UNIVERSITÀ DEGLI STUDI DI TRENTO
Dipartimento di Fisica

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

Investigating Protein Folding Pathways at Atomistic Resolution: from a Small Domain to a Knotted Protein

Supervisor:
Prof. Pietro Faccioli

Candidate:
Roberto Covino

Dottorato di Ricerca in Fisica, XXVI ciclo
Trento, December 16th, 2013

Contents

Contents iii

Introduction vii

List of abbreviations xiii

1 The Protein Folding Problem 1

1.1 Proteins . . . 2

1.1.1 Interactions in proteins . . . 6

1.1.2 Structures . . . 7

1.2 The protein folding problem . . . 8

1.2.1 Proteins are self-assembling systems . . . 9

1.2.2 Investigating the mechanism . . . 10

1.2.3 Folding is described by a two-state kinetics . . . 11

1.2.4 Why is folding so fast? . . . 14

1.2.5 The folding thermodynamics . . . 15

1.2.6 An energy bias towards the native state . . . 19

1.2.7 Rough or smooth landscapes . . . 23

1.2.8 There are many diverse bottlenecks for folding . . . 24

1.2.9 The two views are not incompatible . . . 24

1.2.10 The origin of the funneled landscape . . . 25

1.3 Protein dynamics on a computer . . . 26

1.3.1 The solvent . . . 28

1.3.2 Empirical all-atom force fields 29
1.3.3 Gō-type models 31
1.3.4 All-atom MD simulations in the Anton era 33
1.3.4.1 Accuracy of current AA FF 34
1.3.5 Folding happens through sequential stabilization . . . 39
1.3.6 Role of non-native interactions 40

2 Simulating reactive folding pathways **43**
2.1 Stochastic action 44
2.1.1 Langevin equation 44
2.1.2 Smoluchowski equation 49
2.1.3 Wiener path integrals 51
2.1.3.1 Brownian trajectories are not differentiable . 56
2.1.4 Stochastic action functionals 57
2.2 Diffusion along a reaction coordinate 60
2.3 Characterizing the reactive folding pathways 64
2.3.1 The saddle-point approximation 65
2.3.2 DRP 68
2.3.3 Sampling the path space 70
2.3.4 Sampling and scoring 73
2.3.4.1 Characterizing the folding pathways: the algorithm 73

3 Folding a WW Domain **77**
3.1 Folding pathways of a WW Domain 77
3.1.1 Two folding pathways 79
3.1.2 Little role for non-native interactions 83
3.1.3 Locating the TS 83
3.1.4 Relative weight of the pathways 85
3.1.5 Varying the force 86
3.2 Comparison with experiments 88
3.3 Comparison with numerical investigations 91

3.4 Computational details . . . 92
3.4.1 Atomistic DRP simulations . . . 92
3.4.2 CG native-centric calculations . . . 93

4 Folding a knotted protein **95**
4.1 Knots in proteins . . . 96
4.1.1 Function and evolution . . . 97
4.1.2 Experimental characterization . . . 99
4.1.3 Computational approaches . . . 100
4.2 Folding the smallest knotted protein . . . 102
4.2.1 Characterizing the folding trajectories . . . 103
4.2.1.1 When the knot forms . . . 104
4.2.1.2 Measuring the pathway heterogeneity . . . 104
4.2.2 How the knot forms . . . 107
4.2.3 What happens when knotting fails . . . 109
4.2.4 Discussion: the role of non-native interactions . . . 110
4.2.4.1 Slipknotting vs. direct threading . . . 112
4.2.4.2 Turning non-native interactions on and off . . . 113
4.2.4.3 The role of non-native interactions . . . 114
4.2.5 Computational details . . . 116
4.2.5.1 DRP algorithm . . . 116
4.2.5.2 Coarse grained simulations . . . 117
4.2.5.3 Knot detection . . . 117

5 Milestoning **119**
5.1 The Milestoning algorithm . . . 120
5.2 Refolding a long myosin chain . . . 126
5.2.1 Marginally thermally activated transitions . . . 128
5.2.1.1 Case A (partially folded conformation) . . . 130
5.2.1.2 Case B (almost unfolded initial configuration) 136
5.2.2 Thermally activated transition . . . 139
5.2.3 Discussion . . . 144

5.2.4 Computational details 145

Conclusions **147**

Credits **151**

Acknowledgments **153**

Bibliography **155**

Introduction

> "The dance of life is spontaneous, self-sustaining, and self-creating."
> Paul Davies, *The Fifth Miracle: The Search for the Origin and Meaning of Life*

> "The reductionist hypothesis does not by any means imply a "constructionist" one: the ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe."
> Philip Anderson, *More is different*

One of the most remarkable features of life is its being a very complex self-organizing process. It is no less than astonishing how little and very simple creatures, like bacteria, mosquitoes or snails, are able to do exactly what they need to survive in a highly dynamical external environment. But even at a smaller level, the cell is an exceedingly dense crowd of molecular factories and machines, each accomplishing a complicated task and contributing altogether to form a highly tuned choreography. Proteins are the most important molecular devices in every living being, and are able to carry on an infinite series of different functions in virtue of their ability to self-assembly in a well defined three-dimensional structure.

All these biological systems are constantly drugged out of thermodynamical equilibrium by a flux of energy and matter. In fact, for a living being real equilibrium corresponds to death. Complex systems are by definition made of many components, which can be complex systems themselves, organized in hierarchical levels. In the simplest case, the evolution of each component can be described as a function which significantly varies on a typical scale. A

perfect ideal "microscope" would be able to measure any dynamical observables of a given system at any desired scale (it would be a perfect telescope if used to measure big things). Let us consider for example an observable varying over time. Any real measurement of a time dependent quantity is a discrete process, since the quantity is sampled every given time interval. The latter determines which phenomena we can measure and which we cannot, depending on the typical time scale on which a specific process significantly varies. For example, we could decide to observe the time evolution of a biological macromolecule, like a protein, in water. Thus, we would set the time resolution of our ideal microscope large enough to see little but significant variations in the collective motion of the protein. However, this happens on a time scale that is orders of magnitude larger than the typical variation of the position and velocity of the water molecules. In the time elapsed between a measurement and the following one, water molecules are able to explore their phase space, hence we perceive them at equilibrium. We can thus forget the real dynamics of water and treat it according to the average effect it acts on the molecule. On the other hand, if we measured with a time interval equivalent to the typical time scale of water dynamics, then we would find it out of equilibrium. The same picture holds for the evolution of electrons around a nucleus, where the former are much faster than the latter, and the decoupling of the timescales makes the well known Born-Oppenheimer approximation possible. From a radical different point of view, simple cosmological models are possible because matter can be treated as homogeneous at the length scales of the Universe. Not only stars and planets, but entire galaxies and everything else can be described as a constant mass density filling the Universe as a whole. This very important examples of scale separation illustrate that the concept of equilibrium itself depends on the scale on which we look at the system. Many of the quantitative approaches which nowadays yield a significant understanding of Nature are indeed possible thanks to this fact.

We shall refrain from giving a comprehensive definition of what is life, an indeed daunting task, but we will grab a part of it with safety saying that everything that is living is a very complex and self-organizing process, which stubbornly struggles to stay out of equilibrium.

The reductionism dogma inspires a first straightforward attempt to understand biological systems from a physical point of view. It is hard not to believe that any complex phenomenon is determined by the fundamental

laws of physics, which are nowadays fairly well known and settled, since there is not a single experiment on Earth violating the Standard Model of particle physics. However, as Anderson illustrated in his manifesto [1], deriving from first principles the behavior of a complex system, which is determined by such an overwhelmingly complicated combination of so many fundamental interactions, is practically impossible. "More is different", and any level of complexity is a new "fundamental" level on its own, whose description needs new concepts, new "elementary" objects which obey to "fundamental" laws [1]. Indeed, we will presumably never be able to understand what intelligence is in terms of electromagnetism, but it would be a great success to formulate an at least semi-quantitative effective theory of the brain.

The hope is that physical methods and concepts which are successful at a given time, space, or complexity scale, are also able to provide a useful insight at other scales. This is what happened for example with entropy, originally developed to measure the efficiency of real engines and now one of the most general concepts in quantitative sciences.

However, biological systems are also the result of a long evolutionary selection, which complicates the task to find new effective laws. Indeed, when we investigate a biological system, e.g., a protein, what we see could be the result of general principles, or of a long history of adaptations due to particular events, or, even worse, we could look at a "frozen accident". A physical approach would be presumably effective only in the first case, but distinguishing what is a general feature from what is a particular event is usually hard. Moreover, there is another risk, that is to invoke an evolutionary adaptation to explain a feature that is in fact the subtle result of some general principles, but we are just not able to find that out.

It is a fact that most of the proteins have to display a well defined three-dimensional structure in order to accomplish their biological functions. Small and simple proteins are produced as linear chains and spontaneously fold under physiological conditions to this functional structure, in what is a self-assembling process. This feature makes proteins the smallest biological self-organizing systems, thus an ideal target for a physical investigation.

Decades of experimental and theoretical approaches have produced a common conceptual framework, known as the free energy landscape theory, which states that proteins self-assembly efficiently and rapidly in virtue of an energy bias towards their functional structure. Within this scheme a simple

description of the folding mechanism is still missing, and several open issues are still debated. Among this, it is not clear whether proteins can fold to their functional form by many and diverse mechanisms, as suggested by the free energy landscape theory. In other words, is it true that "all roads lead to Rome"?

Molecular Dynamics simulations at atomistic resolution represent one of the most promising theoretical approach to investigate the theory of protein folding, at least from the theoretical point of view. A model of the molecule and the forces between atoms are considered on a computer, which solves a discrete representation in time of Newton's equations of the system. This technique has reached nowadays a maturity under different points of view, a fact that was also sealed with the Nobel Price in chemistry of this year (2013), which was awarded to Martin Karplus, Michael Levitt, and Arieh Warshel, "for the development of multi-scale models for complex chemical systems".

Despite this success, all-atom Molecular Dynamics still suffers from fundamental limitations. In particular, a sense of impotence rises when we realize that a huge supercomputer running for months and consuming enormous amounts of power will simulate few microseconds of the dynamics of a very small protein in water. This is a severe limitation, since proteins self-assembly on much longer timescales, on the order of milliseconds to minutes. Hence, in such a simulation one would unlikely be able to see a folding event. Molecular Dynamics simulations are so demanding because protein's dynamics is characterized by many relevant timescales, spanning over about twelve orders of magnitude. An hypothetical ideal microscope should be able to look at the system with a time resolution high enough to measure the fastest motion but long enough to appreciate also the slowest ones, i.e., folding itself. In the last few years an approximation of this ideal microscope has been built, the Anton supercomputer, which, thanks to highly specialized hardware and software, is able to simulate a protein in water on the millisecond scale. Folding of small proteins in realistic models have been repeatedly observed, and some issues of Molecular Dynamics have been driven away.

Even if we were able to simulate any protein in a reasonable time, would this mean that the we understand how the protein folds? The answer to this question crucially relies on one's philosophical opinions. However, if we agree that "understanding" means being able to explain with simple principles and having some predictive power, then the answer should probably be negative.

Many points have been clarified so far, a general principle (i.e., the free energy landscape theory) has been proposed, but a simple and predictive theory is still missing. The insight offered by Molecular Dynamics is invaluable, and simulations are more and more considered as a special microscope able to describe the dynamics with a resolution not accessible by experimental techniques yet.

Due to the need to rationalize observations on one hand, and to enhance the sampling over longer timescales on the other hand, many researchers have developed alternative approaches. Usually these techniques give up to a part of all the details contained in a long Molecular Dynamics simulations to focus on few particular aspects. However, not all details are relevant in the same way, and when we develop a simplified approach producing results compatible with experiments, then maybe we understand a bit more some aspects of our problem.

In this thesis we introduce and use a novel algorithm in order to characterize protein folding trajectories. We give up the power to predict the protein's functional structure and to measure physical intervals along the trajectory, but we gain a high efficiency in portraying the sequence of events by which the protein self-assemblies.

Although proteins are the smallest self-organizing biological systems, decades of investigation at the interface of biology, chemistry, physics, and computer science have just began to return a comprehensive picture. Much has still to be understood, and the effort to formulate a quantitative theory of protein folding will demand novel approaches going beyond the separation between old disciplines. Our humble hope is that the work presented in this thesis can be a small contribution in this direction.

This thesis is organized as follows:

Chapter 1 We will briefly review few fundamental facts about proteins and define the protein folding problem. Then, we will introduce and study the current conceptual framework which explains why folding is so fast. Some open issues will be outlined, in particular whether folding happens through parallel pathways and the role of non-native interactions, which will be investigated in this thesis. In the last part, we will explain what a Molecular Dynamics simulation is and review the current achievements and open prob-

lems.

Chapter 2 This chapter is devoted to introduce the method developed and used in this thesis in a self-contained way. We will study in detail the over-damped Langevin equation, which will be assumed to be a valid description for the dynamics of a protein in water. We will sketch the derivation of the Wiener path integral representation of the transition probability for a diffusive process, and introduce the Onsager-Machlup action functional. We will then review some recent results about the existence of a good reaction coordinate for folding, namely the fraction of native contacts. In the last section, we will introduce the Dominant Reaction Pathway (DRP) algorithm, a method to efficiently simulate the reactive folding pathways, which consists in a biased sampling of the folding trajectories that are then ranked according to their probability in the unbiased diffusive dynamics.

Chapter 3 We will use the DRP method introduced in chapter 2 to investigate the folding of a small 35-residue long WW domain, in all-atom resolution and with a realistic force field. This system is a benchmark for the algorithm, and we will show that only two different folding pathways emerges characterizing many microscopic trajectories. There are thus only two possible folding mechanisms, and we will show that this result is compatible with experimental and numerical investigations.

Chapter 4 We will employ the DRP algorithm to study the folding of a natively knotted protein. As a matter of fact, we will report the first case of an all-atom folding simulation of a knotted protein in a realistic force field. We will characterize the folding trajectories, finding that folding happens by a well defined sequence of events, and that knotting can occur through two main different mechanisms, determined by the effect of non-native interactions.

Chapter 5 We will simulate the refolding of a 126-residue long chain of the human cardiac myosin, and analyze the kinetics and thermodynamics of the trajectories by mean of the Milestoning algorithm. We will use the results to gain some insight of the bias' effect in DRP trajectories.

List of abbreviations

AA	All-Atom
CG	Coarse Grained
MC	Monte Carlo
MD	Molecular Dynamics
rMD	ratchet-and-pawl Molecular Dynamics
FF	Force Field
TS	Transition State
NS	Native State
DS	Denatured State
TP	Transition Path
TPT	Transition Path Time
MFPT	Mean First Passage Time
Q	Fraction of native contacts
RMSD	Root Mean Square Deviation (usually compared to a native structure)
CO	Contact Order
PDB	Protein Data Bank
IDP	Intrinsically Disordered Proteins
PDF	Probability Density Function

PDFl	Probability Density Functional
OM	Onsager-Machlup (action functional)
DRP	Dominant Reaction Pathway
NMR	Nuclear Magnetic Resonance
$\mathbf{C}_\alpha$	Central alpha carbon atom
MFP	Maximum (Probability)Flux Path

Chapter 1

The Protein Folding Problem

"Not all those who wander are lost"

J.R.R. Tolkien - *The Lord of the Rings*

In this chapter we will briefly introduce proteins, the main characters of this thesis, by collecting some basic experimental facts about their functions in every living organism, their structure and the relevant interactions responsible for their shape in water. This material is highly standard, and appears in many excellent text books, as for example [2, 3].

We then will extensively introduce the folding problem in the next section, defining the particular point of view we focus on. Folding is remarkably fast in nature, although it involves sampling an astronomically large space. In order to get insight on how this is possible, a historical view of how our understanding has proceeded is very useful. This is particularly true because, differently to many other fields in contemporary science, the currently used terminology, theoretical models and open issues are still deeply related to the first investigations. Plenty of evidences point to the fact that proteins have an energy bias to fold to the native state, where the interactions are optimized to cooperatively stabilize the protein configuration. This result forms the bulk of the current conceptual framework describing protein folding, and we will analyze some still debated issues which will be addressed in this thesis.

These open problems can be tackled by means of theoretical and computational approaches, as Molecular Dynamics simulations in particular, and

we will review the latest achievements in this field, which show that these techniques are mature enough to validly support experimental investigations.

In writing this chapter I took great advantage of Ref.'s [4] and [2], which I would strongly suggest to anyone is approaching this fascinating field for the first time.

1.1 Proteins

Proteins are biomolecules present in all the five kingdoms of life and perform a great variety of functions and complicate tasks, as we will illustrate with few examples. Some proteins are passive building blocks of many structural elements of living beings, as keratin in nails and hair, collagen in cartilages, or the external coats of viruses. They also carry out more active functions, like hormones, which transmit signals across the body, or antibodies, which defend the organism by the malicious threat of viruses and bacteria. Some proteins are natural springs and bundled together in fibrils are the basic constituents of muscles, which make possible for any organism to move and interact with its surrounding environment. Proteins with special shape are embedded in the cellular membrane forming a channel passing through it, and regulates which compounds can enter or exit. Other proteins control when to activate the expression of a gene, the procedure by which the information contained in the DNA strand is translated into RNA, which contains the instruction to build a new protein. These new proteins are assembled in molecular factories called ribosomes, which are themselves made of proteins (and RNA). And newly produced proteins can find shelter in another molecular machine, the chaperon, a barrel-shaped assembly of proteins that can be even closed paying an amount of ATP. Even more impressive protein-based cellular devices exist, which for example actively transport other compounds across the cell.

The amount of different tasks proteins carry out is simply astonishing, and is strictly related to the three dimensional structure that characterizes them. Indeed, experiments show thats that at physiological conditions almost half of all the different types of proteins displays a well defined compact structure, that we will call the native configuration[1]. As we will abundantly discuss, the process of getting this particular native conformations is called folding.

[1] Throughout this work, "conformation" and "configuration" will be used as synonyms.

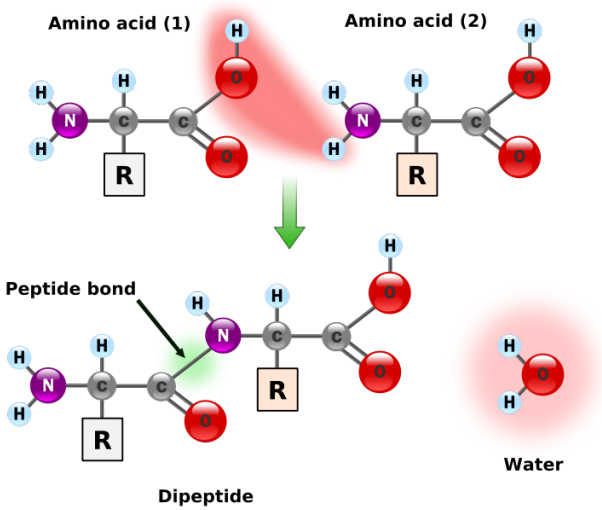


Figure 1.1: Peptide bond formation between two amino-acids. Fig. reproduced with permission from Wikipedia.

Proteins are heteropolymers, i.e., linear chains made of different types of "beads", which are the amino-acids. The shortest polypeptide chain with protein like properties is only 35 amino-acid long (the villin headpice), while on the other side giant proteins made of ∼30,000 amino-acid exist (the titin protein). The average chain length in Eukaryota is ∼300, sensibly bigger than in Prokaryota and Archea [5].

The chemical composition of amino-acids has a part that is always the same, the backbone, and a part that changes, the side-chain or residue. Different amino-acid form peptide covalent bonds linking themselves together (Fig. 1.1).

Amino-acids, which are also called simply residues, are characterized by the chemical and stereochemical properties of their side-chains, which range from the simplest one in glycine, that is simply an hydrogen atom, to very big and complex aromatic rings, as in tryptophan (Fig. 1.2). The sequence of resides defines the so-called primary structure of a protein (Fig. 1.3).

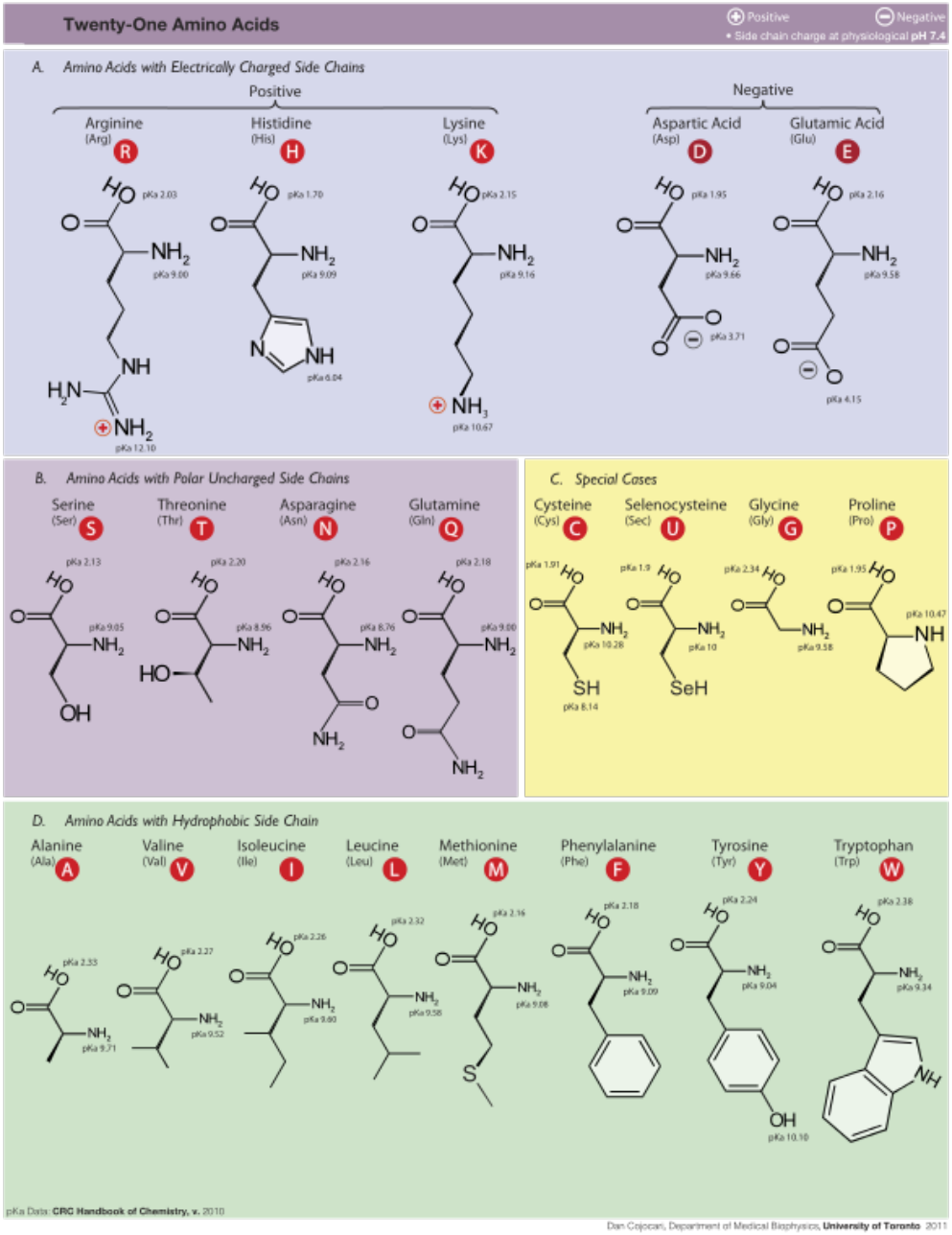


Figure 1.2: The 21 different types of residues. Fig. reproduced with permission from Wikipedia.

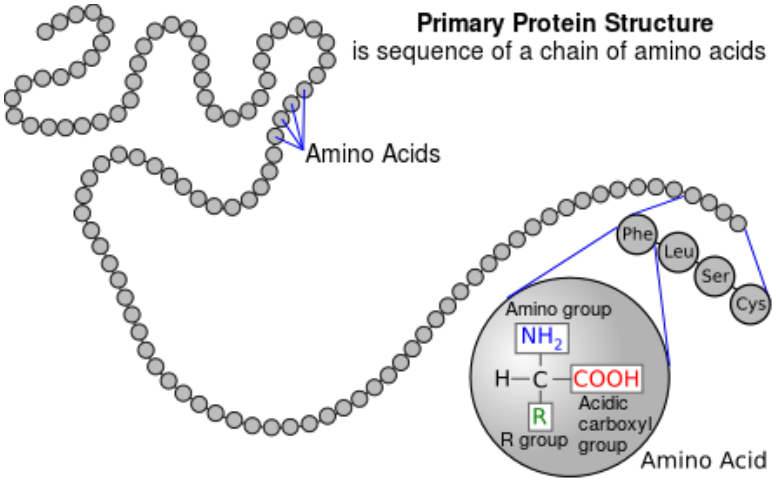


Figure 1.3: Different amino-acids are linked in a chain by covalent peptide bonds, and form the primary structure of a protein. Fig. reproduced with permission from Wikipedia.

Interaction	Energy scale	
	(kcal/mole)	k_BT
covalent bonds	50-150	80-250
electrostatic	10-20	20-35
ionic bonds	4-7	7-12
hydrogen bonds	5	8
Van der Waals	0.2-0.5	0.3-1

Table 1.1: Different interactions relevant for proteins and their energy scales. At room temperature T=300 K $k_BT \approx 0.6$ kcal/mole. Data taken from Ref. [3]

1.1.1 Interactions in proteins

Atoms in the primary structure interact in complicated ways between them and with the atoms of the solvent (which usually is water). Most of these interactions are quantum in nature, but it is common to define and classify them according to empirical criteria (Tab. 1.1).

Covalent bonds are the strongest interactions between atoms in proteins, such that it is impossible to break them at room temperature and at physiological pH, although in the cell, special enzymes can cleave them. The peptide bond linking different residues in the primary structures is of covalent type.

Atoms are usually neutral, unless they are in an ionic state, since the elementary charges are distributed symmetrically. This even distribution can get asymmetric when two or more atoms form a chemical bond. Partial charges are fractional values of the elementary charge that are used to model these local excesses or lacks in the charge distribution. Electrostatic interactions exist between partial charges and dipoles formed by them. Moreover, some residues have a net charge at physiological temperature and pH (Fig. 1.2). Electrostatic forces are calculated according to the usual potential scaling with r^{-1}, which makes them the only long range interactions in proteins. Water effectively screens electrostatic interactions, whose potential energy is scaled down by a Debye factor $\exp(-r/r_0)$. The ionic bond is a particular case of electrostatic interaction, since it involves the attraction between oppositely charged ions.

Hydrogen bonds (h-bonds) arises when two electronegative atoms share an hydrogen, and its nature is yet not completely understood and object of active research [6]. These bonds are formed within the molecule, within the solvent and between the two. They are highly directional and of utmost importance in shaping the native structure of proteins.

