1 for its transitions to latent and substrate forms. *European Journal of Biochemistry*, 268:6274–6283, 2001.
- [165] D.A. Lawrence M.B. Berkenpas and D. Ginsburg. Molecular evolution of plasminogen activator. *The EMBO Journal*, 14:2969–2977, 1995.
- [166] H. Im, E. J. Seo, and M. H. Yu. Metastability in the inhibitory mechanism of human alpha1-antitrypsin. *The Journal of biological chemistry*, 274:11072–7, 1999.
- [167] Y Tsutsui, R. Dela Cruz, and P. L. Wintrode. Folding mechanism of the metastable serpin α 1-antitrypsin. *Proceedings of the National Academy of Sciences USA*, 109:4467–72, 2012.
- [168] A. Fiser and A. Sali. Modloop: automated modeling of loops in protein structures. *Bioinformatics*, 19:2500–2501, 2003.
- [169] F. Lauck, C. A. Smith, G. F. Friedland, E. L. Humphris, and T. Kortemme. Rosettabackrub—a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Research*, 38:W569–W575, 2010.
- [170] O. Fjellstrom and *et al.* Characterization of a Small Molecule Inhibitor of Plasminogen Activator Inhibitor Type 1 that Accelerates the Transition into the Latent Conformation. *The Journal of biological chemistry*, 288:873–885, 2012.
- [171] A. Zhou, J. A. Huntington, N. S. Pannu, R. W. Carrell, and R. J. Read. How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration. *Nature Structural and Molecular Biology*, 10:541–4, 2003.
- [172] O. Carugo and S. Pongor. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, 10:1470–1473, 2001.
- [173] Y. Tsutsui, L. Liu, A. Gershenson, and P. L. Wintrode. The Conformational Dynamics of a Metastable Serpin Studied by Hydrogen Exchange and Mass Spectrometry. *Biochemistry*, 45:6561–6569, 2006.

-
- [174] M. B. Trelle, D. Hirschberg, A. Jansson, M. Ploug, P. Roepstorff, P. A. Andreasen, and T. J. D. Jørgensen. Hydrogen/Deuterium Exchange Mass Spectrometry Reveals Specific Changes in the Local Flexibility of Plasminogen Activator Inhibitor 1 upon Binding to the Somatomedin B Domain of Vitronectin. *Biochemistry*, 51:8256–8266, 2012.
- [175] P. Larsson. Generalized born. http://xray.bmc.uu.se/calle/md_phd/gb.pdf.
- [176] UCLA. The born model. http://voh.chem.ucla.edu/vohtar/spring05/classes/156/pdf/born%20solvation_from%20cherie.pdf.
- [177] S. Koppole. An introduction to continuum electrostatics. http://www.sampath.koppole.com/intro_elec.pdf.
- [178] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112:6127–6129, 1990.

Acknowledgements

First of all, I would to thank my advisor, Dr. Pietro Faccioli, who helped me in a difficult moment and made it possible for me to continue with the doctoral school, giving me a very fascinating project that led me to the USA, my dream since I was 10 years old.

Then, there are a lot of people I would to thank and I hope not to forget anyone: I thank Dr. Tatjana Skrbić, who teached me very patiently and answered all my questions. But Tatjana has been not only a teacher but also a friend for me and I thank her for conversations and important support. I would thank Dr. Silvio a Beccara, who helped me in understanding DRP and running simulations, sometimes a real challenge. Then, thanks to Professor Patrick Wintrode: it has been a pleasure to collaborate with him for his patiente, competence and enthusiasm. Thanks also to Professor Anne Gershenson and Dr. Fang Wang, that collaborate at the project of serpins.

Thanks to Dr. Giovanni Garberoglio and Enrico Tagliavini, who helped me with my problems with computers. Thanks to my colleagues of the PhD room at LISC and of FBK, who sometimes made me laugh sometimes made me angry but eventually teached me a lot about not only physics but also “men’s word”.

Finally, I have to thank two other persons, who are no longer here but helped me in the past years, Professor Gabriele Viliani and Dr. Paolo Verrocchio. They gave me enthusiasm in research and believed in me.