Van der Waals forces are weak attractive and repulsive interactions due to transient inhomogeneities in the atomic charge distribution, both in polar and apolar atoms. They are provoked by different quantum effects, and the kind of interaction depends on the distance separating two atoms. When two atoms are very close, almost in contact, the Pauli exclusion principle prevents the two electronic clouds to overlap. This exclusion manifests as an effective core repulsion. At an intermediate distance, Van der Waals forces are attractive. In fact, the symmetrical charge distribution around an atom is dynamical, and whenever transient inhomogeneities arise, instantaneous

dipoles emerge. One of these dipoles can induce another dipole in a close atom, opening the door to a subsequent dipole-dipole attractive interaction. This effect is known as London dispersion force. At further values of the distance separating two atoms, Van der Waals interactions vanish to zero. The overall behavior is usually approximated by a distance dependent Lennard-Jones potential (see 1.3.2).

To conclude the classification, proteins' dynamics is also affected by entropic effective interactions, which arises when single residues and protein are surrounded by water-like (polar) solvents. Residues are experimentally classified in hydrophobic or hydrophilic depending on whether they tend to attract or not with each other once put in water. This can be qualitatively understood considering the different propensity of different amino-acids to form hydrogen-bonds (h-bonds) with the water molecules. In the bulk of water transient h-bonds with lifetimes on the order of $\sim$ ps form a dynamical network, which is broken when a hydrophobic residue is put in water. H-bonds between water molecules and the residue are less stabilizing than those within water. It is energetically more advantageous for water molecules to saturate all the possible h-bonds with other water molecules, thus forming a rigid network, a sort of "cage" surrounding the hydrophobic residue, known as solvation shell. The frozen configuration of this shell correspond to a much lower entropy, lower than the stability gain granted by the h-bonds. Thus overall this rigid conformation has a higher free energy than the bulk, making it less probable, and the difference in free energy is found to be proportional to the surface of the residue exposed to the solvent. When two hydrophobic residues are put in water, the most probable configuration of the system is the one that minimizes the exposed surface to the solvent. This manifests as an effective attractive interaction between the two residues.

1.1.2 Structures

The native structure can be determined either by X-ray crystallography techniques, if the specific protein actually crystallizes, or by Nuclear Magnetic Resonance (NMR), which measures the coupling between protons of a protein in solution. Experimentally determined protein structures are then deposited and made freely available to anyone on the Protein Data Bank (PDB)[2], which is a world-wide open-access collaboration. The number of deposited struc-

[2] http://www.rcsb.org/pdb/home/home.do

tures has increased exponentially, from the first 13 in 1976 to the 95,644 available while these pages are being written, November 2013.

Protein structures display local regular patterns of h-bonds, which form the so-called secondary structure. The most frequent examples of secondary structures are the α-helix and β-sheet.

In water all the hydrophobic residues tend to attract among each other and as an effect bury themselves in the bulk of the protein. Once they are not longer exposed to the solvent, many h-bonds in the backbone atoms of the chain previously established with water molecules are not anymore saturated. They thus tend to form new h-bonds with other atoms in the backbone. Pauling demonstrated that α-helices and β-sheets are the geometrical patterns which maximize the number of formed h-bonds, hence maximize the stability of the resulting structure. The hydrophobic interaction is thus responsible for a compact and collapsed overall configuration, while h-bonds, being highly directional, yield regular local specific structures [7].

Secondary structures are packed in different ways, or topologies, and the overall resulting three dimensional conformation of the protein is called the tertiary structure.

The relative displacement of different tertiary structures in big and complex proteins or assemblies is know as the quaternary structure.

1.2 The protein folding problem

The so called *central dogma of molecular biology* states that the information flows from DNA to RNA and to proteins[3], which are the molecules entitled to put that information in action [9]. We talk of *transcription* when information is transferred from DNA to a message RNA filament (mRNA). We have *translation* when the information encoded in the mRNA is used to produce a protein. In order to carry out this translation, RNA has to find its way to the ribosome, that is the cellular factory assigned to the production of proteins.

Ribosomes "read" the mRNA and assembles the corresponding amino-acid

[3]The word *dogma* is used in this case in an erroneous way, since this flow has been extensively investigated and proved. Indeed, as declared by Crick himself in his autobiography [8], he used that word just because it conveyed a catchy sentence.

in a linear chain, which has to fold to the native structure in order for the protein to be biologically active. Folding can happen while the chain is being synthesized by the ribosome (co-translational folding [10, 11]), or after (post-translational). In both cases one could think that the linear amino-acid chain is shaped in its native configuration by a plethora of complicated cellular processes and machineries. Investigating folding *in vivo,* in the overcrowded environment of a living cell, is out of reach even using current experimental techniques. Hence, how can we address the study of the folding of proteins with this severe limitation?

1.2.1 Proteins are self-assembling systems

We can treat a simplified version of the problem, *in vitro*, thanks to a series of very famous experiments carried out by Anfinsen in the '60s [12]. In a rather simplified version, he unfolded a sample of the 124-residue long bovine pancreatic ribonuclease enzyme in a test tube. The complete loss of secondary and tertiary structures can be obtained raising temperature or, as Anfinsen did, using urea, that is known to be a very effective denaturant. The sample of denatured proteins showed no sign of enzymatic activity. Once urea is removed, hence the environmental conditions are restored to the native ones, the protein *spontaneously* refolds to its native configuration, displaying again its enzymatic activity.

The usual interpretation of these experiments leads to what is known as the Anfinsen's dogma: all the information that a protein needs to attain its native configuration under native environmental conditions is contained exclusively in its primary structure, i.e., the amino acid sequence which composes the chain. Small globular proteins are able to spontaneously fold once they are in the right environmental conditions.

This postulate, which is by now verified by many experimental evidences, is of utmost importance for any investigation of the protein folding problem. Indeed, by invoking it we can avoid to consider the problem in a living cell and put our efforts to study it in a test tube, where it appears more simple and all the experimental conditions can be under control. It was verified that there is a series of molecular processes modifying the amino acid chain post-translationally in the cell [4]. Moreover, the folding of real proteins is facilitated by the presence of molecular chaperones, like the GroEL-GroES complex, whose shape recalls that of a barrel. As this particular shape sug-

gests, the function of the GroEL-GroES complex is to shelter a folding chain from the disturbances of the overcrowded cellular cytoplasm [13]. It is thus believed that the function of chaperonines is to enhance the folding rate, but they presence is not essential. Unfortunately we still know very little about how the *real* protein folding happens in a cell [10, 14].

Anfinsen's dogma allows us to forget (at least temporary) all these difficulties, and to focus on the simplified version of the protein folding problem. We are thus interested in understanding the spontaneous folding and unfolding of a protein or a sample of diluted proteins in water. Following [4], we quote Fersht [15]:

> We can assume that what we learn about the mechanism of folding of small, fast-folding proteins *in vitro* will apply to their folding *in vivo* and, to a large extent, to the folding of individual domains in larger proteins.

1.2.2 Investigating the mechanism

Having restricted the problem under study, we want now to better specify what exactly we aim to understand. Different authors partition the problem in different ways [4, 16, 17], or even say that the very definition of protein folding problem has little meaning since protein science has become an entire flourishing active area of research [17]. We will say, following Karplus [16], that the problem can be dissected in two main parts:

1. Understanding how the information about the three-dimensional native structure is fully encoded in the primary structure. In other words, the ultimate goal is to develop algorithms that predict the native structure once the amino acid sequence has been given as an input. Great successes have been accomplished in this direction, but it is widely believed that this part of the problem is not suited for a physical approach [4, 16, 17].

2. Understanding which is the mechanism used by a protein to fold. On one hand, this involves the characterization of the folding pathway, i.e., the description of the sequence of events connecting the unfolded to the native conformation. On the other hand, this entails also the understanding of this process dynamics, of all the relevant interactions and, as an ultimate goal, a global comprehension of the general rules.

The second part of the problem is mostly investigated by means of physical methods, and we will limit our attention on it only.

We will focus in this work exclusively on the folding of small, globular, single domain proteins. These molecules are by far the most characterized by experimental and computational investigations [18], since they represent the simplest systems showing all the typical features of folding. Following Finkelstein, we can say that globular one-domain proteins are the simplest biological self-assembling objects [2].

The universe of real proteins is far richer and more complicated [17]. Most of existing proteins are actually composed of multi-domains, each of which is able to independently fold. Then there are membrane proteins, which still present many difficulties upon characterization [19, 20]. Moreover, during the last years we have understood that nearly half of all eukaryotic proteins present a large portion of their chain that never displays a folded structure [21]. It is generally believed that this lack of structure is related with the need of being versatile in carrying on promiscuous biological functions, but nonetheless we know very little of the behavior of this so called *intrinsically disordered proteins* (IDP) [21].

1.2.3 Folding is described by a two-state kinetics

Anfisen's experiments have shown that studying protein folding outside the cell is a well posed problem. We will now briefly review some results about the kinetics of the process, i.e., how fast a protein (un)folds [2]. Usually these experiments are carried out by having a sample of folded proteins in a test tube, then unfolding them by adding chemical denaturants (urea or guanidinium chloride (GuHCl)) or raising temperature, restoring the native conditions and measuring how the refolding proceeds by means of several experimental techniques. It is worth noting that chemical or thermal denatured states are presumably rather different from the initial unfolded state in a cell.

For most of the simple, small, globular proteins which we will focus on, the outcome of experiments is rather well described by a two state kinetics [22, 23]. In this scheme only two states are populated, the native one, $\mathcal{N}$, and the denatured one $\mathcal{U}$, and a protein switches from one state to the other

with a given rate:

$$\mathcal{U} \underset{k_f}{\overset{k_u}{\rightleftharpoons}} \mathcal{N}$$

A notation that is widely used is to write k_f for the rate of transitions from $\mathcal{U}$ to $\mathcal{N}$, and k_u for the unfolding. The equations defining the rates are:

$$\begin{aligned} \frac{dN}{dt} &= -k_u N(t) + k_f U(t) \\ \frac{dU}{dt} &= k_u N(t) - k_f U(t) \end{aligned} \tag{1.1}$$

where $N(t)$ and $U(t)$ are the fractions of proteins that populate respectively the native and the unfolded state at time t, subject to the normalization condition $N(t) + U(t) = 1$. After a transient time the system sets to equilibrium, and the time derivatives in eq. (1.1) vanish. If we define the equilibrium constant

$$K_{\text{eq}} \equiv \frac{N_{\text{eq}}}{U_{\text{eq}}} = \frac{k_f}{k_u}$$

and we take $U(0) = 1$ (i.e., we are analyzing a refolding experiment), then the system (1.1) admits solution

$$N(t) = \frac{K_{\text{eq}}}{1 + K_{\text{eq}}} \left(1 - e^{-(k_f + k_u)t}\right) . \tag{1.2}$$

A kinetic characterized by eq. (1.1) and (1.2) is known as single exponential, and the quantity $k_f + k_u$ is known as the relaxation rate.

Since for a large ensemble of proteins the fractions $N(t)$ and $U(t)$ are proportional to the probabilities to populate the two states, if we suppose to treat the system as in the canonical ensemble, we can express the equilibrium constant as connected to the *stability*

$$\Delta G \equiv G_N - G_U = -k_B T \ln K_{\text{eq}}$$

that is the difference of free energy between the native and the unfolded state. Regarding proteins at ambient temperature and pressure, the term $p\Delta V$ that makes the difference between Helmholtz's and Gibb's free energies is negligible [24], and we will make no distinction from now on. It is found experimentally that proteins are marginally thermostable, since values of ΔG usually range from -15 to -5 kcal/mol, far from the maximum stability they

could attain [25]. The fractions $N(t)$ and $U(t)$ are also proportional to the concentrations of the two species in a volume V, which are usually written as $[N]$ and $[U]$.

The relaxation rate $k_f + k_u$ is determined measuring how it varies with chemical denaturants, representing it in what is known as the chevron plot [22]. Under native conditions $k_u \sim 0$[4], and the dependence of the folding rate on temperature is well reproduced by an Arrhenius relation

$$k_f = k_0 e^{-\frac{\Delta G^\ddagger}{k_B T}}$$

that describes a reaction exponentially hindered by an activation free energy $\Delta G^\ddagger$ [2, 26]. The inverse of the folding rate is the mean folding time, which is in the range of ms to s. Notable exceptions are represented by the ultrafast folders [27], folding in a time on the μs scale, and by proteins that are very big or characterized by a complex topology (e.g., knotted proteins), which fold in minutes or tens of minutes [23].

Plaxco *et al.* found an interesting correlation between the complexity of the topology and the folding time in a protein [28, 29]. Two residues i and j are are said to be in contact in the native structure if their distance is below a given threshold, which is typically of the order of 7 Å . One can calculate the contact order of a native structure

$$CO = \frac{1}{LN} \sum^{N} \Delta S_{ij}$$

where L is the total number of residues in the protein, N the number of native contacts, and ΔS_{ij} the distance in sequence between the residues i and j that are in a contact. Lower values of contact order correspond to proteins where native contacts have usually a local topology, i.e., the involved residues are close. Such proteins are usually molecules rich of α-helix secondary structures. On the contrary, high values of the contact order corresponds to non-local topologies, mostly determined by the presence of β structures. The correlation found by Plaxco *et al.* [28–30] and shown in Fig. 1.4 implies that

$$k_f \sim e^{-CO} .$$

[4] At room temperature, considering a stability of $\sim$ 10 kcal/mol, $K_{\text{eq}} = \exp(-\Delta G/RT) \sim 10^7$, which is hence the ratio between the number of folded proteins over the number of unfolded ones.

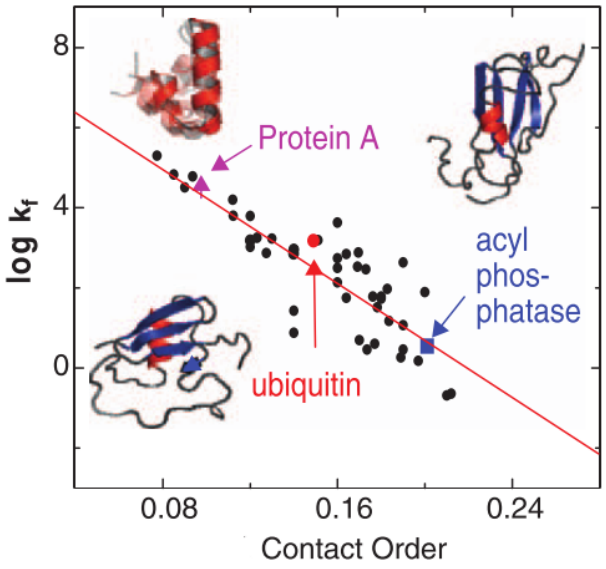


Figure 1.4: Folding rate *vs.* CO. A clear correlation exist spanning over several orders of magnitude. Fig. reproduced with permission from Ref. [18]

1.2.4 Why is folding so fast?

The fast and apparently effortless way proteins fold is astonishing if one considers how complex its dynamics is. Cyrus Levinthal wondered how such a simple and fast behavior can emerge from a complicated interplay of a large amount of interatomic interactions. Since then, in his honor this conundrum has been called the Levinthal's paradox [31].

Levinthal noted that since at physiological conditions a protein has to mostly populate its native state, it is natural to think that this state is the one characterized by the lowest free energy. But a protein can be found in an astronomical large number of configurations. For instance, we can consider that an average protein is 100-residues long, and that each amino acid displays from 2 to 10 degrees of freedom, depending on whether we consider only the backbone or also the side-chain. The most conservative estimate of the number of possible conformations is $2^{100} \sim 10^{30}$ [31]. The fastest rate by which a protein can change conformation is estimated by calculating the frequency associated with thermal energy, i.e.,

$$\nu = \frac{k_B T}{h} \sim 10\,\mathrm{ps}^{-1} \tag{1.3}$$

where h is the Planck constant. But even with this rate, the exploration of the huge conformational space would take $\sim 10^{19}$ s, which is 20 times the age of the Universe.

There are two wrong hypotheses behind this famous estimate [4]. First, the protein can display each conformation with the same probability, therefore they all have the same energy. Second, it can hop from one conformation to any other randomly, and this is for sure not true since we expect the evolution across the conformational space to be continuous. Basically, the Levinthal's paradox presumes a random unbiased sampling of a astronomical large number of conformations all showing the same energy.

As noted in [4], this paradox has to be meant as a *reductio ad absurdum* proof, which clearly shows that the quest for the native state cannot be completely random. As was shown in [32, 33] by using a kinetic argument, a non democratic search can take a biological relevant mean time of arrival. In particular, by exploring the conformational space Ω, individual residues will happen to find their native configuration. If this does not change the probability to switch to another conformation, then the Levinthal's paradox holds unchanged. But the paradox can be avoided if whenever a residue displays a native configuration it tends to stay there. It is enough to impose an energy penalty of few k_BT in leaving a native configuration for a "wrong" one in order to exceedingly reduce the folding time to few seconds. The latter result is a precious hint on how the exploration of Ω proceeds.

1.2.5 The folding thermodynamics

It seems at this point opportune to illustrate in a more rigorous way the thermodynamics of the protein folding. This will also be an occasion to introduce notations and terminology widely used in the literature. We will mainly follow [2, 4, 24].

We shall set the framework of a microscopic description of the equilibrium thermodynamics of a protein. Let us consider a protein made of N atoms in water at fixed temperature T and pressure P and constant number of water molecules N_w, i.e., we will work in the statistical canonical ensemble. The results we are going to carry out will be valid also in the scenario of a diluted solution of non interacting proteins. We will follow the convention to use lower-case letter for the molecules atoms, and upper-case for the water ones.

If the system is ergodic, standard equilibrium thermodynamics suggests that the native state $\mathcal{N}$ is the one with the global lowest free energy. Assuming that the system can be correctly described at a classical level, the whole dynamics is encoded in the Hamiltonian

$$H\left(\mathbf{x},\mathbf{X},\boldsymbol{\pi},\boldsymbol{\Pi}\right)=\sum_{i}\frac{\pi_{i}^{2}}{2m_{i}}+\sum_{j}\frac{\Pi_{j}^{2}}{2m_{j}}+V\left(\mathbf{x},\mathbf{X}\right)$$

with $i=1,...,N$ and $j=1,...,N_w$, where m_k is the atomic mass respectively of the protein (when $k=i$) and water atoms (when $k=j$), and $V\left(\mathbf{x},\mathbf{X}\right)$ denotes the inter-atomic potential energy.

At equilibrium any thermodynamic quantity can be derived from the knowledge of the partition function

$$Z=\frac{1}{h^{N+N_w}N_w!}\int_{\Gamma\times\Gamma_w}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{X}\mathrm{d}\boldsymbol{\pi}\mathrm{d}\boldsymbol{\Pi}e^{-\beta H(\mathbf{x},\mathbf{X},\boldsymbol{\pi},\boldsymbol{\Pi})} \tag{1.4}$$

where as usual $\beta\equiv(k_BT)^{-1}$, h is the Planck constant, Γ and Γ_w are the phase space of the molecule and the solvent, respectively. The factor in front of the integral represents the elementary cell in the phase space, and takes into account for the indistinguishably of water molecules.

At this point it is natural to investigate the dynamics of the molecule by considering an average behavior of the surrounding water. Indeed, it would be uninteresting and useless to derive an equation which explicitly takes into account any possible conformation of all the water molecules. For the average to work, we need for the typical water's equilibration time to be much faster than the molecule's one. Since water sets to equilibrium on the ps scale, this seems a safe assumption.

The standard way to consider an effective dynamics is to average out the water degrees of freedom. It is straightforward for the momenta, since they reduce to a multiple Gaussian integral, returning a multiplicative factor in front of Eq. (1.4) that depends on temperature and mass of the water molecules. In the canonical ensemble, thus at fixed temperature, the multiplicative factor acts as a normalization constant and can be ignored.

Averaging out water's conformational degrees of freedom is not trivial, since they are coupled to the molecule's one by the potential energy. Nevertheless, this can be done at least formally introducing the potential of mean

force:

$$W(\mathbf{x};T) = -k_B T \ln\left(\int_{\Omega_w} \mathrm{d}\mathbf{X} e^{-\beta V(\mathbf{x},\mathbf{X})}\right) \tag{1.5}$$

where the integration is performed over all water conformations Ω_w. We shall stress the fact that this new quantity, which is known as effective energy, is temperature dependent. To understand this we can think for instance that it encodes the dynamics of water, its entropy, and consequently the hydrophobic interaction.

Having introduced the effective energy, we can build an effective Hamiltonian

$$H_{\text{eff}}(\mathbf{x},\mathbf{X},\boldsymbol{\pi},\boldsymbol{\Pi}) = \sum_i \frac{\pi_i^2}{2m_i} + W(\mathbf{x};T)$$

and an effective partition function. In the latter we can integrate out as above the momenta of the molecule's degrees of freedom, obtaining

$$Z_{\text{eff}}(T) = \int_{\Omega} \mathrm{d}\mathbf{x} e^{-\beta W(\mathbf{x};T)}\,.$$

where Ω is the configuration space of the molecule. In the following we will drop the temperature dependence and the effective subscript for the sake of a lighter notation. With the new partition function, the probability density function (PDF) of the system reads

$$p(\mathbf{x}) = \frac{e^{-\beta W(\mathbf{x})}}{Z}\,.$$

We want to match the microscopic description introduced so far with the one based on the state concept. According to the outcome of kinetics experiments, the system can be found either in the native state ($\mathcal{N}$) or in the unfolded state ($\mathcal{U}$).[5] Since $\mathcal{N} \cap \mathcal{U} = \emptyset$, we can consider thermodynamic quantities to be restricted to these states, starting from the partition function

$$Z_i = \int_{\Omega_i} \mathrm{d}\mathbf{x} e^{-\beta W(\mathbf{x})}\,.$$

Any state $\Omega_i \subset \Omega$ has a probability to occur given by

$$P_i = \frac{Z_i}{Z}$$

[5]This is approximately true only in systems displaying a two state kinetics with no long lived intermediates.

and and a free energy

$$G_i = -k_B T \ln Z_i \,.$$

We can now recover the stability that was introduced discussing about the kinetics of protein folding and several valid equalities:

$$\Delta G \equiv G_N - G_U = -k_B T \ln \frac{Z_N}{Z_U} = -k_B T \ln \frac{P_N}{P_U} = -k_B T \ln \frac{[N]}{[U]} = -k_B T \ln K_{\text{eq}} \,.$$

Since the conditional PDF of a configuration $\mathbf{x}$ in a state Ω_i is

$$p_i(\mathbf{x}) \equiv p(\mathbf{x}|\mathbf{x} \in \Omega_i) = \frac{e^{-\beta W(\mathbf{x})}}{P_i}$$

we can calculate the internal energy of a state

$$U_i \equiv \langle W_i \rangle = \int_{\Omega_i} \mathrm{d}\mathbf{x} W(\mathbf{x}) \, p_i(\mathbf{x})$$

and its entropy

$$S_i = -k_B \int_{\Omega_i} \mathrm{d}\mathbf{x} p_i(\mathbf{x}) \ln p_i(\mathbf{x}) \ .$$

From the latter expression it is possible to recover the basic definition of free energy, and therefore write any free energy difference as

$$\Delta G = \Delta U - T \Delta S \,.$$

The mean force acting on a protein that we have written in equation (1.5) can be thought as a hypersurface W defined as

$$\begin{aligned} W: \ & \Omega \to \mathbb{R} \\ & \mathbf{x} \mapsto u = W(\mathbf{x}; T) \end{aligned}$$

an *energy landscape* upon which the protein "navigates" searching for the native state [34]. It is worth to remember that "energy" here has to be meant as potential of mean force or effective energy, since it represents the energy of a configuration $\mathbf{x}$ averaged over all the possible configurations of all the water molecules. The energy landscape is of course not a novel concept *per se*, but we will use it in the following as a useful pictorial way to grasp some understanding on how it should look like regarding the folding problem.

As we described, Levinthal's paradox is based on a random unbiased sampling of the protein's huge conformational space Ω. This searching strategy corresponds to an enormous perfectly flat landscape which displays a deep but extremely narrow well in correspondence of the native state. Wolynes described this scenario using the felicitous picture of a "drunk playing golf" [34], and such an energy landscape is now known as a "golf-course" [35].

1.2.6 An energy bias towards the native state

Levinthal himself proposed a solution to his paradox, by introducing the concept of folding pathway[6]. He speculated that the folded and the unfolded conformations of a protein are connected by a specific sequence of intermediate configurations [36]. This was also motivated by Levinthal's belief that a rapid folding and reaching the global free energy minimum were incompatible. He thus proposed folding to be *kinetically* driven, i.e., the native state is the configuration reachable in the lowest amount of time. As a consequence, initiating the folding from two different denatured configurations implies reaching two different native structures. The opposite view, in which the protein finds its most stable state, is known as *thermodynamically* driven folding. By taking advantage of a pictorial view of the energy landscape, this scenario corresponds to a narrow canyon connecting the unfolded configuration to the folded one (Fig. 1.5 panel (b)). In this scenario the rate-limiting step of the reaction, called Transition State (TS), is a high energy conformation. Levinthal's pathways have been ruled out, since experiments have never found such a deterministic sequence of intermediate states.

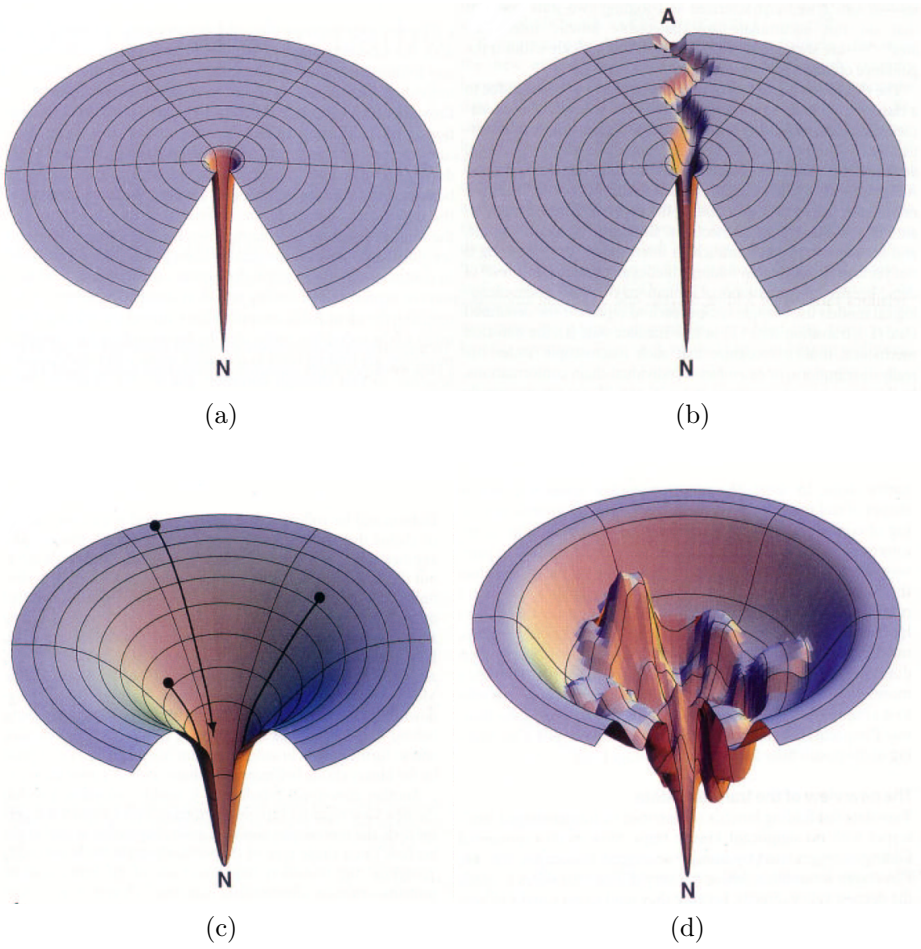
The random wandering on the golf-course landscape (Fig. 1.5 panel (b)) that inspired Levinthal's paradox can be paraphrased according to Finkelstein [2] as follows: before forming any energetic stabilizing contact, the amino acid chain has to attain a huge conformational entropy reduction. This latter yields to an enormous free energy barrier of entropic nature that makes the transition practically impossible to happen. But the work of Zwanzig *et al.* shows that a way to reduce this barrier is to preserve any native interaction that is randomly formed [32, 33] during the folding. In this way, the conformational entropy reduction can be simultaneously and continuously balanced by the formation of stabilizing energetic interactions. We now de-

[6]Actually Levinthal's paper where the idea of pathway had been introduced [36]was published one year before he gave the talk in which he exposed his famous paradox [31].

fine a *native contact* as a couple of residues that in the native state are closer than a given threshold. They interact attractively stabilizing the native state. *Non-native contacts* are transient interactions happening during the folding between residues which are not close in the native state. These contacts can be either repulsive or attractive. The simple model proposed by Zwanzig *et al.* suggests that a way to solve Levinthal's paradox is to consider an *energetic bias towards the native state*.

The same conclusions were drawn from a different perspective by Bryngelson, Wolynes and Onuchic [34, 35, 37–40]. They were inspired by a statistical mechanics approach to spin glasses [41, 42]. These systems display a landscape characterized by the presence of a huge number of similarly deep minima, with no global energy minimum dominating over all. During their evolution these systems remain trapped for an infinite time in these minima, being like a frozen liquid. This happens because there is not a single configuration in which most of the interactions between the components are all simultaneously stabilizing in a cooperative way. Such an inability to optimize the interplay between the different interactions is called *frustration* [43]. Typical random heteropolymers display a frustrated energy landscape, with no hierarchical organization of the minima [44, 45]. Indeed, they are unable to show the characteristics of folding, i.e., its reliability and efficiency. Thus, Bryngelson, Wolynes and Onuchic proposed that proteins are very particular heteropolymers displaying *minimal frustration*. They postulated that folding proceeds by forming native contacts, which in average are always stabilizing; and that non-native interactions, that are considered as a form of frustration, are distributed randomly along the folding. The native energy bias was already anticipated in the work of Gō and Taketomy [46].

In order to give a visual representation of this solution to Levinthal's paradox, Bryngelson, Wolynes and Onuchic have introduced the funneled energy landscape (1.5 panel (c)). To be more precise, this picture represents the *free energy surface*. Indeed, if the vertical axis corresponds to the effective energy, then the horizontal axis represents the conformational entropy . Descending from the top of the funnel, which represents the denatured state (DS) and displays the highest number of conformations, energy and entropy decrease. In other words, going down in energy, there are less and less configurations having that energy. In this sense the famous funnel shape is due to the reduction of entropy and not of energy [48]. We already know that the stability $\Delta G \simeq 10$ kcal/mol, and an estimate of $T\Delta S$ returns ~ 100



(a) (b)

(c) (d)

Figure 1.5: Different types of pictorial representations of the folding energy landscape. Clockwise starting from the upper left panel: unbiased random search (golf-course landscape); Levinthal's pathway; perfectly smooth folding funnel; rugged landscape due to important non-native interactions. Fig. reproduced with permission from Ref. [47], which displays a more rich taxonomy of possible energy surfaces.

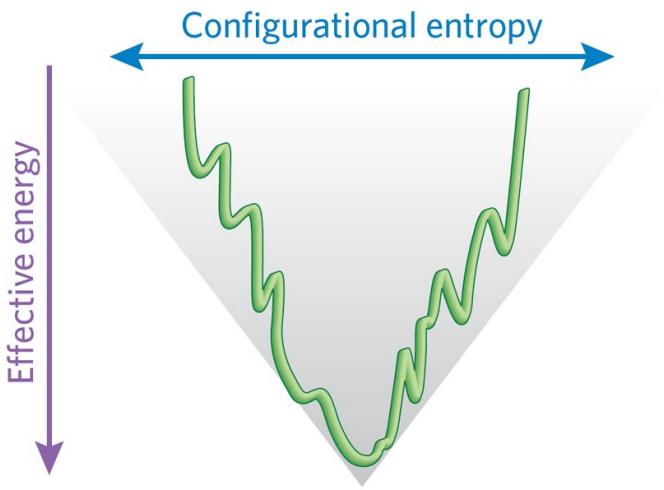


Figure 1.6: Pictorial representation of the free energy landscape of protein folding. The effective energy is represented on the vertical axis. It corresponds to any possible interaction (chemical bonds and angle energy, dihedral angle rotation, inter-atomic Van der Waals, electrostatic, etc.) and entropic term (solvation free energy, solvent entropy, etc.), except the conformational entropy of the protein, which is encoded in the width of the funnel. It is implicitly assumed that there is a good reaction coordinate Q describing the folding of the protein. Energy and entropy both diminish smoothly going towards the native state, which lies on the bottom of the plot. Any difference in the slopes of $U(Q)$ and $S(Q)$ originates a free energy activation barrier according to the elementary relation $G(Q) = U(Q) - TS(Q)$. As noted by several authors, there has been some confusion about this picture. The effective energy U is obtained averaging out any solvent degree of freedom, hence it is technically a free energy itself and is called that way by some authors. We have preferred to follow [4] in order to remove any possible ambiguity by calling free energy only the quantity G. Furthermore, Karplus noted that sometimes it is believed that a protein folds quickly because of the funnel shape, which is due to the reduction of conformational entropy [48]. In fact, this reduction hinders the folding process. Proteins fold fast because there is an energy bias to go towards the native state. Fig. reproduced with permission from Ref. [48].

kcal/mol. Hence, the effective energy decrease has to be $\Delta U \sim 100$ kcal/mol, and the stability is the result of a fine cancellation [24].

Approaching the problem with the instruments of statistical mechanics has introduced a new stochastic view of the process. Pathways are no more necessary, since any possible route on the landscape leads to the native state. Folding is no more a sequential process but an intrinsically parallel one. DS and transition state (TS) are replaced now with ensembles of possibly very diverse configurations. This new approach to the folding problem is known as the "new" view, compared to the old one based on the pathway concepts. Ref. [16] outlines a historical perspective on how the transition between these views happened, investigating the various original contributions. A detailed and illuminating review of the differences between the two scenarios can be found in Ref. [47].

To summarize, *what we have understood in the last decades is that protein folding can happen spontaneously and efficiently because there is an energetic favorable bias towards the native state. Together with the observation that the number of possible conformations decreases as the energy goes down, this lesson can be summarized in a pictorial way in the well-known funneled free energy landscape plot 1.6. This view is now widely accepted as the conceptual framework of the protein folding process.*

1.2.7 Rough or smooth landscapes

Panel (c) of Fig. 1.5 shows a smooth landscape, meaning that random frustrations, that would correspond to a rough surface, are negligible compared to the energetic bias towards the native state. This is an extremely idealized scenario. Even if natural proteins are minimally frustrated, there could still be a residual presence of non-native interactions. In the funnel picture, we can imagine them as little bumps (i.e., repulsive non-native interactions) or little valleys (i.e., attractive interactions), which could act as kinetic traps hindering the folding process. A rather rough landscape due to frustration would appear then more like panel (d) of Fig. 1.5. A lower and more uniform degree of frustration would instead cause the system to feel a sort of friction when evolving on the landscape.

Which is the amount of frustration and which is its effect on folding in real protein is still unclear and under active investigation*, and will be one of*

the main questions we are going to investigate in this thesis.

1.2.8 There are many diverse bottlenecks for folding

By shortly reviewing the two kinetics of two-state proteins, we have seen that the dependence of the rate on temperature is well described by an Arrhenius law, thus by a reaction that has to overcome a free energy barrier $\Delta G^{\ddagger}$. The question that naturally arises is where this barrier is hidden in the funneled energy landscape. Indeed, while in the old view of Levinthal's pathways the TS was a configuration of particular high energy, in this new view it is an ensemble of configurations, all acting in a possible different way as the rate-limiting step to folding. In a perfectly smooth landscape, if there is an activation free energy barrier, then this is of entropic nature. Indeed, it would be caused by a different pace in the decrease of energy and entropy, with the former diminishing not rapidly enough to balance the latter [2, 24]. This can be seen also in a simplified model that is analytically tractable but displays all the characteristics of the folding on a smooth landscape [49]. In presence of significant frustration, the rate-limiting step could be due also to high-energy conformations. In a realistic case, the bottleneck of folding is formed presumably by different kinds of barriers, namely entropic, energetic, and topological [47] (i.e., to attain the right conformation the chain should cross itself).

1.2.9 The two views are not incompatible

The "old" and the "new" views of folding appear to be limit cases of a richer spectrum of possible energy landscapes [4, 50]. In order to explore this idea we want to first eliminate the ambiguity that surrounds the word "pathway". Therefore, we will solely use it to specify any description of the folding mechanism by means of a set of collective variables. Any microscopic realization of the folding in atomistic detail will be simply called "trajectory". Thus, a pathway is a collection of different microscopic trajectories that show a common pattern at a coarser level of description. This pattern could be for instance in the order of formation of the secondary structures, of the hydrophobic nucleus or of the number of native contacts. A pathway is an average behavior to some extent insensitive to what happens in the microscopic trajectories which define it. We will use the words "mechanism" and

"reaction channel" as synonyms for pathway.

In panel (c) of Fig. 1.5 the funnel is perfectly smooth, non-native interactions have no role, and folding is completely heterogeneous, not just at a microscopic level but also concerning the folding pathways. On the other side, panel (b) shows a landscape characterized by an utmost severe frustration, leaving just one open pathway, and folding is perfectly homogeneous. However, it is possible that in real situations a finite number of pathways emerges from all the microscopic realizations, due for example to the specific topology of the protein or the concerted action of non-native interactions. These interactions would shape the topography of the funnel carving deep channels traveled by most of the microscopic trajectories. The landscape would be like in panel (d) of Fig. 1.5, an intermediate version of panel (b) and (c).

Whether such a pathway description has any validation with reality or not has been extensively investigated during the last years. Unfortunately, experimental techniques are mostly based on ensembles analysis, and recovering any information on the actual mechanism (even knowing whether there is just one or many) is not straightforward at all. Hence, since the seminal study by Lazaridis and Karplus [50], this problem has been mainly tackled by means of computational approaches, which we will partially review in the next section.

The main goal of this thesis is to present and discuss a computational method able to find and characterize the folding pathways of globular proteins by means of a numerical simulation of microscopic folding trajectories.

1.2.10 The origin of the funneled landscape

Nowadays, almost everyone accepts that the free energy landscape of folding appears to be funneled. However, we have not said *why* it happens to be that way. In fact, random heteropolymers do not share this peculiarity, displaying a very rough energy landscape with a lack of a clear hierarchical organization of the energy minima [44, 45]. It has been postulated that proteins do have a smooth funneled landscape because they are very particular heteropolymers selected by evolution during billions of years. A pressure for a rapid and reliable folding would have selected only those amino acid sequences displaying the required characteristics. Although it seems clear that real proteins and

their properties are the result of an evolutionary selection, it is still unclear whether there was a positive pressure for a funneled landscape or the latter is just an emergent feature of a more complicate process [23].

1.3 Investigating the protein dynamics on a computer

We will focus now our attention on how to deal with the determination of the protein folding mechanism. From a physical point of view, we are thus interested in solving the dynamics of the amino acid chain in a native-like environment. Since the system we want to study is usually composed of thousands of atoms all interacting in a non-trivial way, it is clear that our purpose is to understand the dynamics of a folding protein by means of numerical approaches on a computer.

First of all, we can consider several levels of spatial resolution possibly adopted to describe the molecule. In a coarse-grained (CG) representation each amino acid is described as a bead located in the position of the Cα-atom. Instead, all-atom (AA) models explicitly represent all the atoms of the protein and are those with the highest spatial resolution. In the following, we will always refer to AA resolution if not differently specified.

In lattice representations, each component of the system can be placed on the sites of a lattice on discrete positions. This drastic approximation makes calculations much faster. On the other side, off-lattice representations simulate the system as embedded in physical space, and are far more accurate but also more expensive.

Atoms and molecules are quantum objects in nature. Solving the dynamics of a protein would mean solving the Schrödinger equation for all the atoms and electrons interacting *via* the Coulomb potential. Such an approach is known as *ab initio,* and is in principle the most accurate way to investigate the time evolution of a molecule. Unfortunately, such an extreme accuracy comes with a huge computational cost, which makes practical quantum *ab initio* calculations feasible only for very small molecules, and absolutely prohibitive for any system of biological interest. Therefore, it is necessary to approximate the system, neglecting quantum effects and treating atoms and molecules classically.

To understand in a rough way whether this approximation is allowed or not, we can calculate the time and length scale associated to a characteristic thermal energy k_BT, that at normal conditions ($T = 25$ °C) is ≈ 0.6 kcal/mol. Indeed, for a system at thermal equilibrium a classical approximation works if

$$k_BT \gg h\nu$$
$$d \gg \lambda = \frac{h}{\sqrt{3mk_BT}}$$

where ν is the fastest rate of the system, d is the typical length scale, λ the De Broglie wavelength associated to thermal energy, h is the Planck constant, and m= 1 atomic mass. We have already seen that the time scale associated with thermal energy is $\sim 10\,\mathrm{ps}^{-1}$ (eq. (1.3)). Typical rates of motions in a molecules range from $100\,\mathrm{ps}^{-1}$ for the stretching of an H-O bonding, to $1\,\mathrm{ns}^{-1}$ for the rotation of dihedral angles. Indeed, our approximation will badly work in dealing with atomic bonds, which are purely quantum, whereas it will be a fairly good one regarding all the other time scales in the system. Since electrons relaxes on a time scale $\ll$1 ps, the Born-Oppenheimer approximation holds and we can avoid to explicitly describe electrons, and treat them collectively as a potential energy surface describing their ground state. On the other hand, λ yields the length scale at which quantum effects arise, and one has $\lambda \sim 10^{-10}\,m \sim$ atomic dimension. Since λ does not overlap over several atoms, we can approximate inter-atomic interactions with classical ones.

We have now understood that in order to describe the folding protein dynamics we need to solve Newton's equations numerically

$$\begin{aligned} \dot{\mathbf{v}}_i &= -\frac{1}{m_i}\nabla U\left(\mathbf{X}_i\right) \\ \dot{\mathbf{x}}_i &= \mathbf{v}_i\left(t\right) \end{aligned} \tag{1.6}$$

for any atom i in the system. This kind of approach is known as Molecular Dynamics (MD) simulation.

Solving the system of differential Eq.'s (1.6) with a time independent potential means to simulate the time evolution of a system where energy in conserved, and is thus equivalent to sample the micro-canonical ensemble. In order to sample the canonical ensemble two main strategies can be used. One can couple the system described by Eq.'s (1.6) to a thermostat and thus let the two to exchange energy. Accomplishing this task is not

trivial, and several algorithms have been proposed [51]. An alternative approach consists in solving a microscopic diffusion equation (the over-damped Langevin equation) instead of Eq. (1.6), which automatically samples the canonical ensemble. This second strategy will be extensively explored in the next chapter.

1.3.1 The solvent

The system we want to simulate is not only composed of the molecule, but also of the environment, that is the solvent where the protein is immersed. In most of the cases this solvent is water with the addition of ions Na^+ and Cl^- to account for the pH of the solution. Solvent has a fundamental role in determining the dynamics, and in particular concerning the folding of proteins. It is responsible for solvent-solvent and solvent-protein hydrogen bonds, electrostatic interactions, changes of dielectric properties, and, most importantly, for the hydrophobic effect, which is considered the main interaction responsible for the initial stages of folding. As a matter of fact, proteins fold only in water-like polar solvents.

There are two main different approaches in dealing with the simulation of the solvent, namely explicit and implicit water. Firstly, one can simulate explicitly each water molecule by means of different models [52] (e.g., SPC, TIP3P, SPC/E, TIP4P, etc.). This high level of detail is fundamental when water's granularity plays an important role in the specific investigated system. AA MD in explicit water is the most detailed description of the protein folding dynamics, and it comes also with the highest computational cost. Indeed, one has to add a number of water molecules that is usually at least ten times the number of atoms in the protein. Most of the CPU time is devoted to simulate the thermal motion of thousands water molecules in the bulk, which has mostly no direct consequences on the dynamics of the molecule. Moreover, by using simulations that adopt an explicit solvent model it is very difficult to get the solvation free energy, that is the free energy of transferring the system from vacuum to the solution [24].

To overcome these difficulties and speed-up simulations, one can treat water on an average level by going further in the mathematical treatment that was sketched in section 1.2.5, which yields an implicit solvent model. Empirical models have been proposed introducing an effective ΔG^{solv} term and accounting for the different effects of the solvent on the molecule [24, 53, 54].

The hydrogen-bonds network that exists between water and molecule's atoms cannot be simulated explicitly and has to be taken into account effectively. Moreover, the hydrophobic effect is implemented by means of an empirical relation, stating that ΔG^{solv} is proportional to the Solvent Accessible Surface Area (SASA). Water viscosity, due to the microscopic impact of water molecules on the protein, is not reproduced. On one hand this enhances the rate of conformational sampling of the protein by accelerating simulations, which are typically 100 times faster than with explicit solvent, but on the other hand this usually disrupts time scales and makes impossible to get the correct folding kinetics. Water implicit models are much faster to use but still suffer from several issues (for example the impossibility to describe buried water) and are less accurate in reproducing the native state of a protein [53].

1.3.2 Empirical all-atom force fields

In order to calculate forces in Eq. (1.6) we need to know the potential energy function $U(\mathbf{X})$. In principle, at a classical level this is given by considering all the electrostatics interactions among all the atoms of the molecule. Using such a level of detail would yield prohibitively long calculations, and a more approximate approach is needed. In order to avoid an infinite computational time and retain a satisfactory accuracy, the so-called empirical Force Fields (FF) are now of common use [55].

FF are defined by a potential energy function depending on the molecule configuration $\mathbf{X}$ and by the set of free parameters appearing in this function. All the FF used to simulate biomolecules display the same functional form, that is

$$U(\mathbf{X}) = \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\chi (1 + \cos(n\chi - \delta)) + \\ + \sum_{\text{impropers}} k_\phi (\phi - \phi_0)^2 + \underbrace{\sum_{i<j} \left\{ \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\}}_{\text{non-bonded}}. \tag{1.7}$$

The first term describes bonds vibrational energy, which is approximated with a harmonic potential acting on the bond distance b, of coupling constant k_b and equilibrium distance b_0. The same approximation is used in

the second term to describe the potential of valence angles θ, with coupling constant k_θ and equilibrium angle θ_0. The third term describes the energy of dihedral angles χ rotation, by means of a sinusoidal potential of constant k_χ, multiplicity n, and phase angle δ. The fourth term is again a harmonic potential acting on the improper dihedral angles ϕ. k_ϕ is the coupling constant and ϕ_0 the equilibrium value. The terms illustrated so far describe all bonded interactions, due to the chain topology. Non-bonded interactions are described in the last two terms contained in the curl brackets, where the summation runs over all pair of atoms (i, j). The first term approximates interatomic Van der Waals interactions by means of a Lennard-Jones potential. For any pair of atoms (i, j), r_{ij} is their distance, and ϵ_{ij} is the depth of the energy well whose minimum is located at $r = r_{ij}^0$. The last term represents the energy of electrostatic interactions between atomic partial charges q_i and q_j, ϵ being the dielectric constant in vacuum.

All the interactions but the electrostatic one are short-ranged, meaning that one can safely neglect them when two atoms are separated by a given threshold distance (namely 7 Å) thus making calculations faster. Electrostatic are the only long-range interactions, since they scale as r^{-1}, and imposing a cut-off would be an exceedingly crude approximation, and more sophisticated tricks are needed. The calculations of this long-range interactions amounts to roughly half of all the CPU time spent to simulate the system.

The free parameters of the energy function (1.7) are obtained by fitting experimental data or results of sophisticated quantum *ab initio* simulations. The currently most used FF, namely CHARM, AMBER, GROMOS, OPLS, have all in common the functional form (1.7), but differ on how the free parameters are calculated and on their values. Refined versions of FF can be found in which only a handful of parameters has been optimized [56, 57].

The main advantage of using the simplified energy function provided by empirical FF is that calculations are relatively efficient and, despite the somehow crude approximations used to write them, they are remarkably accurate in reproducing experimental data [58, 59].

Among the disadvantages we can list the fact that the chemical bond term is described by a harmonic oscillator, thus creation and breaking of bonds are ruled out. Using instead the Morse potential would yield a more realistic description, but with a 3 or 4 orders of magnitude higher computational cost [60].

Empirical FF are relatively computationally cheap because they are fitted to reproduce a given model in a specific range of thermodynamic parameters, for example temperature. Thus, they lack the generality of an *ab initio* quantum mechanical calculation, and are working at best for the system and the thermodynamics parameters they were optimized for [24]. Another *caveat* is that usually current FF are optimized to be used with a specific water model, e.g., TIP3P with AMBER and CHARM, and changing model could yield a loss of accuracy [58]. None of the nowadays standard FF used to simulate biomolecules can take into account atomic polarizability.

Simulating the dynamics of proteins in explicit water in AA resolution by employing empirical FF is the most detailed and accurate way to investigate the problem from a currently feasible computational point of view. Although empirical FF are cheap compared to quanto-mechanical calculations, for most systems it is still impossible to obtain simulations long enough to overlap with time scales of biological interest. Getting this sort of long MD trajectories would take an unreasonable amount of time by using most of the currently available supercomputers, which have routinely access to the microsecond scale for small systems ($\sim$ 1000 atoms). As a matter of fact, the first MD simulation with realistic FF of a folding protein was obtained only in 2010 on a highly specialized machine, which in virtue of a special hardware architecture can simulate small proteins in water on a millisecond scale. The latest achievements and issues will be reviewed in section 1.3.4.

1.3.3 Gō-type models

As we have just mentioned, severe limitations exist on the set of phenomena possible to be investigated by means of AA MD simulations. Besides, even when the folding of a given protein can be studied in the highest resolution, it is necessary to collect many events to draw any statistically significant conclusion.

For these reasons it is very useful and common to employ simplified energy functions, the most successful of which are the so-called Gō-models [61–63]. These adopt a native-centric view of folding and only native contacts play a stabilizing role in the native state. In other words, in the standard Gō-model the energy landscape is a perfect smooth funnel. They can be used both with an AA and a CG representation of a molecule.

A typical Gō energy function appears as [64]:

$$U(\mathbf{X},\mathbf{X_o}) = \sum_{\text{bonds}} k_b (b-b_0)^2 + \sum_{\text{angles}} k_\theta (\theta-\theta_0)^2 + \\ + \sum_{\text{dihedrals}} k_\chi \left\{ \left[1-\cos(\chi-\chi_0) + \frac{1}{2}\left[1-\cos(3(\chi-\chi_0))\right]\right]\right\} + \\ + \underbrace{\sum_{(i,j)\in NC} \epsilon_{NC} \left[5\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^0}{r_{ij}}\right)^{10}\right] + \sum_{(i,j)\notin NC} \epsilon_{\overline{NC}} \left(\frac{\sigma_{\overline{NC}}}{r_{ij}}\right)^{12}}_{\text{non-bonded}}. \tag{1.8}$$

The energy function is knowledge-based, because it depends on both the current and the native configurations. Indeed, all the equilibrium parameters (those with the 0 subscripts) are those of the native structure of the protein one wants to simulate. This automatically guarantees that the native structure is the minimum energy configuration. The first three terms of (1.8) take into account the fact that we are simulating an unbreakable chain, and describe the energy related to stretching of bonds, vibration of angles and rotation of dihedral. These are different if compared to (1.7) because b_0, θ_0 and χ_0 are average values calculated directly from the native structure of a particular protein. The non-bonded part is composed by a Lennard-Jones potential, describing attraction and repulsion, which in this case acts only on couples of residues (i,j) that are in native contact (NC), and a purely repulsive term acting on couples of residues which are distant in the native state ($\overline{\text{NC}}$). In this particular case, since $\sigma_{\overline{NC}}$ is a constant, the non-native repulsion is non-specific. All energy parameters k_b, k_θ, k_χ, ϵ_{NC}, and $\epsilon_{\overline{NC}}$ can be expressed in terms of ϵ_{NC}, and adjusted in order to recover the experimental stability ΔG.

Eq. (1.8) describes the effective energy of a smooth funnel, but Gō-models can be complicated to take into account also for non-native attractive interactions, dependent or not on the type of residues in contact. Further details will be given in Chap.'s 3 and 4.

Gō-models have been successful in reproducing several important observables of the kinetics and the folding mechanism of a wide range of different proteins [65], and are now a standard instrument to investigate the dynamics of biomolecules, and in particular the folding of proteins [61–63].

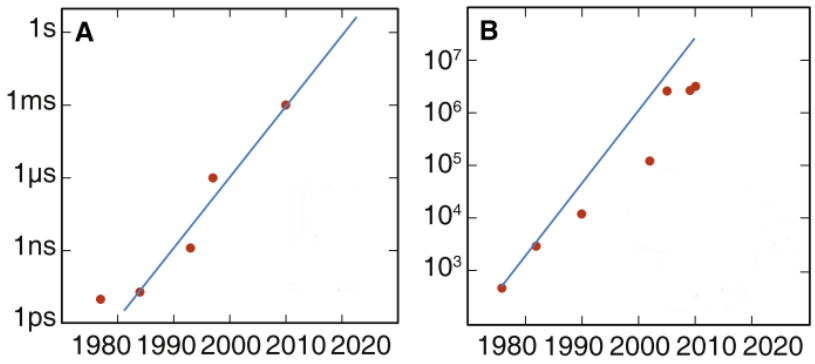


Figure 1.7: Panel A. Total duration of the growth of a protein simulation with time. First point reported is the simulation of Ref. [68], last point is the first millisecond long simulation reported in Ref. [69]. Scaling is exponential, with the accessible time scale doubling every year. It should be possible to simulate the dynamics of a protein on the second scale in 2020. Panel B. Growth of the total number of atoms of the simulated system with time. Fig. adapted with permission from Ref. [67].

1.3.4 All-atom MD simulations in the Anton era

The evaluation of the forces acting on a molecule is an extremely time-demanding computation, making MD calculations highly CPU-intensive. Moore's law is a well known empirical observation stating that computational power of CPUs doubles every 18 months [66]. This enhancement has a direct effect in the size of the system that is possible to simulate on a computer, as well as the total time duration of a simulation [67], which have been steadily increasing since the first MD simulation (Fig. 1.7) .

The attempts of using MD simulations to shed some light on the dynamics of proteins, and in particular on the folding problem, began in 1977. McCammon, Gellin and Karplus simulated for the first time 10 ps of the bovine pancreatic trypsin inhibitor (BTPI) dynamics on a computer using an empirical force field [68]. That effort served as a proof of principle, and somehow contributed to modify the common view at that time of proteins as rigid and static objects, whereas they are highly dynamic even in their native state. As a matter of fact, in 1998 Duan and Kollman accomplished the first serious try to completely fold a protein with a 1 μs long simulation of the villin headpiece subdomain [70].

Until few years ago, it was still under debate whether unbiased MD simulations with empirical FF were accurate enough to simulate the spontaneous folding of a realistic protein in AA resolution [71]. A definitive proof that this is possible was provided by Shaw *et al.* in 2010, who simulated the reversible folding and unfolding of a small WW Domain and of the BTPI by means of unbiased MD using the AMBER FF [69]. The obtained millisecond long equilibrium trajectories show spontaneous folding of the two small proteins to the correct native configurations. Further parameters characterizing the folding trajectories are in good agreement with experimental data. This milestone result by Shaw *et al.* represents the proof that MD and related computational approaches are a valuable strategy to cope with the protein folding problem.

Simulating the millisecond scale in AA resolution with explicit solvent has been possible thanks to several technological improvements [72]. MD calculations are extremely demanding and have to be distributed on a large amount of single CPUs by means of parallel algorithms and architectures, which are both under continuous development. However, the most important advancement has been the realization of the Anton supercomputer [73], named after the Dutch scientist Antonie van Leeuwenhoek. This computer is build employing special-purpose designed hardware that optimizes the computational steps of an MD simulation. This extreme level of hard-coded optimization permits to simulate MD steps two orders of magnitude faster than commercial clusters (Fig. 1.8).

1.3.4.1 Accuracy of current AA FF

The presence of Anton and the possibility to get millisecond long trajectories has changed the panorama of AA MD simulations, and modified the current view on three important related issues: the lack of sufficient sampling, the accuracy of modern empirical FF and a robust interpretation of the high resolution folding trajectories [60, 71, 74].

Sampling efficiently enough to simulate time scales relevant to biology has been the most severe limitation of AA MD approaches. This problem is far from being solved, since reaching the millisecond scale is limited to a unique supercomputer, whereas all commercial cluster are bounded to the microsecond scale. Even the millisecond scale is badly insufficient if one considers that an average protein folds on the second scale. However, looking

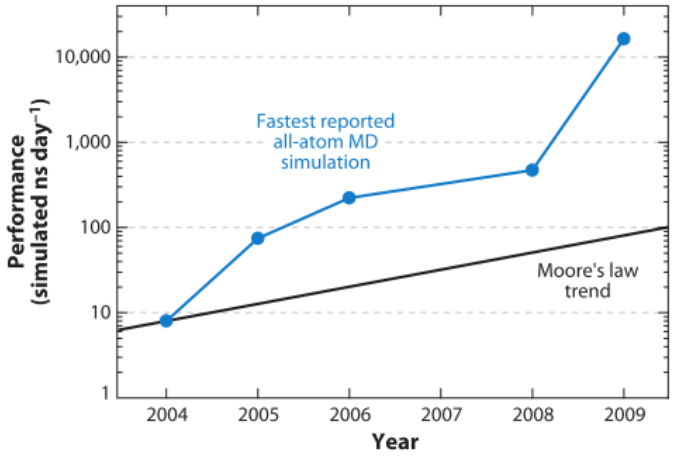


Figure 1.8: Fastest MD simulation evolution during the last years (regardless of the system's dimension, measured in simulated ns per day). Last point corresponds to the performance of the Anton supercomputer, namely several μs per day. Fig. reproduced with permission from Ref. [72]

at the growth of the simulated time scale during the years (Fig. 1.7) we can say that a reasonable hope exists that technological developments will mitigate the sampling problem in the next future [67].

Nonetheless, there is still much room to develop alternative sampling methods, which give up to all the details of a long equilibrium MD simulation to focus and enhance the sampling of a specific part of the dynamics under investigation, as we will see in the next chapter.

Empirical FF of the last generation, based on the functional form reported in Eq. (1.7), are somehow a drastic simplification of the exceedingly rich interactions in a real molecule surrounded by water. Therefore it is justifiable that until few years ago many researchers doubted that such energy functions could be accurate enough to reproduce experimental results. FF have been improving over the years, and although issues still exist, they have proved to be reasonably accurate [58]. Moreover, it is worth observing that FF are not guaranteed to properly act on timescales they were not optimized for. The fact that they yield reasonable results on longer available timescales is not trivial at all. Issues in AA FF (how transferable are they? how good are they in reproducing thermodynamic quantities? can they properly describe unfolded configurations?) have been investigated by means of ultra-long

simulations on the Anton supercomputer in a series of papers [69, 75–79] .

As we mentioned, Shaw *et al.* have demonstrated in Ref. [69] that spontaneous folding and unfolding of two small proteins happen on a long MD simulation. They extended this result by investigating the folding of 12 different (ultra)fast folder proteins, ranging from 10 to 80 residues in length (see Fig. 1.9), and all but one spontaneously folded to the experimental structure [75]. In a total simulated time of $\sim$ 8 ms about 400 folding and unfolding events have been observed, with kinetic and thermodynamic data in good agreement with experiments. The simulated proteins represent three different families of structures, namely α, β, and α/β proteins, and were folded by using the same FF (CHARM22*). This is remarkable considering that reproducing the correct interplay between α and β secondary structures is another open issue of realistic FF [57, 80].

Usually the folding rate is the first observable that is measured to attest the validity of a given MD simulation, and during the years many results obtained with different techniques have yielded a fairly good match with experiments. The existence of ultrafast folders as the villin headpice, which folds in few microseconds, permits to collect tens of (un)folding events on a millisecond long trajectory, enough to directly compute thermodynamic averages. Piana *et al.* used these simulations to show that, beside rates, also other kinetic and thermodynamic observables are in good agreement with experiments [77]. This conclusion can be extended beyond fast-folding proteins, as was shown by long MD simulations of ubiquitin, which folds on a millisecond scale [78].

Correctly reproducing the dynamics of unstructured proteins, which can be either found in the DS or because they are IDP, is a challenge. Lindorff-Larsen *et al.* investigated the performance of an AA FF by simulating a well characterized protein in its DS with a 200 microsecond long simulation [76]. As already reported in the literature, the FF was slightly too hydrophobic, yielding a DS more compact that the experimental one. Moreover, some transient α secondary structures persisted during all the duration of the simulation, pointing out that serious sampling issues still persist. Despite these limitations, the comparison of the dynamics in the DS with NMR measurements was fairly good [76].

To summarize, during the last three years, by using the unique features of the Anton supercomputer, Shaw and his group extensively tested the accuracy of modern FF and particularly addressed the most important long-

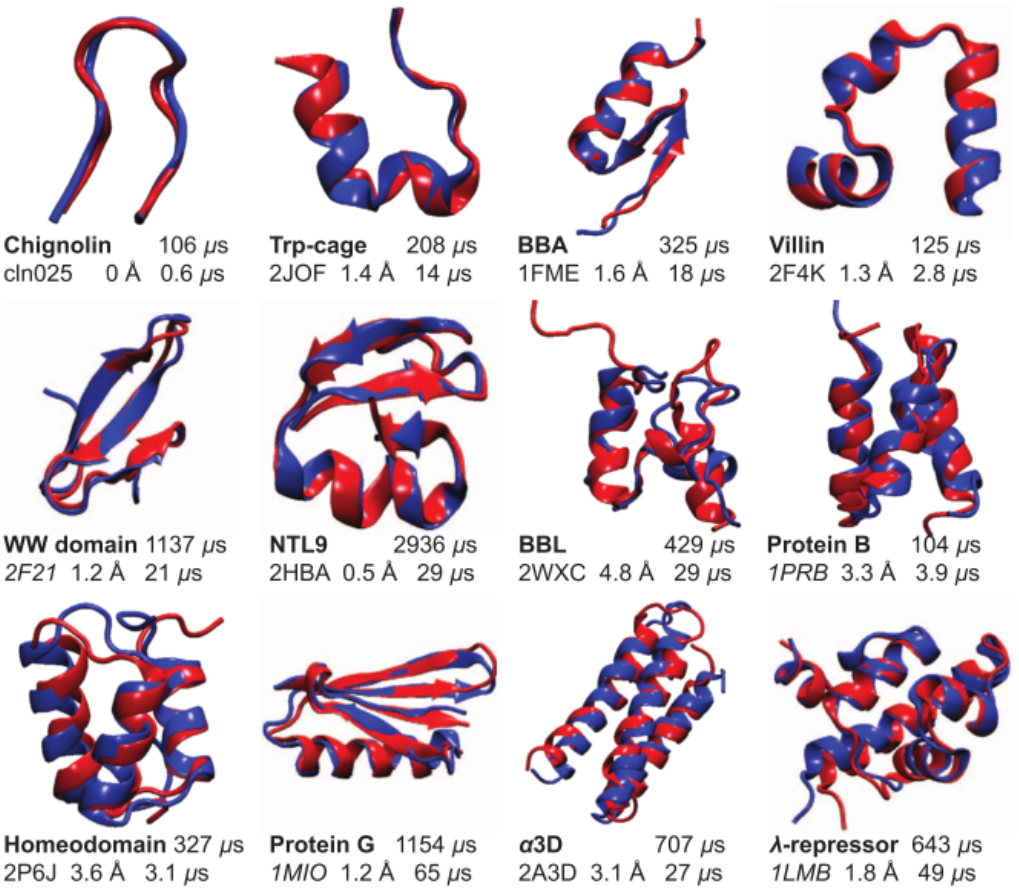


Figure 1.9: Twelve (ultra)fast folding proteins simulated on Anton by using the same FF [75]. For each protein the simulated native structure is shown in blue, superimposed to the experimentally determined one, which is in red. The structure representation is accompanied by the name of the protein, the total amount of simulated time, the RMSD-to-native calculated on all the residues Cα's and the average folding time. As it can be seen, all the simulated native structures are equivalent to the experimental one within the atomistic resolution of few Å. Fig. reproduced with permission from Ref. [75].

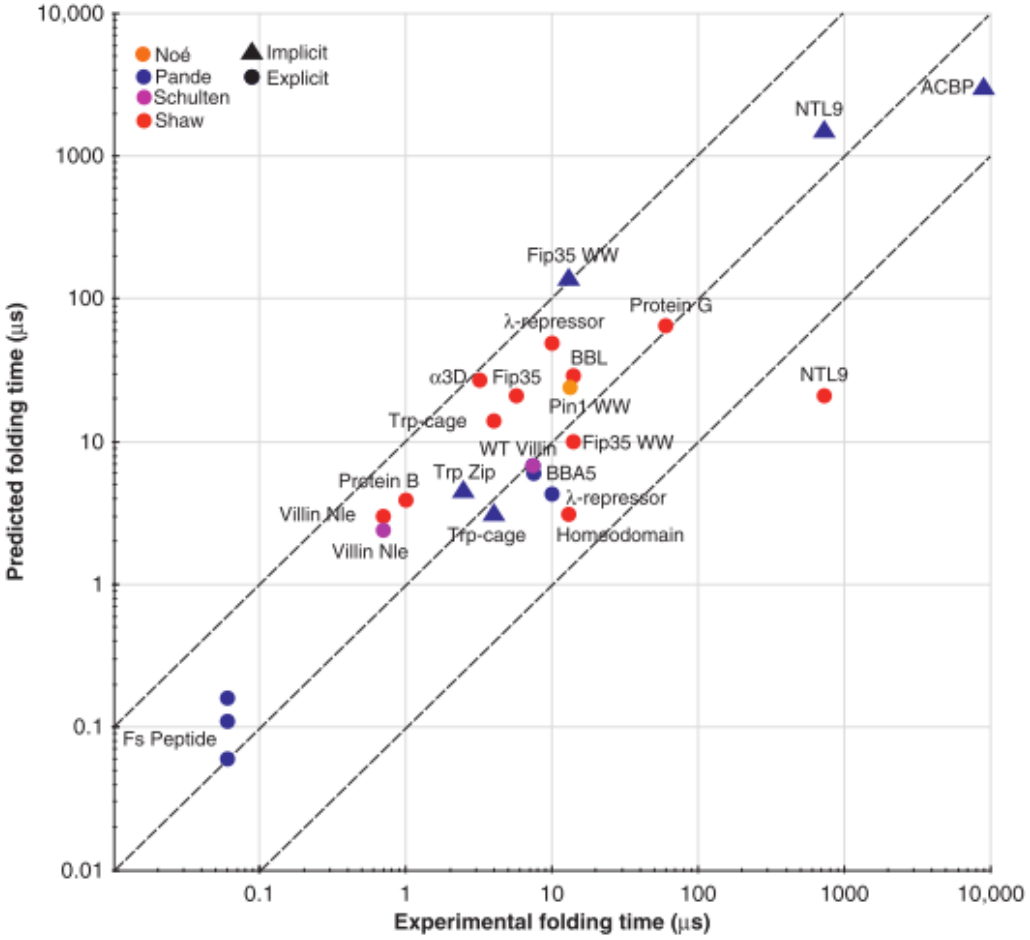


Figure 1.10: Comparison of simulated folding time with experimentally measured one. Different colors refer to different authors. Triangles and circles represent simulations carried on in implicit and explicit water, respectively. All but one results are in agreement within one order of magnitude. As noted by the authors of the review in Ref. [74], this is a reasonable agreement considering that simulations were carried on different temperature and environmental conditions in general compared to experiments. Reproduced with permission from Ref. [74].

standing open issues. The results of these investigations confirm that current FF are accurate enough to correctly reproduce many experimental measurements, and observed deficiencies highlight areas that need improvements in the future.

Long MD trajectories are highly complex objects, and extracting all their dynamical content is not a trivial task [74, 81]. Simplified models and dimensionality reduction techniques are extremely useful instruments to rationalize and characterize all the details of an AA folding simulation, as we will see in Chapter 2. Furthermore, sophisticated techniques based on a Markov Chain reduction have been very popular during the last years, as they provide a rigorous quantitative and human readable reduction of the complex dynamics encoded in MD trajectories. An example based on the Milestoning approach will be reviewed in Chapter 5.

1.3.5 Folding happens through sequential stabilization

There is plenty of excellent reviews focusing on many different aspects of the protein dynamics and folding phenomenology. In particular, several controversial aspects still exist, namely the order of events in folding, the role of transient secondary structures in the DS, the heterogeneity of folding pathways, and the role of non-native contacts [18]. Usually, experimental techniques measure properties of whole ensembles of proteins, and it is unfortunately extremely difficult to understand the mechanism of folding in atomistic detail, since transient structures cannot be resolved yet. The long MD trajectories produced by Anton and extensively validated represent a useful opportunity to summarize some general and common features of the observed folding events.

It appears that local native contacts are formed earlier than non-local ones, although some of the latter play a key role in stabilizing the TS. Residual secondary structures are present in the DS, and live on many different timescales. The longer a given secondary structure lives in the DS, the more likely it can be found in the NS, thus residual structures and native conformations are correlated. According to the definition of pathways we have already specified, folding as observed on Anton's simulations is rather homogeneous and somehow sequential. Structural elements are mainly formed in the same order, and a small number of different pathways is found. These conclusions emerge both from the insight on (ultra)fast folders [75] and on

the millisecond scale folder ubiquitin [78].

There is an apparent clash between the theoretical expectation based on the funneled energy landscape theory (i.e., that folding is a highly parallel phenomenon) and the conclusions of AA MD simulations. It is difficult to discriminate this by means of experiments, because the only way to detect multiple pathways is by analyzing the ϕ-value, although it is not at all straightforward.

The sequential view of folding, where local native contacts form earlier than non local ones, is in qualitative agreement with the "foldon" view. A foldon is defined as a "cooperative formation of pieces of secondary structures or loops, often contiguous in sequence, in a process of sequential stabilization" [18, 82]. Many experimental evidences point out that folding proceeds by forming this sort of quanta of secondary structures in a sequential way , suggesting that the effective energy of a protein is modulated by many local minima, each of which corresponding to a different foldon [18, 82, 83].

1.3.6 Role of non-native interactions

During the last years different researches carried on several investigations about the possible role of non-native interactions, i.e., roughness on the folding energy landscape.

Experimental attempts to shed light on this problem are very difficult, since it is impossible to follow the evolution of single contacts in time. Therefore, investigations have been carried on mostly by numerical approaches, although a frustrated landscape has been detected measuring the so called *internal friction* [84]. By employing Gō-models and lattice protein representations, a controversial picture emerged, where some results find a negligible role for non-native contacts, whereas other claim their importance [64, 85–94]. Those studies finding that the presence of non-vanishing frustration enhances the folding rate agree on the fact that non-native contacts have to be an energy perturbation compared to native ones. Indeed, when the contact energy in the two cases is equivalent, the amino acid chain behaves like a random heteropolymer [87].

We will shortly review the two most recent studies, since they represent the most systematic analyses carried on so far and stress two different possible roles of frustration.

Several authors found evidence that a low degree of frustration can actually accelerate folding [64, 87, 94, 95]. In particular, in Ref. [64] Contessoto *et al.* used a Gō-model with non-specific tunable non-native interactions to study kinetics and stability of 19 different proteins in Cα resolution by increasing the magnitude of frustration.

The proteins under attention can be divided in two groups, depending on the effect of non-native interactions: in one group a moderate frustration enhances folding rates, in the other it hampers them. The first group contains proteins displaying a high CO and a high folding activation free energy, $\Delta G^{\ddagger}$. The second group, on the contrary, is composed of proteins rich of local secondary structures, hence with low $\Delta G^{\ddagger}$ and CO. Thus, Contessoto *et al.* found a correlation between topology and frustration, suggesting that proteins with non-local native conformations take advantage of non-specific non-native interactions which stabilize the TS and therefore enhance the folding rate.

By considering a complementary point of view, Best *et al.* investigated the role of non-native interactions in determining the folding mechanism [93]. By exploiting the long AA folding trajectories simulated on Anton [75], they measured how long a given contact lives while a protein folds compared to its lifetime in the DS. They found that contacts showing high values of this ratio are positively correlated with native contacts, suggesting that the longer a contact exists during the folding, the more important it is to determine the mechanism. In all cases but one they found no statistically significant non-native contacts compared to the native ones. An additional Bayesian approach showed that, by following non-native contacts, one is not able to discriminate between the DS and the reactive part of the folding trajectory. The authors concluded that non-native interactions play no role in shaping the mechanism of all the proteins under consideration. The only exception is represented by α3D, which is a synthetic protein folding to a *de novo* structure. Presumably, this is because it was designed to display a stable folded configuration and this does not automatically imply that it has also a smooth energy landscape.

A similar difference between natural and artificial proteins was also found by Zhang and Chan [96], who simulated two almost homologous proteins (i.e., same native topology, same length), differing for their kinetics. The natural protein folds according to an exponential kinetics, whereas the artificial one folds through a complex multiphase kinetics. By simulating the two

molecules by means of a Gō-model enriched with non native interactions, the authors found that frustration has no role in the folding of the natural protein, whereas it causes a multitude of trapped meta-stable states in the folding of the artificial one.

Clarke's group obtained what is presumably the first experimental direct evidence of a rough landscape [84], which manifests as a measurable internal friction. The researches extensively characterized three homologous spectrin domains, which are three α-helices bundle proteins [97–99]. Two of them display the same folding times, whereas the third needs a longer time to fold. A solid explanation for this difference is that the slower spectrin domain is characterized by a much rougher energy landscape, which causes a friction term that is not related to the solvent and slows down the folding kinetics. On a microscopic level, this sort of frustration is possibly due to mis-docking events between the three α-helices [97–99]. Best has tried to rationalize these findings by employing a Gō-model [100], which yields to a correct description of the mechanism but is unable to reproduce the different folding times of the three spectrin domains, compatibly with what found in [93].

Chapter 2

Simulating reactive folding pathways on a computer

"Everything that is living can be understood in terms of the jiggling and wiggling of atoms."

R. Feynman

It is surprising to think that one of the richest area in physics and mathematics origins from the observations of the botanist Robert Brown concerning the restless dance of microscopic pollen in water. These small particles are being scattered around following what everybody would say to be a random walk. In the first section of this Chapter, we will consider this initial experiment to derive the Langevin and Smoluchowski equations and briefly outline their properties. Then we will understand how in Brownian motion the probability to observe a given transition between two points is given by summing the probabilities of all the paths connecting the two points. This path integral representation is known as the Wiener integral. It will be then natural to introduce the stochastic action functionals, which will play a central role in the method developed and tested in this thesis. This first section is almost self-contained, and represents a simple way to derive and understand the stochastic action and path integral formulations. The material shown is not original, but based on Ref.'s [101–103].

Historically protein folding is represented as a diffusive process along a thermally activated free energy barrier. Such a picture relies on the existence

of a good reaction coordinate for folding. The second section will be entirely devoted to review some recent theoretical and experimental results that show that indeed a satisfactory reaction coordinate exist, that is the fraction of native contacts.

We will then assume that diffusion is a good description for folding, and discuss how it is possible to take advantage of the instruments developed in the first section to find the statistically dominant folding trajectories and pathways. Eventually, we will introduce in detail the core-algorithm of this thesis, which yields representative folding pathways by a sampling and scoring procedure in the functional space of folding trajectories, given that the denatured and native configurations are provided.

2.1 Stochastic action

The first attempt to theoretically describe Brownian motion would be to solve Newton's equations for all the particles in the system, i.e., both the pollen particles and the water ones. This description considers the solution of a large number of coupled differential equations. However, we have seen in Chapter 1 that considering the evolution of the solvent on average greatly enhances the possibility of a theoretical description.

2.1.1 Langevin equation

We can avoid to explicitly describe the motion of water using an experimental fact. When an object is dragged in water it feels a resistance that is proportional to its velocity, which is the friction γ. Averaging the motion of water molecules, which are far smaller than the pollen one, turns out in an effective action on the pollen particle.

We thus can write an approximate version of Newton's equation for the Brownian particle

$$m\ddot{\mathbf{r}} = -\gamma\dot{\mathbf{r}} \tag{2.1}$$

since the only force acting from outside the system is gravity, that in this case is irrelevant. The solution of this equation reads

$$\dot{\mathbf{r}}\left(t\right) = \dot{\mathbf{r}}\left(0\right) e^{-\frac{\gamma}{m}t},$$

which implies that after a time

$$t > m/\gamma$$

the original velocity would be completely dispersed by the resistance of water. This is clearly in contradiction with the observation of the constant random scattering typical of Brownian motion. It is worth noting that the random collisions of water molecules cannot always happen in such a direction to hamper the motion of the pollen particle. In fact in many occasions they would actually impact in the same direction, transferring momentum and energy from the heat bath to the particle. Hence, we must add a force to Eq. (2.1) in order to take into account this effect

$$m\ddot{\mathbf{r}} = -\gamma\dot{\mathbf{r}} + \mathbf{f}(t) \tag{2.2}$$

This force is intuitively time dependent, and random in nature. This apparently innocent observation bears deep consequences. Indeed, we are now not able anymore to solve this equation, since we actually do not have any control on $\mathbf{f}(t)$, which is a stochastic function (i.e., it assumes random values in time). Let us consider again the over-damped time scale. After a rescaling

$$\gamma \to \frac{\gamma}{m}$$

we obtain the equation

$$\dot{\mathbf{r}}(t) = \frac{1}{\gamma}\mathbf{f}(t)$$

and its solution

$$\mathbf{r}(t) = \mathbf{r}(0) + \frac{1}{\gamma}\int_0^t \mathrm{d}\tau \mathbf{f}(\tau) \ .$$

Since we are dealing with random effects, the most we can ask for is to study the evolution of the system on average, i.e.,

$$\langle \mathbf{r}(t) \rangle = \langle \mathbf{r}(0) \rangle + \frac{1}{\gamma}\int_0^t \mathrm{d}\tau \langle \mathbf{f}(\tau) \rangle \ .$$

The meaning of the average is two-fold. Suppose we have N different jars containing water, and in each of them we prepare a Brownian particle in the same initial position $\mathbf{r}(0)$. We measure for each of them the different time evolution $\mathbf{r}(t)$. What we expect is this evolution to be isotropic, that

is to have an equal number of displacements in any direction. Therefore the average over a high number of different realizations of the same experiment will return

$$\langle \mathbf{r}(t) \rangle = \langle \mathbf{r}(0) \rangle$$

and in this sense we are averaging on an ensemble.

On the other hand, the typical time scale between two different collisions is much smaller than the one characterizing the evolution of $\mathbf{x}(t)$. Therefore, we demand the time average of the random force on any time interval Δt such that

$$t_{\text{coll}} \ll \Delta t \ll t_{\text{evo}}$$

where t_{coll} is the typical time of a collision and t_{evo} that of the particle's evolution. Hence, we can set the first characteristic the random force has to satisfy

$$\langle \mathbf{f}(\tau) \rangle = 0 \tag{2.3}$$

where the average is a time or an ensemble one.

The next interesting quantity to study is the mean squared displacement of the particle

$$\left\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \right\rangle = \frac{1}{\gamma^2} \int_0^t \mathrm{d}\tau \int_0^t \mathrm{d}\tau' \, \langle \mathbf{f}(\tau) \, \mathbf{f}(\tau') \rangle \, . \tag{2.4}$$

We now demand that two consecutive collisions are uncorrelated after a time interval such that $\Delta t \gg t_{\text{coll}}$. This can be approximated as

$$\langle f_i(\tau) \, f_j(\tau') \rangle = 2D\gamma^2 \delta_{ij} \delta(\tau - \tau') \tag{2.5}$$

which is different from zero only considering kicks at the same time and in the same direction, $2D\gamma^2$ being the strength of the random force. This approximation means that the thermal bath instantaneously loses memory of the direction of a collision. Using Eq. (2.5) in (2.4) we obtain

$$\left\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \right\rangle = 6Dt \, . \tag{2.6}$$

The last equation shows a famous result, stating that the average displacement has a very peculiar $\sqrt{\langle \mathbf{r}(t) \rangle} \sim \sqrt{t}$ evolution with time, which is a hallmark of Brownian motion and diffusion. Furthermore, it clarifies the meaning of D, the diffusion constant, which has the dimension of the square of a length on a time.

Intuitively, the strength $2D\gamma^2$ of a kick that the immersed particle suffers should be related to the physical features of the heath bath. Since we all have learnt that the macroscopic parameter called temperature T is directly proportional to the average velocity of the microscopic components of the system, we expect that $D\gamma^2$ depends also on T. To understand this, we have to recall the Langevin equation (2.2) and solve it. This can be done by multiplying (2.2) by the factor $\exp(\gamma t/m)$, and nothing that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(m\dot{\mathbf{r}}e^{\gamma t/m}\right) = \mathbf{f}(t)\, e^{\gamma t/m}$$

which can be immediately solved

$$\dot{\mathbf{r}}(t) = e^{-\gamma t/m}\dot{\mathbf{r}}(0) + \frac{1}{m}\int_0^t \mathrm{d}\tau \mathbf{f}(\tau)\, e^{\gamma(\tau-t)/m}\,.$$

Now, recalling condition (2.3), we can take the average and obtain

$$\langle\dot{\mathbf{r}}(t)\rangle = e^{-\gamma t/m}\dot{\mathbf{r}}(0)\ .$$

This means that for long times the memory of the initial velocity is completely lost. Again, that does not mean that after a time $t \gg m/\gamma$ the particle will be at rest, but that the velocity due to the constant bombardment of the environment will be equally likely in any direction. We can now consider the velocity correlation function

$$\begin{aligned}
&\langle\dot{r}_i(t)\,\dot{r}_j(t')\rangle = \\
&e^{-\gamma t/m}\dot{r}_i(0)\, e^{-\gamma t'/m}\dot{r}_j(0) + \frac{1}{m}\int_0^t \mathrm{d}\tau \int_0^{t'} \mathrm{d}\tau' e^{-\gamma(\tau+\tau'-t-t')/m}\langle f_i(\tau)\, f_j(\tau')\rangle = \\
&\langle\dot{r}_i(t)\rangle\langle\dot{r}_j(t')\rangle + \frac{1}{m}\int_0^t \mathrm{d}\tau \int_0^{t'} \mathrm{d}\tau' e^{-\gamma(\tau+\tau'-t-t')/m}\langle f_i(\tau)\, f_j(\tau')\rangle
\end{aligned}$$

where the linear terms in the noise are discarded thanks to Eq. (2.3). By using property (2.5) and by integrating in τ', we get

$$\langle\dot{r}_i(t)\,\dot{r}_j(t')\rangle = \langle\dot{r}_i(t)\rangle\langle\dot{r}_j(t')\rangle + \delta_{ij}\frac{2D\gamma^2}{m}e^{-(t+t')\gamma/m}\int_0^t \mathrm{d}\tau e^{-2\gamma\tau/m}$$

which can be rewritten as

$$\begin{aligned}
\langle\dot{r}_i(t)\,\dot{r}_j(t')\rangle = &\quad \langle\dot{r}_i(t)\rangle\langle\dot{r}_j(t')\rangle + \delta_{ij}D\gamma e^{-\gamma(t'+t)/m}\left[e^{2\gamma t/m}-1\right] \\
= &\quad \langle\dot{r}_i(t)\rangle\langle\dot{r}_j(t')\rangle + \delta_{ij}D\gamma\left[e^{-\gamma(t'-t)/m} - e^{-\gamma(t'+t)/m}\right]\,.
\end{aligned}$$

For long times the second exponential in the square brackets dies, and we retain

$$\langle \dot{r}_i(t)\, \dot{r}_j(t') \rangle = \langle \dot{r}_i(t) \rangle \langle \dot{r}_j(t') \rangle + \delta_{ij} D\gamma \left[e^{-\gamma(t'-t)/m} \right] .$$

From the last equation we see that m/γ sets the velocity decorrelation time, i.e., the time after which the Brownian particle has lost the memory of its initial velocity.

We now want to consider the equilibrium, when enough time has passed to damp any initial velocity, i.e. $\langle \dot{r}_i(t) \rangle = 0$, and the residual one is due only to the interaction with the environment. Equilibrium establishes for long times, and mathematically speaking we have to take the limit

$$\lim_{t-t' \to \infty} e^{-\gamma(t'-t)/m} = 1$$

and the velocity correlation function reduces to

$$\left\langle \dot{\mathbf{r}}^2 \right\rangle = 3D\gamma\,.$$

We know that at equilibrium the equipartition theorem holds, that is, any degree of freedom appearing quadratically in the Hamiltonian brings an average thermal energy of $1/2 k_B T$, where k_B is the Boltzmann constant. Hence in this case we can write

$$E = \frac{1}{2} m \left\langle \dot{\mathbf{r}}^2 \right\rangle = \frac{3}{2} m D\gamma = \frac{3}{2} k_B T,$$

using which we can get

$$D = \frac{k_B T}{m\gamma}$$

which is known as the Einstein relation and is one of the simplest instances of the so-called fluctuation-dissipation theorem. In this case, the theorem states that the scale of mobility due to the thermal kicks D is intimately related to the dissipation due to the friction γ through temperature. Indeed, both are different manifestations of the effect of the thermal bath, depending on whether the microscopical collisions transfer energy and momentum to the system or remove them.

2.1.2 Smoluchowski equation

The over-damped Langevin equation describing diffusion (Eq. (2.2)) is a stochastic differential equation, and the subtle instruments of stochastic calculus are needed in order to deal with it. We look thus for an alternative description in terms of an ordinary differential equation, describing the same physics of Eq. (2.2) but on an statistical level. The easiest way to find this equation is by exploiting again an experimental fact. In a system of particles which diffuse in a solvent at equilibrium one has that

$$\mathbf{j}(\mathbf{r},t) = -D\nabla\rho(\mathbf{r},t)$$

where $\mathbf{j}(\mathbf{r},t)$ is the probability density current (probability over time and unit area) and $\rho(\mathbf{r},t)$ the probability density function (PDF) to find a particle of the system.

In a closed system particles are never created nor destroyed and thus their number is constant. Any variation in a given volume $\Omega_i \subset \Omega$ with given boundary $\partial\Omega_i$ is due to the current passing through this boundary. Gauss's theorem yields

$$\int_{\Omega_i} \mathrm{d}\omega \frac{\partial\rho}{\partial t} = -\int_{\partial\Omega_i} \mathrm{d}\sigma \cdot \mathbf{j} = -\int_{\Omega_i} \mathrm{d}\omega \nabla \cdot \mathbf{j}\,.$$

But since the integration volume and boundary are arbitrary, this relation has to hold also for the integrand functions, and we have

$$\frac{\partial\rho}{\partial t} = D\nabla^2\rho(\mathbf{r},t) \tag{2.7}$$

which is called the diffusion equation, and describes the statistical evolution in time of the system we were looking for. The diffusion equation can be solved by using standard techniques [102, 103], and the result is

$$\rho(\mathbf{r},t) = \frac{1}{(4\pi Dt)^{3/2}} e^{-\frac{\mathbf{r}^2}{4Dt}}$$

which satisfies the normalization condition

$$\int_{\Omega} \mathrm{d}\mathbf{r}\rho(\mathbf{r},t) = 1$$

since the particles of the system have to be somewhere. For the sake of notation we focus on the one-dimensional case. Generalization to higher dimensions is straightforward.

The *transition probability* or *propagator* or *kernel* K is defined by

$$K\left(x_t, t|x_0, 0\right) \mathrm{d}x_t \equiv \mathscr{P}\left\{x\left(t\right) \in \left[x_t, x_t + \mathrm{d}x_t\right], \, x\left(0\right) = 0\right\} \tag{2.8}$$

and represents the conditional PDF that the system prepared in x_0 at time t_0 is found at position x at time t. Brownian motion has all the characteristics of a memoryless or Markovian process, in which the propagator (2.8) time-evolves the PDF of the system according to

$$\rho\left(x, t\right) = \int_{-\infty}^{+\infty} \mathrm{d}x_0 K\left(x, t|x_0, t_0\right) \rho\left(x_0, t_0\right) \tag{2.9}$$

and the future evolution of the system does not depend on all the previous history. Using Eq. (2.9) in Eq. (2.7) yields

$$\frac{\partial K\left(x, t|x_0, t_0\right)}{\partial t} = D \frac{\partial^2 K\left(x, t|x_0, t_0\right)}{\partial t^2}$$

which is again a diffusion equation, that by considering the initial condition

$$K\left(x, t|x_0, t_0\right) \longrightarrow_{t \to t_0} \delta\left(x - x_0\right)$$

is solved by

$$K\left(x, t|x_0, t_0\right) = \frac{1}{\sqrt{4\pi D\left(t - t_0\right)}} e^{-\frac{\left(x - x_0\right)^2}{4D\left(t - t_0\right)}} . \tag{2.10}$$

The kernel has to respect the normalization condition

$$\int_{-\infty}^{+\infty} \mathrm{d}x K\left(x, t|x_0, t_0\right) = 1$$

i.e., at any time t the system has to be somewhere.

The space and time dependence in Eq. (2.10) simplifies if the system is homogeneous

$$K\left(x, t|x_0, t_0\right) \to K\left(x - x_0, t|t_0\right)$$

and if the diffusive process is stationary

$$K\left(x, t|x_0, t_0\right) \to K\left(x, t - t_0|x_0\right)$$

as in the case of Brownian motion.

The diffusion equation (2.7) can be derived also from the microscopic Langevin equation [102, 103]. This permits to generalize the result to an over-damped Langevin equation in an external potential

$$\dot{\mathbf{r}} = \frac{1}{\gamma}\left(\mathbf{f}\left(t\right) - \nabla U\left(\mathbf{r}\right)\right) \tag{2.11}$$

which has a corresponding so-called Smoluchowski equation for the transition probability (2.8)

$$\frac{\partial K}{\partial t} = \frac{1}{\gamma}\nabla \cdot \left(K \nabla U\right) + D\nabla^2 K \tag{2.12}$$

which is a particular form of the Fokker-Planck equation. The equilibrium solution for a general potential U is

$$\rho\left(\mathbf{x}\right) \propto e^{-\beta U(\mathbf{x})}$$

i.e., the equilibrium Boltzmann distribution.

2.1.3 Wiener path integrals

The transition kernel of Eq. (2.10) satisfies the very general Chapmann-Kolmogorov relation. Let us consider three PDFs describing the system, $\rho(x_0, t_0)$, $\rho(x', t')$ and $\rho\left(x, t\right)$, s.t. $t_0 < t' < t$. We can use the definition of kernel of Eq. (2.9) to write them down

$$\begin{aligned}
\rho\left(x,t\right) &= \int_{-\infty}^{+\infty} \mathrm{d}x' K\left(x,t|x',t'\right)\rho\left(x',t'\right) \\
\rho\left(x,t\right) &= \int_{-\infty}^{+\infty} \mathrm{d}x_0 K\left(x,t|x_0,t_0\right)\rho\left(x_0,t_0\right) \\
\rho\left(x',t'\right) &= \int_{-\infty}^{+\infty} \mathrm{d}x_0 K\left(x',t'|x_0,t_0\right)\rho\left(x_0,t_0\right) \ .
\end{aligned}$$

If the last relation is used in the first, and the result is compared with the second line, this combination yields

$$K\left(x,t|x_0,t_0\right) = \int_{-\infty}^{+\infty} \mathrm{d}x' K\left(x,t|x',t'\right) K\left(x',t'|x_0,t_0\right) \tag{2.13}$$

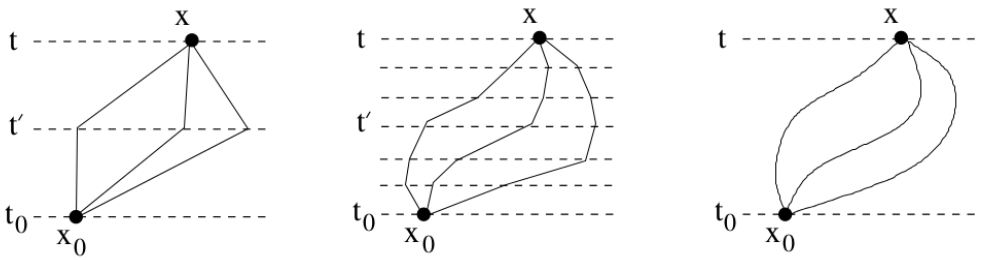


Figure 2.1: Pictorial representation of the Chapmann-Kolmogorov relation (Eq. (2.13)). On the left panel only one intermediate time is considered. In the central panel, the total time is sliced more often but still in finite number, whereas in the panel on the right the number of slices is infinite. Adapted with permission from Ref. [102].

which states that when a Brownian particle diffuses to x at time t, provided that it was at position x_0 at time t_0, if we check at an intermediate time t', we will find the particle at point x' somewhere in between (Left panel of Fig. 2.1). This very general result is based on the causality principle, according to which the system cannot disappear during its evolution from an initial to a final point, and holds for any Markovian process. If we check more often, i.e., we thicken the time slicing, Eq. (2.13) becomes

$$K(x,t|x_0,t_0) = \int_{-\infty}^{+\infty} \mathrm{d}x_{N-1} \ldots \int_{-\infty}^{+\infty} \mathrm{d}x_2 \int_{-\infty}^{+\infty} \mathrm{d}x_1 K(x,t|x_{N-1},t_{N-1}) \times \ldots \\ \ldots \times K(x_2,t_2|x_1,t_1)\, K(x_1,t_1|x_0,t_0)$$

and the total propagator is given by the convolution of all the propagators to intermediate positions x_1, $x_2, \ldots, x_{N-1}$ corresponding to the time slices t_1, $t_2, \ldots, t_{N-1}$, as represented in the middle panel of Fig. 2.1. We can intuitively guess that by thickening infinitely the time slicing, the total propagator would be given by the contribution over all microscopic paths connecting the initial and the final positions (Right panel Fig. 2.1).

We know want to formalize this intuitive insight. By the definition of PDF, the probability to find a Brownian particle at any time t in a given interval of positions is given by

$$\mathscr{P}\{x(t) \in [A,B]\} = \int_A^B \mathrm{d}x \rho(x,t)\ . \tag{2.14}$$

Brownian motion is memoryless, hence each displacement has a probability to occur that is independent on the others. It is thus easy to calculate the joint probability of a set of events as the product of the probabilities of each event, i.e.,

$$\begin{aligned}
&E \equiv \mathscr{P}\left\{x\left(t_1\right) \in \left[A_1, B_1\right], x\left(t_2\right) \in \left[A_2, B_2\right], \ldots, x\left(t_N\right) \in \left[A_N, B_N\right]\right\} = \\
&= \mathscr{P}\left\{x\left(t_1\right) \in \left[A_1, B_1\right]\right\} \times \mathscr{P}\left\{x\left(t_2\right) \in \left[A_2, B_2\right]\right\} \times \ldots \times \mathscr{P}\left\{x\left(t_N\right) \in \left[A_N, B_N\right]\right\} \\
&= \int_{A_1}^{B_1} \frac{\mathrm{d}x_1}{\sqrt{4\pi D t_1}} \exp\left\{-\frac{x_1^2}{4Dt_1}\right\} \int_{A_2}^{B_2} \frac{\mathrm{d}x_2}{\sqrt{4\pi D\left(t_2 - t_1\right)}} \exp\left\{-\frac{\left(x_2 - x_1\right)^2}{4D\left(t_2 - t_1\right)}\right\} \times \ldots \\
&\times \int_{A_N}^{B_N} \frac{\mathrm{d}x_N}{\sqrt{4\pi D\left(t_N - t_{N-1}\right)}} \exp\left\{-\frac{\left(x_N - x_{N-1}\right)^2}{4D\left(t_N - t_{N-1}\right)}\right\}.
\end{aligned} \tag{2.15}$$

This equation is the joint probability that the trajectory of a Brownian particle passes through N gates $[A_i, B_i]$, which are located at different times. We now shall thicken the gates, shrinking the time interval $\Delta t_i = t_i - t_{i-1} = \Delta t$ and raising N, thus considering the limit

$$\begin{cases} \Delta t \to 0 \\ N \to \infty \end{cases}$$

keeping the total time $N\Delta t$ constant. We shall apply this limit to Eq. (2.15), and consider gates of infinitesimal width $\mathrm{d}x_i$

$$\begin{aligned}
\lim_{\Delta t \to 0, N \to \infty} E &= \lim_{\Delta t \to 0, N \to \infty} \exp\left\{-\sum_{i=1}^{N} \frac{\left(x_i - x_{i-1}\right)^2}{4D\Delta t}\right\} \prod_{i=1}^{N} \frac{\mathrm{d}x_i}{\sqrt{4\pi D\Delta t}} \\
&= \lim_{\Delta t \to 0, N \to \infty} \exp\left\{-\frac{1}{4D} \sum_{i=1}^{N} \left(\frac{x_i - x_{i-1}}{\Delta t}\right)^2 \Delta t\right\} \prod_{i=1}^{N} \frac{\mathrm{d}x_i}{\sqrt{4\pi D\Delta t}} \\
&= \exp\left\{-\frac{1}{4D} \int_0^t \mathrm{d}\tau \dot{x}^2\left(\tau\right)\right\} \prod_{\tau=1}^{t} \frac{\mathrm{d}x\left(\tau\right)}{\sqrt{4\pi D \mathrm{d}\tau}}
\end{aligned} \tag{2.16}$$

Note that the infinitesimal gate width removes the integration over the arrival position. Therefore, the last line of the expression represent the probability that a Brownian particle goes through an infinite number of infinitesimal gates located at any time instant between the initial and the final points. Stated more simply, last line of Eq. (2.16) is the probability for a Brownian

particle to follow a given trajectory $x(\tau)$. Elementary principles of probability prescribe that to evaluate the probability to go from one point to a set of points $[A, B]$ one has to sum the contribution given by each trajectory,

$$\begin{aligned}\mathscr{P}\{x(t) \in [A,B]\} &= \int_{x(0)=0}^{x(t)\in[A,B]} \prod_{\tau=1}^{t} \frac{\mathrm{d}x(\tau)}{\sqrt{4\pi D \mathrm{d}\tau}} \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} \\ &= \int_{x(0)=0}^{x(t)\in[A,B]} \mathcal{D}x(\tau) \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} \\ &= \int_A^B \frac{\mathrm{d}x}{4\pi D t} \exp\left\{-\frac{x^2}{4Dt}\right\}\end{aligned} \tag{2.17}$$

where the last equivalence is due to Eq. (2.14). The integral sign is a formal way to say sum over all possible paths connecting the initial to the final point. We have introduced the formal measure

$$\mathcal{D}x(\tau) \equiv \prod_{\tau=1}^{t} \frac{\mathrm{d}x(\tau)}{\sqrt{4\pi D \Delta t}}$$

which represents the "volume" associated to each path $x(\tau)$. We have to note that this is really just a formal way of writing, since being a product of infinite terms (Eq. (2.16)), its finiteness and existence are not guaranteed. Integrals of the type of Eq. (2.17) are known as Wiener integrals.

If we ask the probability to go just to a single point, that is we put $A = B$, then by definition (2.8) we get the kernel

$$\begin{aligned}K(x_t, t|x=0, t=0) &= \int_{x(0)=0}^{x(t)=x_t} \mathcal{D}x(\tau) \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} \\ &= \frac{1}{4\pi D t} \exp\left\{-\frac{x^2}{4Dt}\right\}\end{aligned} \tag{2.18}$$

where in the last line we used Eq. 2.10. The link between the kernel and the probability is given by

$$\begin{aligned}&\int_{x(0)=0}^{x(t)\in[A,B]} \mathcal{D}x(\tau) \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} = \\ &= \int_{-\infty}^{+\infty} \mathrm{d}x_t \chi_{[A,B]}(x_t) \int_{x(0)=0}^{x(t)=x_t} \mathcal{D}x(\tau) \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\}\end{aligned}$$

where we have introduced the characteristic function of the interval $[A, B]$ defined as usual as

$$\chi_{[A,B]}(x) = \begin{cases} 1 & x \in [A, B] \\ 0 & \text{otherwise} \end{cases} .$$

Eq.s (2.17) and (2.18) express the probability of a given transition as a sum over all possible paths connecting the initial to the final configurations, and for this reason they are known as (Wiener) path integrals. These results, which at this point seem just a more complicate way to write down basic results, are the naturally extension of a PDF of functions to the realm of *functionals,* which are "functions of functions", and are usually represented as $F[x(\tau)]$. More precisely we define a functional F acting on a set of functions defined in a domain $\mathcal{D}$

$$\begin{aligned} F: \quad & \mathcal{D} \to \mathbb{R} \\ & x(\tau) \mapsto F[x(\tau)] . \end{aligned}$$

The Wiener path integral contains the generalization of the PDF to a Probability Density Functional (PDFl). Indeed, if we define

$$\Psi[x(\tau)] \equiv \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} \tag{2.19}$$

then it is natural to make the analogy

$$\rho(x_1, x_2, \ldots, x_N) \longrightarrow \Psi[x(\tau)] .$$

A PDF is defined to convey the probability s.t.

$$\rho(x_1, x_2, \ldots, x_N)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_N \propto \quad \mathscr{P}\{y_1 \in [x_1, x_1 + \mathrm{d}x_1], \ldots, y_N \in [x_N, x_N + \mathrm{d}x_N]\}$$

so extending the analogy we can write that

$$\Psi[x(\tau)]\,\mathcal{D}x(\tau) \propto \mathscr{P}\{y(\tau) = x(\tau) + \delta x(\tau)\}$$

which describes the probability of finding a function $y(\tau)$ in the infinitesimal tube around a given function $x(\tau)$, with $y(\tau)$, $x(\tau)$, and $\delta x(\tau)$ living in the same domain. Going on further, we can average any functional $F[x(\tau)]$ over all the Brownian trajectories writing

$$\langle F[x(\tau)]\rangle_\Psi = \frac{\int \mathcal{D}x(\tau)\,\Psi[x(\tau)]\,F[x(\tau)]}{\int \mathcal{D}x(\tau)\,\Psi[x(\tau)]} .$$

Mathematicians define the Wiener measure as

$$\mathrm{d}_W x(\tau) \equiv \mathcal{D}x(\tau)\,\Psi[x(\tau)]\,.$$

which has to be normalized according to

$$\int_{x_i}^{\Omega} \mathrm{d}_W x(\tau) = \int_{x_i}^{\Omega} \mathcal{D}x(\tau) \exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau \dot{x}^2(\tau)\right\} = 1.$$

2.1.3.1 Brownian trajectories are not differentiable

We shall recall that Eq. (2.6) is a hallmark of Brownian motion and diffusive processes. But it is also cause of a problem, since the limit

$$\lim_{t\to 0} \frac{\sqrt{\langle x(t)^2\rangle}}{t} \sim \lim_{t\to 0} \frac{t^{1/2}}{t} \to +\infty$$

diverges, thus derivatives are not defined and Brownian trajectories are not differentiable. On the other hand, the transition probability (2.10) satisfies

$$\lim_{t\to 0} K(x,t|x_0,t_0) = \lim_{t\to 0} \frac{1}{\sqrt{4\pi D(t-t_0)}} e^{-\frac{(x-x_0)^2}{4D(t-t_0)}} = \delta(x-x_0) \tag{2.20}$$

which means that after an infinitesimal time the Brownian particle is infinitesimally close to x_0. But Brownian motion is memoryless, hence the initial point x_0 can be any point along a trajectory $x(\tau)$, which means that the trajectory is continuous.

Brownian trajectories are continuous but non differentiable, therefore all the time derivatives symbols $\dot{\mathbf{r}}$ that we have used starting from Eq. (2.2) are just a formal expression, and remain ambiguous unless a discretization criterion is defined. This has also be provided to numerically solve the Langevin equation.

It is useful to rescale the stochastic noise function

$$f(t) = \gamma\sqrt{2D}\eta(t)$$

s.t. the new noise η has zero mean and unity variance. One of the most common choice is the Itō discretization rule, in which the over-damped Langevin equation describing free Brownian motion becomes

$$x_{i+1} - x_i = \sqrt{2D\Delta t}\eta_i\,. \tag{2.21}$$

This result yields also an easy way to show that the noise is described by Gaussian random variable η_n. From Eq. (2.21) it follows that the propagator expressed in terms of the noise and in term of a finite displacement are related by

$$K(\eta_i) = J(\Delta x, \eta_i) K(\Delta x, \Delta t)$$

where $\Delta x \equiv x_{i+1} - x_i$ and $J(\Delta x, \eta_i) \equiv \partial \Delta x / \partial \eta = \sqrt{2D\Delta t}$ is the Jacobian of the variable transformation. Thus, we get that

$$K(\eta_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\eta_i^2}{2}} \tag{2.22}$$

which is exactly the PDF of a Gaussian with unity variance. The stochastic force that we have introduced at the beginning of our analysis of Brownian motion is thus simply a Gaussian random variable with a variance fixed by the fluctuation-dissipation theorem. This simple model of the stochastic force is also known as white noise.

2.1.4 Stochastic action functionals

We want know to generalize the Wiener path integral to trajectories which are solution of the over-damped Langevin equation in presence of an external potential (Eq. (2.11)). This is obtained performing the change of variable $\dot{x}(\tau) \to \dot{x}'(\tau') + \frac{1}{\gamma} \nabla U(x')$ s.t.

$$\begin{aligned} &\mathcal{D}x(\tau) \exp\left\{ -\frac{1}{4D} \int_0^t \mathrm{d}\tau \dot{x}^2(\tau) \right\} \to \\ &J(\tau, \tau') \mathcal{D}x'(\tau') \exp\left\{ -\frac{1}{4D} \int_0^t \mathrm{d}\tau' \left(\dot{x}'(\tau') + \frac{1}{\gamma} \nabla U(x') \right)^2 \right\} \end{aligned}$$

in Eq. (2.19), but the Jacobian is not trivial to calculate. It can be shown that [102, 103]

$$J(\tau, \tau') = \frac{\delta x(\tau)}{\delta x'(\tau')} = \exp\left\{ \frac{1}{2\gamma} \int \mathrm{d}\tau \nabla^2 U \right\}$$

and the new propagator corresponding to a diffusive dynamics in presence of an external field is

$$K\left(x_t,t|x=0,t=0\right)= \int_{x(0)=0}^{x(t)=x_t}\mathcal{D}x\left(\tau\right)\exp\left\{-\frac{1}{4D}\int_0^t \mathrm{d}\tau\left(\dot{x}\left(\tau\right)+\frac{1}{\gamma}\nabla U\left(x\right)\right)^2+\frac{1}{2\gamma}\int_0^t \mathrm{d}\tau\nabla^2 U\right\}.$$

The double product resulting from the square in the exponential can be immediately integrated by part

$$\frac{1}{2D\gamma}\int_0^t \mathrm{d}\tau\dot{x}\left(\tau\right)\nabla U\left(x\right)=\frac{1}{2D\gamma}\int_0^t \mathrm{d}\tau\dot{U}\left(x\right)=\frac{1}{2D\gamma}\left[U\left(x_t\right)-U\left(x_0\right)\right]$$

and since it is not path-dependent it can be written outside the integral symbol. We can recollect the remaining terms in the exponential to get

$$K\left(x_t,t|x=0,t=0\right)= \exp\left\{\frac{1}{2D\gamma}\left[U\left(x_t\right)-U\left(x_0\right)\right]\right\}\int_{x(0)=0}^{x(t)=x_t}\mathcal{D}x\left(\tau\right)\exp\left\{-\int_0^t \mathrm{d}\tau\frac{\dot{x}^2\left(\tau\right)}{4D}-V\left[x\left(\tau\right)\right]\right\} \tag{2.23}$$

where we have defined the effective potential

$$V\left[x\left(\tau\right)\right]=\frac{1}{2\gamma}\nabla^2 U\left(x\right)-\frac{1}{4D\gamma^2}\left(\nabla U\left(x\right)\right)^2. \tag{2.24}$$

and the path independent contribution $\Delta U\equiv U(x_t)-U(x_0)$.

To avoid the calculation of the nasty Jacobian, by paying the price of a smaller insight, we can derive Eq. (2.23) by exploiting another analogy. Indeed, if we perform the change of variable $K\rightarrow K'$ s.t.

$$K\left(x_t,t|x_0,t_0\right)=e^{-\frac{U(x)}{2\gamma D}}K'\left(x_t,t|x_0,t_0\right)$$

in the Smoluchowski equation (2.12), we obtain

$$\frac{\partial K'}{\partial t}=D\nabla^2 K'+\left(\frac{1}{2\gamma}\nabla^2 U-\frac{1}{4\gamma^2 D}\left(\nabla U\right)^2\right)K'$$

which is analogous to the Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\psi+V\left(x\right)\psi$$

once we introduce the effective potential defined in Eq. (2.24). It is well known that the evolution of the wave function ψ is described alternatively by Feynman's path integrals, which express the quantum probability amplitude as

$$\langle x_t, t|x_0, 0\rangle \propto \int_{x(0)=0}^{x(t)=x_t} \mathcal{D}x\,(\tau) \exp\left\{\frac{\imath}{\hbar}\int_0^t \mathrm{d}\tau \frac{\dot{x}^2\,(\tau)}{2m} - V\,[x\,(\tau)]\right\}. \tag{2.25}$$

If we consider the Schrödinger equation in imaginary time, by performing a so-called Wick rotation

$$t \to \imath\tau$$

and we identify

$$\frac{\hbar}{2m} \longleftrightarrow D$$

the Wiener and Feynman path integrals coincide. We must bear in mind that there are important differences between the two path integrals. In quantum mechanics Eq. (2.25) yields a probability amplitude describing the propagation, whereas Eq. (2.23) yields already a probability. Moreover, the potential term appearing in the standard Feynman integral is the actual potential energy characterizing the system. In the stochastic counterpart written in Eq. (2.23), it is instead an effective potential related to the potential energy through Eq. (2.24), and does not have the dimension of an energy but that of a frequency.

Inspired by the formal analogy, we can define in Eq. (2.23) an effective Lagrangian

$$L = \frac{\dot{x}^2\,(\tau)}{4D} - V\,[x\,(\tau)]$$

having the dimension of an inverse time, and the corresponding action

$$\boxed{S_{\mathrm{OM}}\,[x\,(\tau)] = -\frac{\beta}{2}\Delta U + \int_0^t \mathrm{d}\tau \frac{\dot{x}^2\,(\tau)}{4D} - V\,[x\,(\tau)]} \tag{2.26}$$

which is a dimensionless number, known as the Onsager-Machlup *stochastic action functional* (OM action) [104, 105], that will have a central role in this work. The path integral formulation of the propagator of a diffusive process in presence of an external field can be thus cast in the attractive form

$$\boxed{K\,(x_t, t|x_0, 0) = \int_{x(0)=0}^{x(t)=x_t} \mathcal{D}x\,(\tau)\, e^{-S_{\mathrm{OM}}[x(\tau)]}} \tag{2.27}$$

To summarize, the last equation states that the probability for a Brownian process to evolve from an initial position to a given final one at time t is given by all the possible paths connecting the initial and final positions, each contributing with a weight proportional to $\exp\{-S_{\mathrm{OM}}[x(\tau)]\}$.

2.2 Folding as diffusion along a reaction coordinate

As we have seen in Chapter 1, protein folding is a complicate process which depends on a huge amount of degrees of freedom. From a formal dynamical point of view, its evolution takes place in a highly dimensional space, which is qualitatively represented to gain a more intuitive grasp as a bidimensional funneled landscape. On the other hand, experimental kinetics of small globular proteins is mostly well described as characterized by two stable states separated by a high free energy barrier. It is thus very convenient and widely used to interpret the folding as a thermally activated reaction between the native and unfolded states, which are both minima of a one-dimensional free energy function.

The simplest description is to consider folding as a diffusive process determined by a Langevin equation. Thanks to the so-called Mori-Zwanzig formalism, it is always possible to formally project the dynamics of a system described by Classical Mechanics on a smaller number of degrees of freedom. In particular, this projection technique can yield the extreme dimensional reduction needed to transform the exceedingly complicated dynamics of a protein in water into one evolving along a single coordinate. In general, though, the result is described by a non-Markovian generalized Langevin equation displaying memory kernels [101], and is thus depending on the history of the system. However, the projection would be memoryless if the system admits a good reaction coordinate. This is in general a function Q of the conformational space Ω of the system, s.t. $Q(\mathbf{x}) : \Omega \to \mathbb{R}$, and we can intuitively think of it as the privileged direction along which most of the probability current describing the evolution of the system flows. A good reaction coordinate not only separates the folded and the unfolded state, and measures the proceeding of the reaction, but also controls the dynamics [106].

There is not guarantee that for a generic system a good reaction coordinate exists. But since protein folding proceeds forming native contacts, it is

natural to consider the fraction of formed native contacts Q as the putative reaction coordinate. Indeed, on a perfectly smooth funnel, if the denatured configurations were uniformly distributed on the top of it, then Q would be the exact reaction coordinate [49], suggesting that Q should be a satisfactory choice at least for the folding simulated by using Gō-like models. Best and Hummer have tested this hypothesis in a series of papers [95, 106–108]. The 1-d Smoluchowski equation projected on Q can be rewritten as

$$\frac{\partial K}{\partial t} = \mathcal{L}K$$

where we have introduced the diffusion operator

$$\mathcal{L} = \frac{\partial}{\partial Q} D\left(Q\right) e^{-\beta G(Q)} \frac{\partial}{\partial Q} e^{\beta G(Q)} \tag{2.28}$$

where $K(Q_t, t|Q_0, 0)$ is the projected propagator, β is the usual Boltzmann factor, $G\left(Q\right)$ is the potential of mean force obtained by averaging out all the other degrees of freedom. Note that in general the projected Eq. (2.28) displays a position-dependent diffusion constant $D\left(Q\right)$. This implies that, in order to describe the dynamics, the knowledge of two independent functions is needed, whereas the interpretation of $G\left(Q\right)$ alone is no more straightforward. By analyzing reversible folding trajectories simulated using CG Go-models, Best and Hummer found that Q is a good reaction coordinate according to a maximum likelihood criterion [106]. In general it is possible to find a coordinate transformation such to have a position-independent diffusion coefficient. The latter is a well appreciated feature since it permits to have a direct interpretation of the free energy profile $G\left(Q\right)$, and this is particularly important especially with experimental results. In another paper, the two authors showed that by using Q as a reaction coordinate, D is almost position-independent [95]. This two results prove that Q is a good reaction coordinate at least for Gō-models, which is a reasonable conclusion since they are based on a smooth natively biased energy function. Moreover, in Ref. [93] the accuracy of Q was tested by means of the same maximum likelihood criterion as in Ref. [106] also on AA MD trajectories produced by Anton [75]. Even in this case, the fraction of native contacts turned out to be a satisfactory reaction coordinate, although the realistic FF determining the dynamics of these trajectories is rougher than a Gō-model.

This series of results makes us more confident that a decent reaction coordinate exists describing the folding of proteins, at least for the small,

globular ones with a simple topology. Furthermore, these findings put on a more solid theoretical ground the possibility to model the folding as a diffusion on a 1-d free energy barrier. This model is analytically tractable and widely investigated in the literature. Kramers found a closed expression for the mean first passage time (MFPT) to diffuse from the unfolded to the native state, that in a two-state model is the inverse of the folding rate, that is the barrier hopping frequency. If the process is thermally activated, i.e., the two basins are separated by a free energy barrier $\Delta G_f^{\ddagger}(Q) \gg k_B T$, then the MFPT is [109, 110]

$$\tau_f = \frac{1}{k_f} = \frac{2\pi}{\beta D^{\ddagger} \omega_u \omega^{\ddagger}} e^{\beta \Delta G_f^{\ddagger}} \tag{2.29}$$

where $D^{\ddagger}$ is the diffusion coefficient and $\omega^{\ddagger}$ is the curvature both evaluated on the top of the barrier, while ω_u is evaluated in the unfolded basin. The segment of a long trajectory which exits the DS and goes straight to the NS before going back again to the DS is called Transition Path (TP), as its time reversal. This is the part where the protein actually "hops" the barrier. By simulating the process on a computer, it can be immediately seen that the protein spends the overwhelming majority of the time oscillating by thermal motion in the two basins. The transition along the TP is actually extremely short. This observation agrees with the expression of the transition path time (TPT) τ_{TP}, as it was analytically calculated by Szabo [112]

$$\tau_{TP} \approx \frac{1}{\beta D^{\ddagger} \left(\omega^{\ddagger}\right)^2} \ln\left(2 \mathrm{e}^{\gamma_E} \beta \Delta G_f^{\ddagger}\right) \tag{2.30}$$

where γ_E is the Euler-Mascheroni constant ($\approx 0.577...$). The big difference between the two time scales is caused by the fact that Eq. (2.29) depends exponentially on $\Delta G_f^{\ddagger}$ whereas Eq. (2.30) only logarithmically. Only very recently single molecule experiments gained access to the TPT time scales of real folding proteins. Chung *et al.* reported the measurement of the TPT for the folding of two different proteins [111, 113], a short WW domain displaying a folding time $\sim 10^{-4}$ s, and the GB1, an α/β protein with a folding time ~ 1 s. By using sophisticated FRET techniques, which employ counting of single photons, the authors measured the TPT for the two proteins, which turned out to be $\sim 2\,\mu$s for the WW domain and $< 10\,\mu$s for GB1. Thus, although the folding times differ by for four orders of magnitude, the TPT are almost the same, that is compatible with Eq.s (2.29) and (2.30).

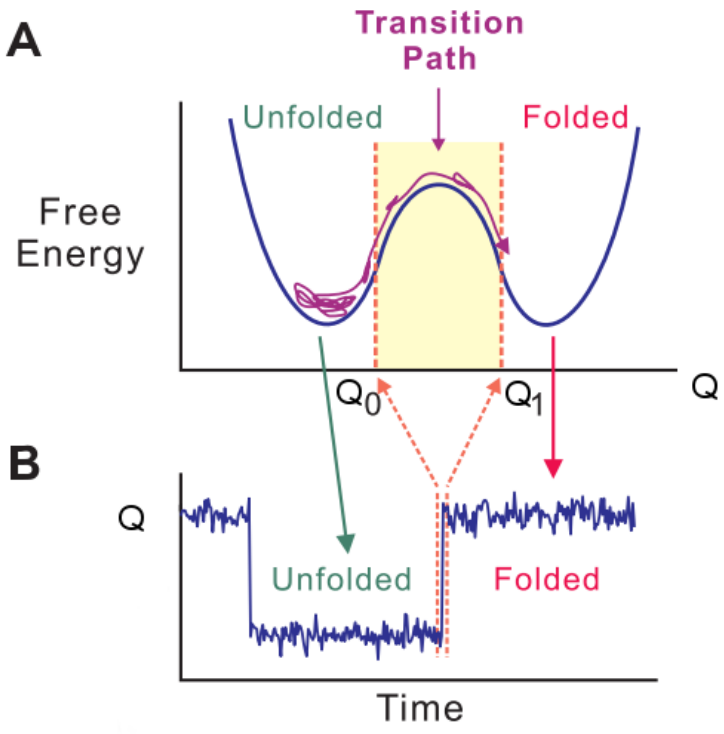


Figure 2.2: Schematic representation of the folding of a two-state protein as diffusion over a 1-d free energy landscape. Panel A: The double well free energy vs. Q, fraction of native contacts. The unfolded and folded basins are separated by a barrier $\gg k_B T$. The TP is a successful jump from one basin to the other. Panel B: Value of Q *vs.* time. The protein is spending most of the time in the basins, whereas overcoming the barrier seems an instantaneous process. Adapted with permission from Ref. [111].

2.3 Characterizing the reactive folding pathways

Given the results reviewed in last section, we are legitimate to think of a protein that folds to its native state as a Brownian particle in a double well. The protein wanders around in the DS basin, due to thermal motion, until a rare series of fluctuations push all in the same direction and the molecule can overcome the barrier $\Delta G_f^{\ddagger}$, and eventually the protein folds. In a long equilibrium microscopic trajectory, the almost instantaneous TP contains all the relevant information regarding the folding mechanism, and is thus the most interesting part to elucidate the mechanism. As we have seen in Chapter 1, an MD simulation is the most straightforward and standard way to numerically obtain such a long equilibrium trajectory. Unfortunately, MD is also extremely time-demanding, and a large part of the biologically relevant timescales is yet not accessible. This is due to the presence of many time scales in the dynamics of the protein folding: chemical bonds vibrate on the fs scale, dihedral angles rotate on the ns one; the formation of $\alpha-$helices and $\beta-$sheets takes $\sim$ 100 ns and $\sim$ μs respectively; folding occurs on a scale ranging from μs to minutes. In order to have a numerically stable MD simulation, the time step in Newton's discretized equations has to be of the order of the fastest motion in the protein, and therefore usually a 2 fs time step is used. This implies that a MD simulation has to overcome a huge gap in the time scales, and that for a millisecond folder a number of 10^{12} time steps has to be evaluated.

Most of the intense computational effort of an MD simulation is devoted to simulate the thermal motion of the protein in the DS, whereas the more interesting part in which the protein actually folds is contained in the very short successful transition over the free energy barrier. Thus, during the years, several groups have proposed diverse methods which aim to focus the computational effort to directly obtain and characterize the TP [114].

In particular, in the following we will introduce and discuss a recent TP-based method which we will use to characterize the folding of two different realistic proteins. We assume that the over-damped Lanvengin equation correctly describes the dynamics of each atom which composes the protein in water. Then, we are able to use the instruments derived in Section 2.1 to give a path integral representation of the probability to be in the native

configuration at time t as

$$P_f(t) = \int_\Omega \mathrm{d}\mathbf{x}_f \chi_{\mathrm{NS}}(\mathbf{x}_f) \int_\Omega \mathrm{d}\mathbf{x}_i \chi_{\mathrm{DS}}(\mathbf{x}_i) \rho_i(\mathbf{x}_i) \int_{\mathbf{x}_i}^{\mathbf{x}_f} \mathcal{D}\mathbf{x}(\tau) e^{-S_{\mathrm{OM}}[\mathbf{x}(\tau)]} \tag{2.31}$$

where Ω is the entire conformational space of the protein; $\chi_{\mathrm{NS}}(\mathbf{x}_f)$ and $\chi_{\mathrm{DS}}(\mathbf{x}_i)$ the characteristic functions of the NS and DS, respectively; $\rho_i(\mathbf{x}_i)$ is the equilibrium distribution of the protein's initial configuration; the last path integral is the propagator $K(\mathbf{x}_f, t|\mathbf{x}_i, 0)$, where $\mathbf{x}_f \in \Omega_{\mathrm{NS}}$ is the native conformation while $\mathbf{x}_i \in \Omega_{DS}$ is the denatured one.

2.3.1 The saddle-point approximation

Eq. (2.31) provides a microscopic representation of the folding probability, formulated in terms of the Langevin trajectories $\mathbf{x}(\tau)$ in configuration space Ω. We now want to develop a scheme to find the statistically most representative trajectories, that is those which account for most of folding probability $P_f(t)$.

Firstly, we shall approximate $P_f(t) \approx \sum_i K(\mathbf{x}_f, t|\mathbf{x}_i, 0)$, where the sum runs over different initial configurations, hence considering that the result of the remaining integrations in Eq. (2.31) contributes in the same way for each folding trajectory. Now we can focus on the propagator K, and look for the most probable microscopic trajectory connecting $\mathbf{x}_i$ to $\mathbf{x}_f$, that is clearly the one that minimizes the OM action 2.26. Since usually folding is a thermally activated process, thus it requires to overcome a free energy barrier $\gg k_B T$, we can take advantage of the saddle point approximation to determine the most probable folding trajectories.

The saddle-point approximation (also known as stationary phase approximation) is an extension of the standard Laplace approximation method for real-valued Riemann integrals, and is a standard tool in theoretical physics. The basic idea is the following: if the integrand function displays one clear maximum, then it can be approximated with the value it assumes in such an extremal point.

Without loss of generality, we consider only one dimension and firstly suppose the existence of just one path $\bar{x}(t)$ which extremizes $S_{\mathrm{OM}}[x(\tau)]$.

That saddle-point trajectory is defined by requiring that

$$\left.\frac{\delta S_{\mathrm{OM}}\left[x\left(\tau\right)\right]}{\delta x\left(t\right)}\right|_{x=\bar{x}}=0 \tag{2.32}$$

together with the boundary conditions

$$\begin{aligned} x\left(0\right) &= x_i \\ x\left(t\right) &= x_f \end{aligned}$$

Note that in Eq. 2.26 the path independent term can be ignored in the functional derivative 2.32, and in the following we shall consider only the integral part. It is well known that when the action is in the so-called Lagrangian form, i.e., when the integrand function is the difference of a kinetic and a potential energy, as in the case of S_{OM}, then the functional derivative (2.32) is given by the Euler-Lagrange equations

$$\frac{\delta S_{\mathrm{OM}}\left[x\left(\tau\right)\right]}{\delta x\left(t\right)}_{x=\bar{x}}=0\Longleftrightarrow\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial\dot{x}}L-\frac{\partial}{\partial x}L=0\,. \tag{2.33}$$

Hence, the stationary request is equivalent to solving the equation

$$\frac{\ddot{\bar{x}}\left(\tau\right)}{2D}=\nabla V\left[\bar{x}\left(\tau\right)\right] \tag{2.34}$$

whose solution is the saddle-point trajectory $\bar{x}\left(t\right)$.

Now we can perform a functional Taylor expansion of the action around the stationary path $\bar{x}\left(\tau\right)$

$$\begin{aligned} S_{\mathrm{OM}}\left[x\left(\tau\right)\right] &= S_{\mathrm{OM}}\left[\bar{x}\left(\tau\right)+\delta x\left(\tau\right)\right] \\ &= S_{\mathrm{OM}}\left[\bar{x}\left(\tau\right)\right]+\frac{1}{2}\int_{0}^{t}\mathrm{d}\tau\int_{0}^{t}\mathrm{d}\tau'\delta x\left(\tau'\right)F\left(\tau,\tau'\right)\delta x\left(\tau\right)+\mathcal{O}\left(\delta x^{3}\right) \end{aligned} \tag{2.35}$$

where we have introduced the fluctuations around the saddle-point trajectory

$$x\left(\tau\right)=\overline{x}\left(\tau\right)+\delta x\left(\tau\right)$$

which have to obey the boundary conditions

$$\delta x\left(0\right)=\delta x\left(\tau\right)=0$$

and we have also introduced the fluctuation operator

$$F(\tau,\tau') \equiv \left.\frac{\delta^2 S_{\mathrm{OM}}[x(\tau)]}{\delta x(\tau)\,\delta x(\tau')}\right|_{x=\bar{x}} . \tag{2.36}$$

We can now insert Eq. (2.35) in Eq. (2.27), obtaining

$$K \approx e^{-S_{\mathrm{OM}}[\bar{x}(\tau)]} \int \mathcal{D}\delta x(t)\, e^{-\frac{1}{2}\int \mathrm{d}\tau \int \mathrm{d}\tau' \delta x(\tau') F(\tau,\tau') \delta x(\tau)} . \tag{2.37}$$

The remaining part of the integral is a Gaussian path integral of the fluctuations $\delta x(\tau)$, and it can be solved by calculating the so-called fluctuations determinant

$$\int \mathcal{D}\delta x(\tau) \exp\left[-\frac{1}{2}\int \mathrm{d}\tau \int \mathrm{d}\tau' \delta x(\tau') F(\tau,\tau') \delta x(\tau)\right] \propto (\det F)^{-1/2} . \tag{2.38}$$

Finally, the saddle-point approximation yields the result

$$K \approx \int \mathcal{D}x(\tau)\, e^{-S_{\mathrm{OM}}[x(\tau)]} \approx (\det F)^{-1/2}\, e^{-S_{\mathrm{OM}}[\bar{x}(\tau)]} \tag{2.39}$$

whereas if many saddle-points $\overline{x}_i$ exist, it can be straightforwardly generalized as

$$K \approx \int \mathcal{D}x(\tau)\, e^{-S_{\mathrm{OM}}[x(\tau)]} \approx \sum_i e^{-S_{\mathrm{OM}}[\bar{x}_i(\tau)]} (\det F_i)^{-1/2} \tag{2.40}$$

In order to understand the last two equations, we can assist our intuition thinking at how high mountains, as the Alps, were historically crossed. The probability of a given path is given by the fraction of people using it in a given time interval. Clearly, most of the people used paths going through the pass, that is a saddle point, i.e., the lowest accessible point on the top of a mountain. Considering a steep and high mountain, with a narrow pass on the top of it, the total probability to cross it is well approximated by the probability of the path going through the pass. A qualitatively similar picture holds also in the highly dimensional conformational space Ω of a protein, if there is a high barrier and a number of clearly distinct non-overlapping saddle-points on the top of it.

In Eq. 2.39 the exponential is calculated on the saddle-point trajectory $\bar{x}_i(\tau)$, which is solution of the ordinary differential Eq. 2.34, and is therefore a smooth differentiable trajectory. However, a stochastic process as diffusion

would never happen through such a trajectory. In fact, in Eq. 2.39 we have to take into account also the determinant of fluctuations 2.38, which measures the volume in a path space of a small bundle of trajectories in the functional vicinity of $\bar{x}_i(\tau)$. The latter trajectory has to be considered as a motif around which stochastic trajectories take place, with most of them contained in the tube due to thermal fluctuations, at least in a low temperature regime [115].

2.3.2 The Dominant Reaction Pathway approach

The most probable microscopic trajectories connecting the unfolded and the folded basins can be in principle found by minimizing the path-dependent part of the OM action functional in Eq. (2.26) for different initial configurations. If the saddle-point approximation holds, each resulting trajectory $\bar{x}_i(\tau)$ is a representative of a set of stochastic trajectories which only differ for the effect of small thermal fluctuations $\sim k_B T$. We get as many tubes of microscopic trajectories as the different initial configurations that we consider. We can now characterize each tube at a more coarse-grained level, by measuring a given set of order parameters, and eventually cluster together all the tubes which are equivalent in this coarse description. Each cluster represents a different pathway, that is a different folding mechanism as we have defined it in Chapter 1.

However, if we directly minimized the action (2.26), we would clash again with the timescale separation that is the reason for which MD simulations are so demanding. Indeed, the path-dependent part of the action would be discretized as

$$\Delta t \sum_{i=1}^{N_t} \left[\frac{(\mathbf{x}(i+1) - \mathbf{x}(i))^2}{4D\Delta t^2} - V[\mathbf{X}(i)] \right] \tag{2.41}$$

where V is the effective potential (Eq. (2.24)). In order to cover a timescale relevant for folding by using a time step $\Delta t \sim$ fs, the total number of steps N_t would be enormous, making any numerical minimization practically infeasible.

A possible way to overcome this severe limitation is given by the fact that Eq. (2.34), with the effective potential (2.24), conserves an effective energy

$$E_{eff} = \frac{1}{4D}\dot{\bar{x}}^2(t) + V[\bar{x}(t)] \ .$$

Thanks to this observation, we can use the Hamilton-Jacobi formulation of classical mechanics, and minimize, instead of Eq. (2.41), an Hamilton-Jacobi effective action

$$S_{HJ} = \sum_{i=1} \Delta l_{i,i+1} \sqrt{\frac{1}{D} \left(E_{eff} + V_{eff}[\mathbf{X}(i)] \right)} \tag{2.42}$$

where time is not the independent variable anymore. Indeed, Eq. (2.42) is still a functional of the path, but the latter is written now in terms of the curvilinear abscissa $\mathrm{d}l = \sqrt{\mathrm{d}\mathbf{x}^2}$, that is the elementary displacement in configuration space. It is possible to use large length steps Δl since there is no separation in length scales in the Euclidean distance l covered during the folding. Thus, it is sufficient to discretize Eq. (2.42) with $\sim \mathcal{O}(100)$ steps, by making a direct numerical relaxation feasible at least for small molecules. This idea has been developed in a series of papers [115–128], and in particular tested on chemical reactions [116], conformational changes [117, 121] of small molecules, and protein folding in simplified CG Gō models [122, 124]. The results of these tests and investigations are in good agreement with standard MD simulations, which are feasible since the systems are relatively simple and small, although obtained at a much lower computational cost.

In the natural following step, we applied the outlined minimization scheme to a small protein in AA resolution with realistic FF. The procedure requires to produce some initial trajectory, which can be obtained for example by an MD simulation at high temperature that unfolds the protein rather quickly. Then a minimization algorithm is applied on the target functional (2.42) calculated on the initial trajectory. However, the minimization of Eq. (2.42) is a hard task, since it can assume complex values and depends on all the degrees of freedom of a folding trajectory. For a small protein of $\sim 10^3$ atoms, considering 10^2 displacement steps Δl, the degrees of freedom entering the functional are $\sim 3 \times 10^5$. We have carried on several attempts to obtain folding trajectories of a realistic protein, but unfortunately the problem turned out to be intractable. Even by using state-of-the-art minimization algorithms, like e.g. [129], the initial trajectory and the minimized one are strongly correlated, almost identical, differing for little local relaxations only. Presumably the exceedingly high number of degrees of freedom and the complexity and roughness of realistic FF cause the initial trajectory to remain trapped in local minima, thus making an efficient sampling of the path space practically impossible .

2.3.3 Biased sampling of the path space

A direct numerical relaxation of the target functional (2.42) seems impossible, and an alternative strategy has to be envisaged. The solution we propose requires to change the point of view of the problem: instead of minimizing the action to get a trajectory with two fixed endpoints, let us consider many trajectories connecting the given initial and final configurations and then score them according to their value of the action. Of course, if we want to characterize the folding in AA resolution and in realistic FF we cannot use a standard MD simulation, since this is often infeasible, and we still want to avoid to consume computational resources to simulate thermal fluctuations in the DS. Moreover, in order to significantly sample the functional trajectory space, we prefer to use a rather inexpensive algorithm, which yields a significant number of folding trajectories in a reasonable computational time.

For these reasons, we adopted a biased MD algorithm known as the ratchet-and-pawl Molecular Dynamics (rMD) [130–136]. This computational scheme can be easily implemented upon any standard MD simulation using any FF, and permits to efficiently obtain a large number of trajectories going from an initial to a final configuration. It consists of two parts, the functional form of the biasing potential, and the coordinate which sets the direction along which the system is biased. The functional form was originally proposed by Marchi and Ballone [130] and used by Paci and Karplus to bias the protein unfolding in implicit solvent [131, 132]. Camilloni and Tiana then used it to fold proteins in realistic FF and explicit solvent [134], by biasing along a particular distance introduced in Ref. [133], which is built by considering the native contacts formation in a protein. In the following, we will present and use the rMD formulation proposed by Tiana and Camilloni [134].

Following Ref. [133], we shall define a protein contact map $C_{ij}\left[\mathbf{x}\left(t\right)\right]$ of a trajectory, which at each time step is a matrix $N_a \times N_a$ with entries calculated according to

$$C_{ij}\left[\mathbf{x}\left(t\right)\right] = \begin{cases} \frac{1-\left(\frac{r_{ij}}{r_0}\right)^6}{1-\left(\frac{r_{ij}}{r_0}\right)^{10}} & r_{ij} < r_{cut} \\ 0 & r_{ij} > r_{cut} \end{cases} \tag{2.43}$$

where r_0, r_{cut} are constant values, and r_{ij} is the Euclidean distance between atoms i and j in the conformation $\mathbf{x}\left(t\right)$. Note that the entries are calculated for each couples of atoms in the protein. Eq. (2.43) is a smooth interpolation

of the step function which is usually used to calculate a contact map. Indeed, if two atoms are close and form a contact, i.e., $r_{ij} < r_0$, then $C_{ij} \approx 1$; in the other case, i.e., $r_0 < r_{ij} < r_{cut}$, $C_{ij} \approx 0$. The native contact map $C_{ij}(\mathbf{x}_{\text{nat}})$ is simply obtained evaluating Eq. (2.43) on native conformation. We can now define a distance $z[\mathbf{x}(t)]$ separating a configuration at time t from the native one [133]

$$z[\mathbf{x}(t)] = \sum_{i>(j+35)}^{N} [C_{ij}[\mathbf{x}(t)] - C_{ij}(\mathbf{x}_{\text{nat}})]^2 \tag{2.44}$$

where the summation is extended over all atoms with a separation in sequence > 35. The value of this distance is ≈ 0 if the protein has reached its native state. Indeed, any time a native contact (i, j) forms, then $C_{ij}[\mathbf{x}(t)] - C_{ij}(\mathbf{x}_{\text{nat}}) \approx 0$ and $z[\mathbf{x}(t)]$ diminishes. If a native contact has not formed yet, or a non-native contact forms, then $[C_{ij}[\mathbf{x}(t)] - C_{ij}(\mathbf{x}_{\text{nat}})]^2 \approx 1$ and the value of $z[\mathbf{x}(t)]$ increases. Eq. (2.44) thus measures the geometrical and topological similarity of a configuration with the native one.

The biasing potential, instead, is a time dependent harmonic potential defined as

$$V_{\text{rMD}}(\mathbf{x}, t) = \begin{cases} k(z[\mathbf{x}(t)] - z_{\min}(t))^2 & z[\mathbf{x}(t)] > z_{\min}(t) \\ 0 & \text{otherwise} \end{cases} \tag{2.45}$$

characterized by

$$z_{\min}(t) = \min_{t'<t} z[\mathbf{x}(t')]$$

that is the lowest reached value of the distance in the contact map space up to time t, and the coupling constant k, which has the dimension of an energy. The potential (2.45) has then just to be added to the energy function of the particular FF that are used to perform an MD simulation. The value of k is a result of a trade-off between having very fast simulations and keeping the bias as soft as possible. Usually a reasonable choice is to have a biasing force that is two or three orders of magnitude smaller than the typical forces acting on the protein.

In a rMD simulation, the system is let to fluctuate spontaneously whenever it decreases the distance (2.44) to the native configuration, whereas the time-dependent harmonic potential (2.45) hinders fluctuations decreasing the overall similarity with the native state. In other words, the system is free to follow its spontaneous dynamics, while a Maxwell's demon uses a spring

to select only the fluctuations with the right "sign", s.t. z is decreased. In this sense rMD is different than a steered MD simulation [137], where the molecule is pulled with a constant force or velocity along a given biasing coordinate.

On the other hand, it is well known that biasing along a "wrong coordinate" yields highly artificial trajectories. To understand this intuitively, we can think at a bi-dimensional free energy landscape, displaying two deep wells which are separated by a barrier. By definition, the most likely trajectories connecting the two wells are located along the reaction coordinate of the system. If we biased along the direction perpendicular to the reaction coordinate, then the system would never cross the barrier and stay stuck in the reactant state. This is the worst case, but even by biasing along a direction which forms an angle with the reaction coordinate, we would force the molecule to visit unlikely high free energy conformations, and the resulting trajectories would not be representative of the spontaneous transition. Therefore, we understand that it is of utmost importance to bias along a direction close enough to the system's reaction coordinate. The great advantage of the rMD scheme as proposed by Camilloni and Tiana [134] is that the biasing potential (2.45) acts on the distance in the contact map (2.44), which is very similar to the fraction of native contacts. Although there is no systematic investigation on how this two quantities are quantitatively related, *we shall assume as a working hypothesis that the distance in the contact map (2.44) is a satisfactory reaction coordinate for the folding of small and globular proteins.*

As a matter of fact, Camilloni, Tiana and a Beccara have carried out several attempts to fold proteins by using the potential (2.45) acting along diverse geometrical order parameters, but despite high values of the biasing force, trajectories remain stuck in non-native conformations and folding is never observed.

The rMD biasing scheme has proved to be very efficient in sampling the folding trajectory space [134]. Folding trajectories are obtained at a rather cheap computational cost, since fluctuations are filtered out and meta-stable states are removed. This high efficiency comes with a high cost, since time intervals measured along the trajectories are highly unphysical, and thus both kinetics and thermodynamics are disrupted. This is a severe drawback indeed, and we cannot directly compute rates and free energies, which are usually employed to compare experimental results.

2.3.4 Sampling and scoring

We have now and instrument, the rMD algorithm, to produce many protein folding trajectories connecting an initial unfolded configuration to a folded one. These are the results of a bias, pushing the system along a direction that we assume to be related to the real reaction coordinate of the system. Nonetheless, this is only approximately true, and the bias will always introduce spurious effects in the trajectories. In order to soften them, we can rank the biased folding trajectories according to the value of the OM action functional (2.26). If we consider a rMD folding trajectory $x(\tau)$, then $\exp\{-S_{\mathrm{OM}}[x(\tau)]\}$ is approximately proportional to the probability for that trajectory to happen in the *unbiased* diffusive dynamics. We score a set of rMD trajectories with fixed endpoints and select the one which minimizes S_{OM}, which is thus the one with highest weight in the unbiased dynamics. In other words, we produce a set of reasonable folding "trial" trajectories, score them with an unbiased weight, and pick up the best one.

2.3.4.1 Characterizing the folding pathways: the algorithm

We are now ready to sketch the algorithm we will test and use to portray the folding of realistic protein models.

1. Provide the protein native structure at atomistic resolution.

2. Produce as many as possible denatured conformations, e.g. by doing a short high-temperature MD followed by an unbiased MD at environmental temperature to relax the structure. Denatured conformations should be selected according to the equilibrium Boltzmann distribution.

3. For each unfolded configuration, produce as many as possible folding trajectories running rMD simulations, which we shall call *trial trajectories*. Any realistic FF can be used, although in implicit solvent only.

4. Calculate the value of the OM action for each of the successful rMD trajectories, retain only the one displaying the lowest value, hence the less biased one, which we shall call a *dominant folding trajectory* and discard all the others. This scoring and selecting procedure returns our best guess for the folding trajectory connecting fixed initial and final configurations.

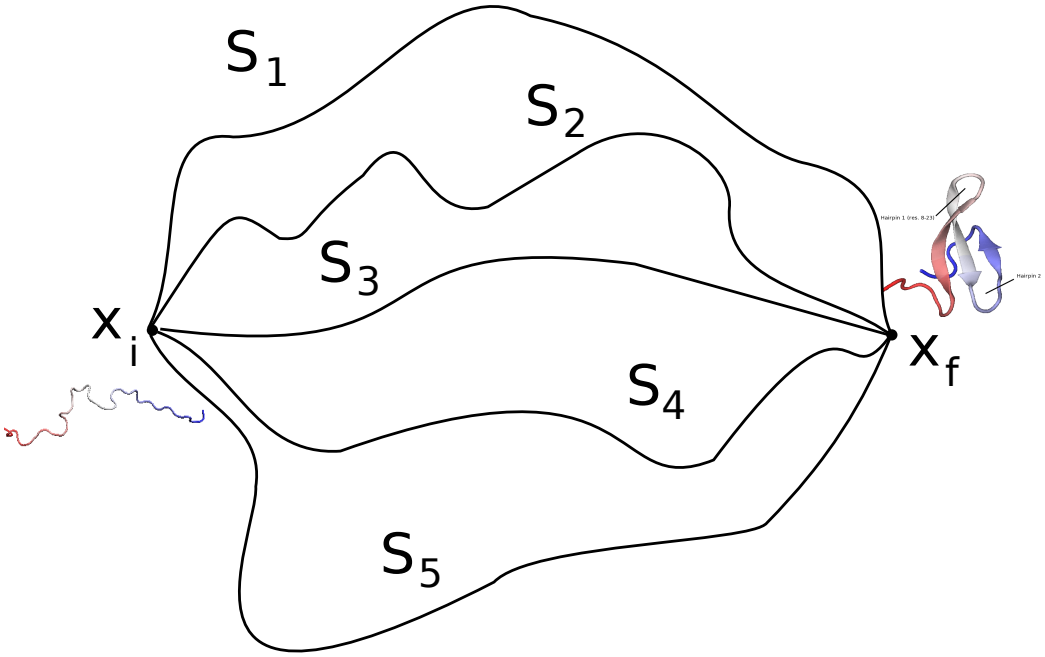


Figure 2.3: For given initial denatured and final folded configurations we produce many folding trial trajectories by meas of biased rMD simulations. We score the trajectories according to their probability in an unbiased diffusive dynamics, and select the less biased one.

5. Characterize all the trajectories according to order parameters, by clustering in a *dominant reaction pathway* those which share the same folding mechanisms.

Chapter 3

Folding a WW Domain

In this chapter we investigate the folding mechanism of the WW domain Fip35 using a realistic atomistic FF, by applying the DRP algorithm introduced in the last section of chapter 2. In the first section we will show evidence for the existence of only two folding pathways, which differ by the order of formation of the two hairpins composing the WW Domain. We will then show in the second section how this result is consistent with the analysis of the experimental data on the folding kinetics of very similar WW domains. Then, the we will compare our results with those obtained from ultra-long equilibrium MD simulations of this system performed on the Anton supercomputer.

We will show how free energy calculations performed in two CG models support the robustness of the two pathways picture. Moreover, turning on and off non-native interactions, we will find evidence that the qualitative structure of the folding pathways is mostly shaped by the native interactions and the chain topology.

This chapter is based on the original research paper of Ref. [138]:

- S. a Beccara, T. Škrbić, R. Covino, and P. Faccioli, PNAS 109 (2012)

3.1 Folding pathways of a WW Domain

Ultrafast folding proteins are a perfect target for any MD based approach, and indeed are often used as a benchmark system. The Fip35 WW Domain

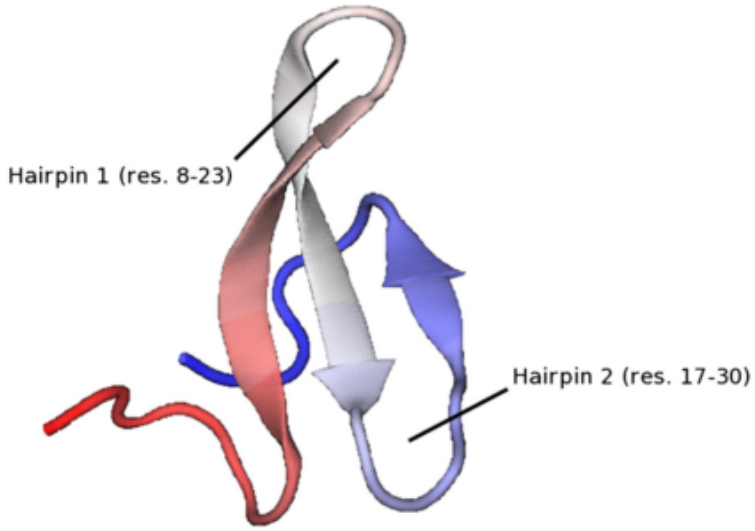


Figure 3.1: Native structure of Fip35 WW Domain [139], mutant of protein human pin1 (pdb code: pin1). The primary sequence of Fip35 is: EEKLPPGWEKRMSADGRVYYFNHITNASQWERPSG. Fig. reproduced with permission from Ref. [138]

is a engineered mutant of the human Pin 1 protein, which folds in only 14 μs, being the fastest folding WW domain [139]. It is only 35-residues long, and displays a simple topology, made of two $\beta-$hairpins of different length sharing a common strand. The first hairpin (H1) spans over the residues 8-23, and the second hairpin (H2), spans over residues 17-30.

Fip35 and other very similar WW domains have been widely characterized both experimentally [139–141] and numerically [75, 142–157]. In particular, the first AA MD simulation on the millisecond scale performed on Anton has shown reversibly folding and unfolding [69]. In this first paper, Shaw *et al.* described only one folding mechanism, where H1 folds first, followed then by H2. On the other hand, Ensign *et al.* found a completely heterogeneous folding by means of a Markov-state-model analysis of short out-of-equilibrium trajectories [142].

3.1.1 Two folding pathways

In order to investigate the folding mechanism of the Fip35 WW domain, we produced several dominant folding trajectories by means of the DRP algorithm. In Fig. 3.2 we show our set of atomistic dominant folding trajectories, projected onto the plane defined by the Root-Mean-Square-Deviation to native (RMSD) [70] of the C_α-atoms in H1 and H2. We can clearly identify two distinct folding pathways, which differ by the order of formation of the hairpins: in about half of the computed dominant folding trajectories H1 consistently folds before H2 (Left part of Fig. 3.3), while in about the other half, we observe that the two hairpins form in the reversed order (Right part of Fig. 3.3).

It is noteworthy that not all the rMD trial trajectories computed starting from a given initial condition follow one of the two folding pathways discussed above. Indeed, as shown in Fig. 3.4, many of them involve a simultaneous formation of native contacts in both hairpins. These are systematically excluded when scoring and selecting the dominant trajectories according to the lowest OM action criterion. We can thus say that folding events in which the hairpins form simultaneously are much less frequent than those in which the two secondary structures form in sequence.

Another result emerging from our simulations is the existence of a correlation between the structure of the initial configurations from which the transition is initiated and the pathway taken to fold. If at the beginning of

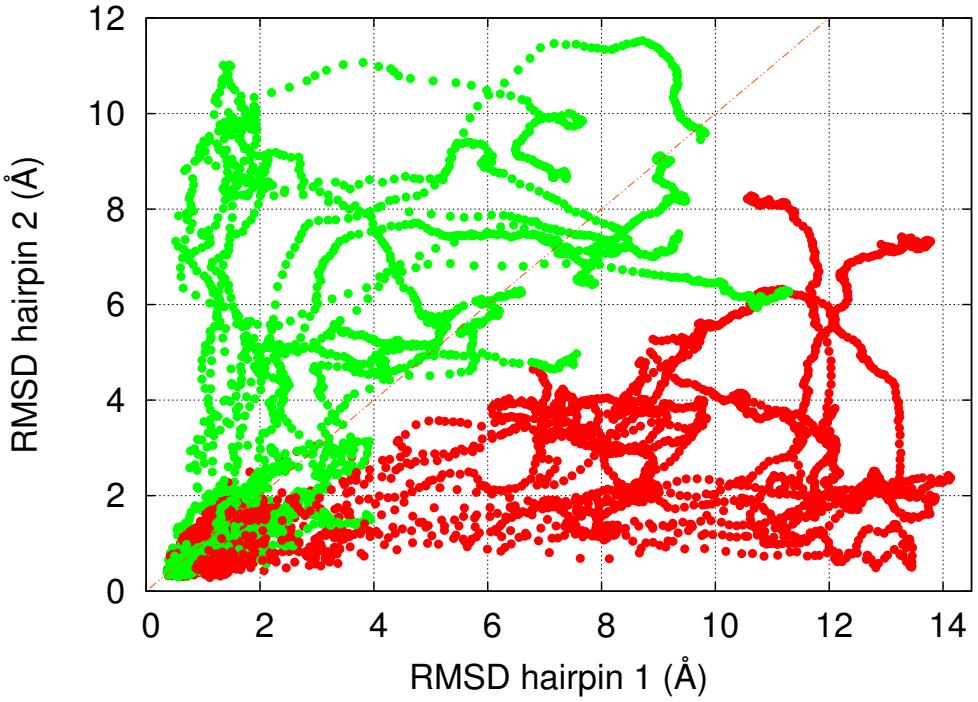


Figure 3.2: The set of dominant folding trajectories for Fip35, obtained from atomistic DRP simulations, projected on the plane defined by the RMSD of the two hairpins to the corresponding native structures. Fig. reproduced with permission from Ref. [138].

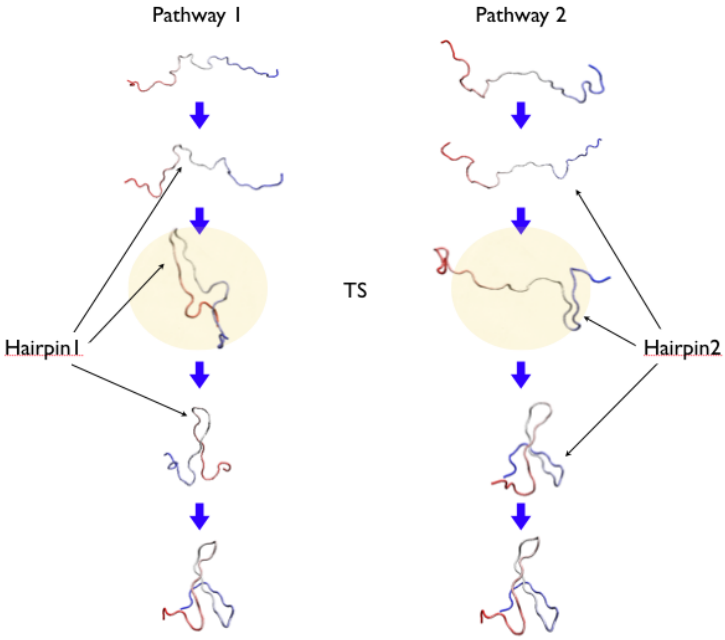


Figure 3.3: Schematic representation of the structure of the two dominant folding pathways obtained in our simulations. Fig. reproduced with permission from Ref. [138]

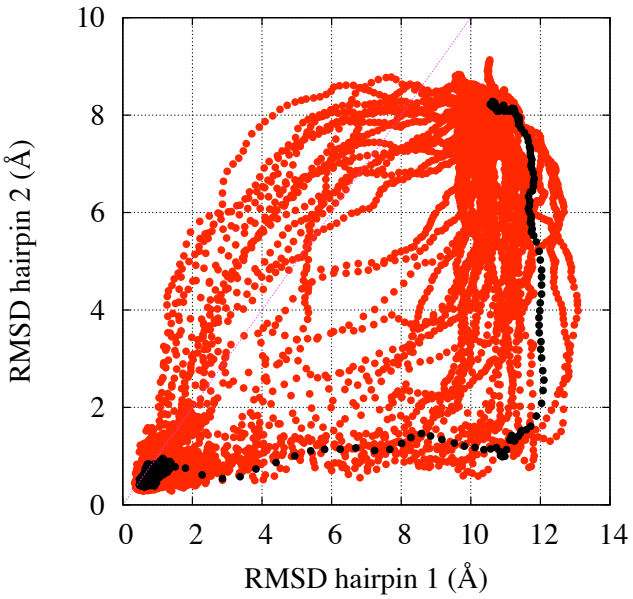


Figure 3.4: The set of trial trajectories connecting a given denatured configuration, on the upper right corner of the plot, to the NS, which is in the lower left corner, used in the search for the dominant trajectory, projected on the plane defined by the RMSD to native of the two hairpins. The darker path is the selected dominant reaction trajectory. Fig. reproduced with permission from Ref. [138].

the transition H1 has a RMSD smaller than H2, then the first pathway is most likely chosen. In the opposite case, i.e., when H2 has a smaller RMSD to native than H1, then the second pathway is generally preferred.

3.1.2 Little role for non-native interactions

In order to further support these results and gain insight into the folding mechanism, we have performed simulations in a native centric Gō-like simplified model, computing equilibrium properties using the CG models described in Section 3.4. In Fig. 3.5 we show the free energy landscape at the 300 K, as a function of the RMSD to native of the two hairpins for the two models. The upper panel considers only stabilizing native contacts, whereas the lower panel was calculated by considering also attractive non-native interactions. In both cases, we observe the existence of two valleys in the free energy landscape, which correspond to the two folding pathways discussed above.

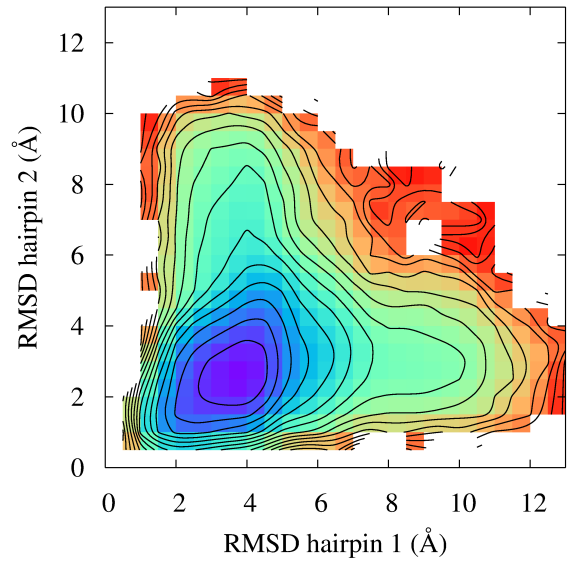
The fact that the two free energy landscapes in Fig. 3.5 are remarkably similar suggests that non-native interactions have a vanishing role in the folding the Fip35 WW Domain.

3.1.3 Locating the TS

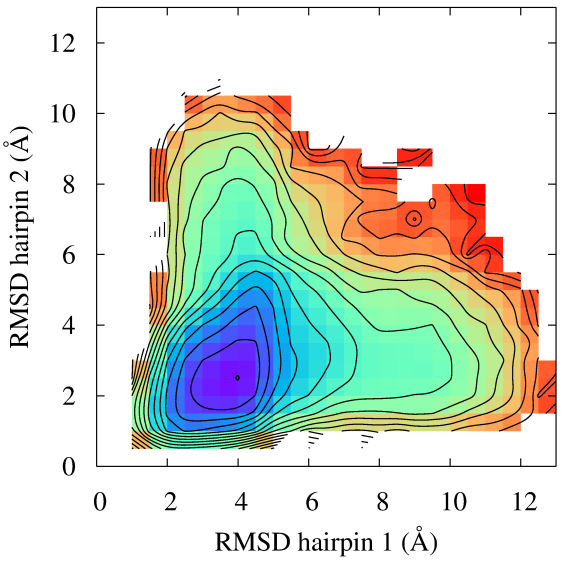
According to a widely used definition, the TS is the set of configurations $\mathbf{x}_{\mathrm{TS}}$ such that the probability to reach the NS is equal to that of going back to the denatured configuration [121]. In order to locate the TS we consider the approximation

$$\frac{\mathscr{P}\left(\mathbf{x}_{\mathrm{TS}} \to \mathbf{x}_i\right)}{\mathscr{P}\left(\mathbf{x}_{\mathrm{TS}} \to \mathbf{x}_{\mathrm{f}}\right)} \approx \frac{e^{-S_{\mathrm{OM}}(\mathbf{x}_{\mathrm{TS}} \to \mathbf{x}_N)}}{e^{-S_{\mathrm{OM}}(\mathbf{x}_{\mathrm{TS}} \to \mathbf{x}_D)}} = 1$$

where $\mathbf{x}_N$ and $\mathbf{x}_D$ are the first native and denatured configurations visited along the dominant trajectory, starting from $\mathbf{x}_{\mathrm{TS}}$. We thus only take into account the "reactive" part of the trajectory, that is the one which leaves the DS and, without recrossing, goes straight to the NS. To satisfy this requirement, we considered the total RMSD *vs.* frame index curve. The typical trend of this curve for most of the dominant trajectories is shown in Fig. 3.6. It consists in an initial plateau, followed by a rather steep fall, and then by another flat region, where the system oscillates in the NS. The reactive part of the path was identified with the region of steep fall in this



(a)



(b)

Figure 3.5: The free energy surface at $T = 300$ K as a function of the RMSD to native of the two hairpins, obtained from the MC simulations in two CG models, described in Section 3.4. In the upper panel the model accounts for native interactions only [158], whereas in the lower panel both native and non-native attractive interactions are considered [159]. The free energy as a function of the RMSD of the two hairpins (potential of mean force) was obtained from the frequency histogram calculated from long MC trajectories. Fig. reproduced with permission from Ref. [138].

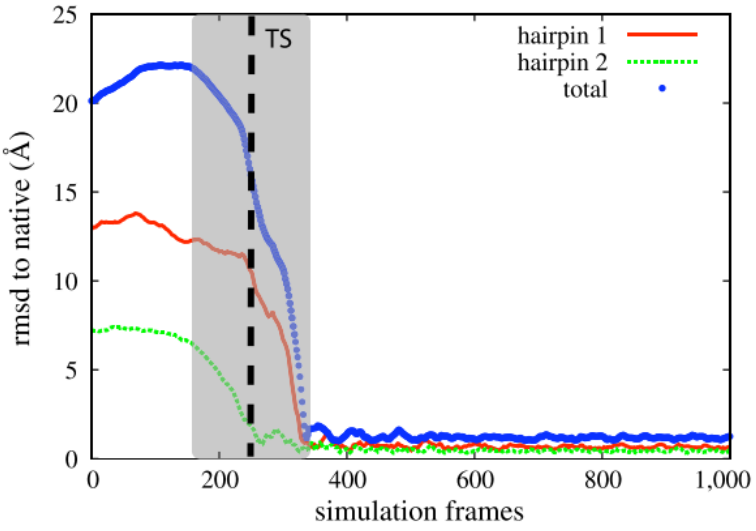


Figure 3.6: The typical evolution of the RMSD in the second folding pathway is shown. H2 (green) folds first followed by H1 (red). The gray shaded area corresponds to the reactive part considered to calculate the TS, which is represented with the dashed black line. Fig. reproduced with permission from Ref. [138].

curve. In particular, the beginning of the reaction was set to the frame at which the derivative of the total RMSD curve changes sign, from positive to negative.

Following this prescription we are able to qualitatively locate the TS, which is found at the "turn" of the pathways; i.e., is formed by configurations in which H1 is folded while H2 is largely unstructured in one pathway, and by configurations where the opposite is true in the other (see Fig. 3.3).

3.1.4 Relative weight of the pathways

Trajectories simulated by means of the rMD display unphysical time intervals and are far from equilibrium. Thus, it is not possible to directly calculate kinetic quantities as the folding rate, or thermodynamics as free energy differences. However, in order to be able to compare our results with experimental investigations, we have to provide an at least qualitative estimate of the fold-

ing rate. We therefore decide to consider the following approximation

$$\frac{k_1}{k_2} \approx e^{-\beta\left(G_1^{\ddagger}-G_2^{\ddagger}\right)}, \tag{3.1}$$

where k_1 and k_2 are the folding rates, $G_1^{\ddagger}$ and $G_2^{\ddagger}$ are the free energies of the TS's of the first and second folding pathways respectively. Eq. (3.1) expresses the ratio between the probabilities to see a folding event along one of the two pathways as a free energy difference. Such an expression is based on the assumption that each folding pathway is a thermally activated process. The free energy barriers corresponds to the two different TS's located along the dominant folding trajectories, where, as discussed in section 3.1.3, either H1 or H2 are formed. We then project the location of the TS's obtained on the dominant folding trajectories on the free energy landscapes represented in Fig. 3.5, and measure $G_1^{\ddagger} - G_2^{\ddagger}$. In this way we find that the probability of the pathway where H1 forms first is $\sim 70\%$, whereas the probability of the second pathways is $\sim 30\%$. We stress that this result has to be considered only as semi-quantitative, but nonetheless it opens the door to a comparison with other kind of studies of the same system.

Furthermore, this simple scheme enables us to address the question of the dependence on temperature of the relative weights of the two channels. Repeating the calculation at a higher temperature of 380 K, assuming that the structure of the TS's is not significantly modified, we find $k_1/k_2 \approx 1.6$, which corresponds to a branching ratio of the first pathway of about 60%. Hence, the rate limiting role of the second channel grows with temperature, and this can be understood as follows. The folding of one of the hairpins generates an entropy loss proportional to the number of native contacts formed. The TS in the first folding channel involves forming a longer hairpin, namely H1, hence reaching it produces a larger entropy loss (but also larger gain of native energy). The role of the entropy loss relative to the energy gain in forming the hairpins grows with temperature, hence disfavoring the first folding channel compared to the second.

3.1.5 Varying the force

The rMD simulations depend on an external parameter k, which sets the strength of the biasing force. At very low values of k, rMD trajectories are minimally biased, and the system does not perform any folding transition in

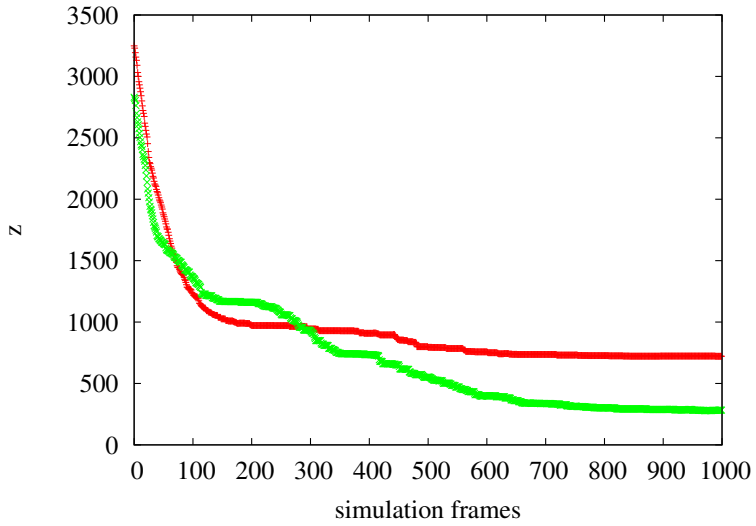


Figure 3.7: Typical evolution of the ratcheting coordinate $z(t)$ in two rMD simulations. Fig. reproduced with permission from Ref. [138]

a typical simulation time. Thus, there is an insufficient gain in this regime in using this biasing scheme, compared to an unbiased MD simulation.

In the opposite high k limit, the dynamics is affected by a significant bias since the external force becomes comparable with the physical internal forces acting on the atoms. In this regime, if the ratcheting coordinate z is a bad reaction coordinate, the system is driven into large free energy regions. The unbiased statistical weight given by the exponent of the OM action is expected to penalize these trajectories.

Fig. 3.7 shows the evolution of the biasing coordinate z, in two typical folding rMD trial trajectories.

It is important to study to what extent a given folding trajectory depends on the specific value of the bias constant k adopted in rMD simulations. In Fig. 3.8 we plot the dominant reaction trajectories obtained starting from the same initial condition, using different values of k, which span over almost two orders of magnitude. We see that in most simulations the folding is described by the same qualitative mechanism, in which H1 forms before H2. Only in one case, for a low value of the coupling constant k, we find that the protein travels across the DS before taking a different pathway, in which the order of formation of the hairpins is reversed.

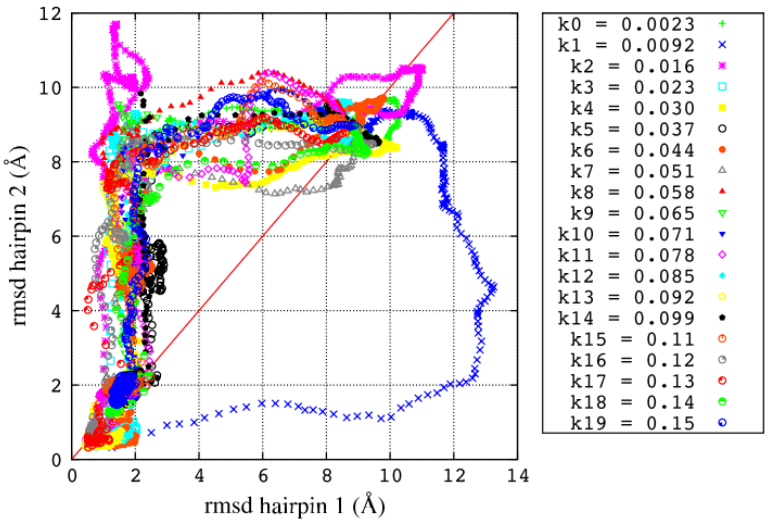


Figure 3.8: Dominant trajectory connecting the same initial and final configuration calculated with a rMD coupling constant k varying over two orders of magnitude. Fig. reproduced with permission from Ref. [138].

3.2 Comparison with experiments

The so-called $\phi-$values analysis represents the main experimental technique that is able to characterize the TS [15, 160, 161]. The latter is by definition made of transient short-lived configurations, which cannot be resolved by standard approaches as X-ray crystallography and NMR.

Let us consider a small one-domain two-state folding protein. If we mutate a residues in the wild-type protein we can measure the $\phi-$value defined as

$$\phi = \frac{\left(\Delta G^{\ddagger-\mathrm{DS}}_{\mathrm{wt}} - \Delta G^{\ddagger-\mathrm{DS}}_{\mathrm{m}}\right)}{\left(\Delta G^{\mathrm{NS}-\mathrm{DS}}_{\mathrm{wt}} - \Delta G^{\mathrm{NS}-\mathrm{DS}}_{\mathrm{m}}\right)} = \frac{\Delta\Delta G^{\ddagger-\mathrm{DS}}}{\Delta\Delta G^{\mathrm{NS}-\mathrm{DS}}} \tag{3.2}$$

where $\Delta G^{\ddagger-\mathrm{D}}$ is the folding free energy barrier, which separates the DS from the TS, calculated for the wild-type (wt) and the mutant (m). $\Delta G^{\mathrm{NS}-\mathrm{DS}}$ is instead the free energy difference between the NS from the DS, i.e., the stability, again calculated for the wild-type and the mutant. Thus, a $\phi-$value measures to which extend a given mutation of a residue perturbs the TS compared to the NS.

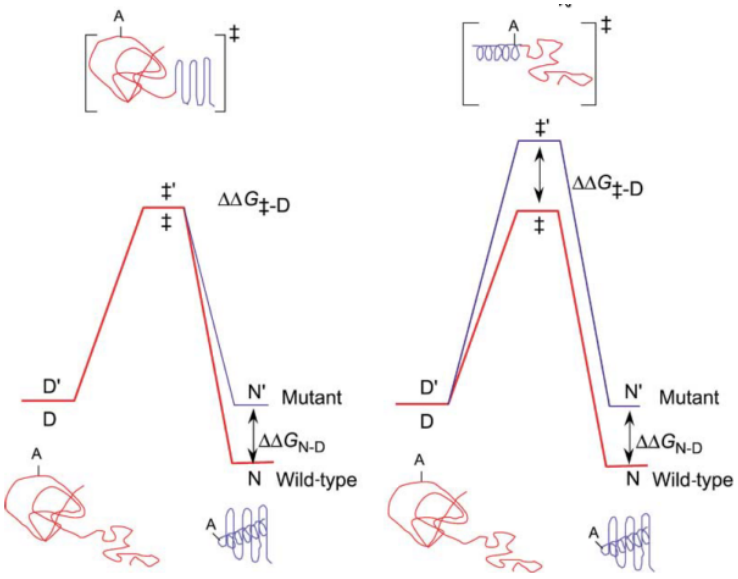


Figure 3.9: Schematic representation of how $\phi-$values are used to characterize the TS. Fig. reproduced with permission from Ref. [15].

To interpret the resulting numbers, a series of hypothesis is usually considered: that there is a strict correlation between the interactions in the TS and its structure; that stabilizing interactions in the TS are only native-like; that a mutation can only destabilize the TS or the NS, thus excluding a possible stabilizing effect. If these key assumptions hold, then a $\phi-$value measures the amount of native structure formed in the chain around the mutated residue (Fig. 3.9 left panel). If $\phi = 0$, then the NS is destabilized but the TS is not, and this means that the mutated residue is still unstructured in the TS. On the contrary, if $\phi = 1$, then the TS and the NS have been destabilized in the same way, thus a native structure is already formed in the TS in the location of the mutated residue (Fig. 3.9 right panel).

Non-canonical $\phi-$values, which are neither 0 nor 1, cannot be interpreted unequivocally. Negative values can be caused by a stabilizing effect of the mutation either on the NS or the TS. On the other hand, values between 0 and 1 may be caused by an heterogeneous TS, which is usually interpreted as the existence of more folding pathways. In this sense, the $\phi-$values analysis is also the main experimental instrument to investigate the folding mechanism.

Jäger *et al.* extensively measured the $\phi-$values on different temperatures

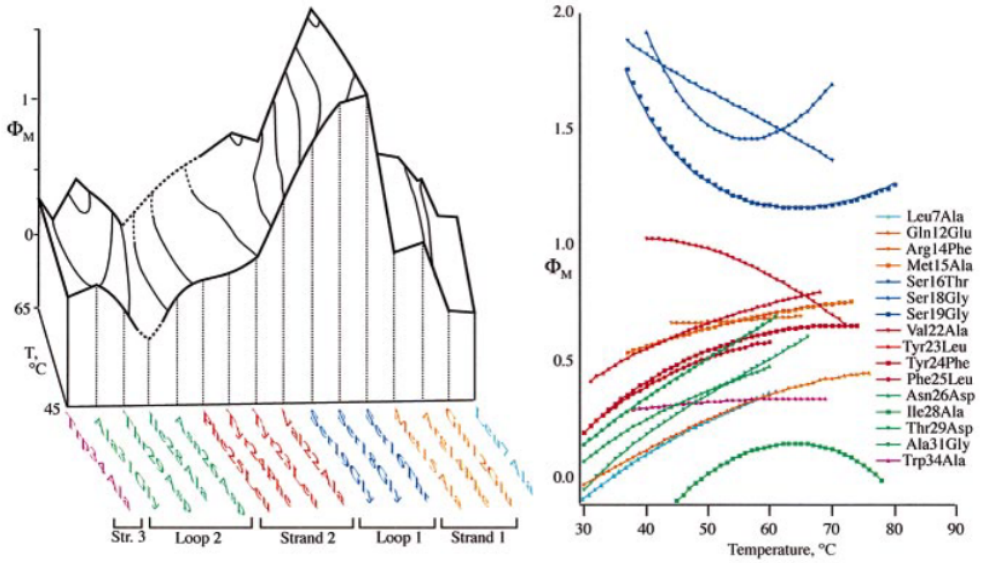


Figure 3.10: $\phi-$values analysis of the human Pin 1 WW Domain. Left panel: $\phi-$values are represented on the vertical axis, while the location of the mutated residue is on the horizontal one. The third dimension represents the temperature at which the folding has been characterized. H1 corresponds to the region including Strand 1, Loop 1 and Strand 2. H2 is composed of Strand 2, Loop 2 and Strand 3. Right panel: same data represented in a $\phi-$values vs. temperature plane to clearer show the temperature dependence. Fig. reproduced with permission from Ref. [140].

for the human Pin 1 WW Domain, the wild-type precursors of Fip35 [140]. Their result, reported here in Fig. 3.10, shows a majority of non-canonical $\phi-$values, which can be interpreted only qualitatively. The TS is heterogeneous, with most of formed structures located on H1, while on H2 a smaller amount is also present, which notably increases with rising temperatures. This picture can be explained invoking two possible folding pathways where the hairpins form in a different order, and is thus compatible with our findings, but we have to use a model to quantitatively interpret the experimental data of Fig. 3.10.

Weikl provided such a model and the related interpretation in a series of papers [151, 162, 163]. The core assumption is inspired by the foldon picture, since H1 and H2 are considered cooperative folding units that can be either folded or unfolded in the TS. Thus, a mutation can affect (stabilizing or destabilizing) one of the two hairpins, and Eq. (3.2) can be accordingly recast

$$\phi = \frac{\rho_1 \Delta\Delta G_1^{\ddagger-\mathrm{DS}} + \rho_2 \Delta\Delta G_2^{\ddagger-\mathrm{DS}}}{\Delta\Delta G^{\mathrm{NS}-\mathrm{DS}}},$$

where ρ_1 is the probability, or fraction, of the TS conformations in which H1 is formed, and $\rho_2 = 1 - \rho_1$ is the probability of the TS conformation with H2 formed. This model displays a single free parameter, and it is enough to fit it to the experimental data to satisfactorily explain the $\phi-$values in Fig. (3.10). The fitting procedure yields the probability of each pathway, with $\rho_1 = 0.69 \pm 0.05$ and $\rho_2 = 0.36 \pm 0.05$. The pathway in which H1 forms first is about twice more populated than the other one, where instead H2 forms first.

This results provides a solid interpretation of the data shown in Fig. (3.10) and is in good agreement with the semi-quantitative estimate calculated in section 3.1.4 on our folding trajectories.

3.3 Comparison with numerical investigations

In the last decade, the folding of WW Domains has been investigated extensively by considering different models and FF's. Some studies reported the coexistence of the two folding pathways where hairpins form sequentially [144, 146, 152, 155, 164–166], whereas other described a more heterogeneous picture displaying many possible folding mechanisms [142, 147–149].

Fip35 has been one of the first proteins investigated by means of millisecond long MD simulations performed on Anton. In a first paper, Shaw *et al.* reported the existence of only one dominant folding mechanism, where H1 folds before H2 [69]. An independent analysis of the folding trajectories [153], and further simulations by the same group [75], revealed also the presence of the less frequent second pathway, where the order of formation of the hairpins is inverted. In particular, Krivov estimated the probability of each reaction channel, finding 80% for the pathway where H1 forms first, and 20% for the second one [153]. Moreover, Krivov built the free energy landscape surface projected on two optimized coordinates, which are obtained from the native contacts in each hairpin. He found that the TS in each pathway displays conformations where the tip of the corresponding hairpin is formed [153]. Berezovska *et al.* found a qualitatively similar picture by using a Markov-state-model [154].

Noteworthy, it is still debated whether Fip35 is a "downhill folder" or a standard two-state folder. In the first case there is no thermally activated barrier separating the DS from the NS, and folding time is due presumably to diffusion on a rough energy landscape [69, 139, 141], whereas in the second case, instead, a clear barrier exist [153, 154].

To summarize, the result of our investigation by means of the DRP algorithm shows that the Fip35 WW Domain folds along two different pathways, defined by the order of formation of the two main structural elements, H1 and H2. This finding is in agreement with the interpretation of experimental ϕ-values data and long unbiased MD simulations, as well as with alternative methods.

Remarkably, simulating the set of folding trajectories we have studied in this chapter took only two days of calculations on 48 CPU's.

3.4 Computational details

3.4.1 Atomistic DRP simulations

The AA DRP calculations were performed using DOLOMIT, a home written code, which calls a librarized version of GROMACS 4.5.2 [167], in order to calculate the molecular potential energy and its gradient. We employed the AMBER ff99SB FF [56] in implicit solvent with Generalized Born formalism.

The Born radii were calculated according to the Onufriev-Bashford-Case algorithm [54].

We defined the NS as the set of conformations with a RMSD to the crystal structure of the C_α in the hairpins smaller than 3.5 Å. A configuration was considered denatured if the RMSD to native of both hairpins was larger than 6 Å. The stability of the NS within the present FF was checked by running 12 unbiased 2 ns long MD simulations at the room temperature (300 K). In all such trajectories the protein remained in the NS.

We then generated 44 independent initial conditions, by running a 50 ps MD at 1,600 K, starting from the energy minimized NS, followed by a 100 ps relaxation at 300 K. The time step employed in all the simulations was 1 fs. From 24 starting configurations we computed 96 trial trajectories each consisting of 50,000 rMD steps. For each of the remaining 20 initial conditions, the number of trial paths was limited to 48. In such rMD simulations, the ratchet spring constant was set to $k = 0.02$ kcal/mol. Using this value, the modulus of the biasing force was always found to be at least one order of magnitude smaller than the modulus of the total force acting on the system. We observed that 5 of the 44 initial conditions did not correspond to DS, hence they were rejected. In addition, in 13 of the remaining 39 sets, more than 80% of the trial trajectories did not reach the native state within the simulation time. In these cases the exploration of the path space was limited to very few trial trajectories, so the corresponding dominant trajectories were discarded. For the remaining 26 sets of trajectories, we identified the most probable by computing the OM action.

3.4.2 CG native-centric calculations

To study the equilibrium properties of the folding of the Fip35WW domain we used the CG model recently developed in Ref.'s [106, 159]. In that model, amino acids are represented by spherical beads centered at the C_α positions. The non-bonded part of the potential energy contains both native and non-native interactions. The former are the same used in the Gō-type model of Ref. [158], while the latter consist of a quasi-chemical potential, which accounts for the statistical propensity of different amino acids to be found in contact in native structures, and of a Debye-screened electrostatic term. In this model, the average potential energy due to native interactions in the folded phase is typically one order of magnitude larger than that due

to non-native interactions. Above the folding temperature, this ratio drops to about four. This model was shown to provide an accurate description of protein-protein complexes with low and intermediate binding affinities [159]. We calculated the specific heat, evaluated from MC simulations at different temperatures, which indicates that this model yields the correct folding temperature for this WW domain [138].

The simulations were performed using a MC algorithm based on a combination of Cartesian, crankshaft and pivot moves. Details can be found in Ref. [168].

Chapter 4

Folding a knotted protein

The existence of proteins which spontaneously fold to a self-tied conformation of their backbone is a surprising fact that has been discovered and investigated during the last decade. Unraveling the subtle interplay that is required between all the interactions in order to fold and avoid any sort of traps is still an open issue. Simulating and understanding the folding mechanism of knotted proteins is a challenge and a testing ground for any computational approach.

In the first section we will briefly review what is currently known about knots in protein. In particular, we will discuss how to operatively define them and what we learnt exploring the knotting process by using both experimental and computational methods.

We will completely devote the second section to describe our original results, obtained by investigating the folding of the smallest known knotted protein by means of the DRP computational scheme. We will explain how our findings point out a crucial role for non-native interactions in determining the probability and mechanism of folding in the little knotted protein.

This chapter is based on the original research paper of Ref. [169]:

- S. a Beccara, T. Škrbić, R. Covino, C. Micheletti, and P. Faccioli, PLoS Computational Biology 9 (2013)

and the perspective article of Ref. [170]:

- R. Covino, T. Škrbić, S. a Beccara, P. Faccioli and C. Micheletti, Biomolecules 4, 1 (2013).

Furthermore, in writing this chapter I took great advantage of fruitful discussions with Patrícia Faísca and Joanna Sułkowska, whom I deeply thank.

4.1 Knots in proteins

From a mathematical point of view, a knot can be defined only in closed strings. Indeed, in an open one it is always possible to find a set of moves able to untangle it. Nonetheless, the concept of *physical knot* has been introduced to describe long-lived self-entangled configurations in open strings.

Loop closure is the main strategy to systematically search knots in proteins [171]. One can artificially close the open polypeptide by extending the termini of the protein far enough to be able to connect the chain without crossing it by using an arc at "infinity". This approach is suitable because in most of the cases the protein termini are exposed to the surface and can be extended with no risk to create spurious knots. Once the chain is closed, it is possible to calculate topological invariants defined for mathematical knots, for instance the Alexander polynomial, which we have employed in our work [171].

A slipknot is another interesting non-trivial topology that can be found in an open chain and thus also in a protein [172, 173]. It is obtained by threading a loop through another one, in such a way that the chain is globally unknotted. In other words, and to give a more operative definition, a knot is tightened if its termini are pulled, whereas the same action would untie a slipknot.

In 1994 Mansfield was the first to systematically survey the PDB looking for the presence of knots [174](curiously, at that time the database contained only 400 structures), but concluded that there were none. Only in 2000 Taylor reported the finding of the first deeply embedded knot in a protein [175]. Since then, the number of protein structures deposited in the PDB has increased exponentially. Several groups surveyed again the data bank [176**?** –181], and many structures representing knotted proteins have been found. According to the latest data [**?**], 620 knots and slipknots were found considering 74.223 structures, accounting hence for about the 0.85% of the total amount of known protein structures. All knots are found in enzymes [182]. One can hence conclude that knots in proteins do exist, although they are very rare.

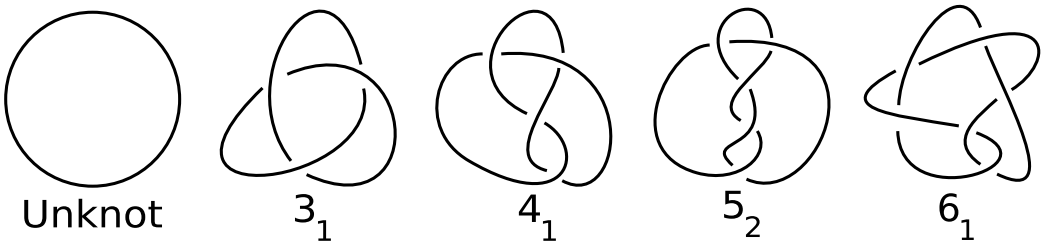


Figure 4.1: Knot types found in natural protein structures stored in the PDB. Figure adapted from Wikipedia.

A knot is classified according to the minimum number of crossings found projecting it on a plane. The simplest topology is the trefoil knot 3_1, with three crossings. Then, raising the number of crossings, one gets more types, namely 4_1, 5_1, 5_2, 6_1, 6_2, 6_3. So far only the 3_1, 4_1, 5_2, 6_1 knot types have found to occur in proteins [182](see Fig. 4.1)

The size of the knot is the smallest number of amino acids that have to be cleaved at either termini in order to make the knot disappear. We say the knot is *shallow* when the size measures up to about 20 residues, otherwise it is a *deep* one.

Most of the knots found in native structures are the simplest possible one, the trefoil topology, and $\sim 2/3$ are indeed very shallow [182, 183].

Knots in naturally-occurring proteins differ for at least two major aspects with respect to flexible polymers of equivalent length. First, they are statistically much rarer [184]. Secondly, the type, location and length of knots occurring in open flexible homopolymers have a stochastic character [185], whereas for natively-knotted proteins they are specific and robustly reproduced in repeated folding experiments [186].

4.1.1 Function and evolution

It is still under debate whether the presence of a knot provides a protein with a precise biological functional advantage. Several possible effects have been proposed, in particular an enhanced thermal, kinetic or mechanical stability [176, 179, 180, 187–192], but no clear proof exists and the questions is still open.

The latter debate is related with the understanding of which evolutionary

path led to the emergence of knotted proteins. It is interesting to note, following Shacknovich [184], that the use of an evolutionary argument has been completely reversed in few years. Indeed, it was believe that, as in collapsed homopolymers, knots would have to be ubiquitous also in proteins. To explain their actual absence, researchers invoked an evolutionary pressure eliminating knots in proteins, since they would have severely hindered the folding reliability and efficiency. On the contrary, after that a significant number of knotted structures were identified, an opposite selective pressure is now considered. This is supposed to preserves these occurrences because of their presumable role in a still unknown important function. This idea is also supported by the evidence that knots are preserved among and across protein families [176, 177, 179**?**].

Recently, the creation of an artificial protein able to spontaneously fold to a knotted native conformation [173] suggested that such a pressure might not be necessary [184]. Indeed, the Yeats group was able to create and artificially knotted protein starting from an existing dimer using the domain fusion technique [173]. The original dimer is composed of two proteins which naturally intertwine, and by adding a linker made of 8 residues connecting them, a new trefoil knotted protein was obtained. King *et al.* also engineered a version of the dimer connected by a cysteine bond, thus linked but topological unknotted. This permitted them to experimentally characterize the thermodynamic and folding kinetic properties of the artificially knotted protein, and compare them with the test cases represented by the natural and the linked dimer. Having such a control case, usually represented by couples of homologous proteins, is of utmost importance when investigating the cause and effect of a given characteristic, as also shown in [168, 190]. It was found that the new artificial protein folds more slowly than the natural one, albeit displaying a folding time of $\sim 20\,\mathrm{s}$, compatible with most unknotted proteins, and only 20 times higher compared to the test cases. Furthermore, there are evidences that the protein unfolds by untying the knot in chemical denaturants, a characteristic that is not shared by naturally knotted proteins [193], as will be soon described. Hence this artificial protein could be a particular case showing peculiar characteristics. Nonetheless, the domain fusion mechanism by which it was obtained suggests a general molecular mechanism possibly exploited by evolution in shaping knotted proteins.

4.1.2 Experimental characterization

The most experimentally investigated knotted proteins are the bacterial α/β-knotted methyltransferases YibK from *Haemophilus influenzae* and YbeA from *Escherichia coli*[1]. Both are single-domain homodimeric proteins and display a deep trefoil knot ($\sim$ 40 a.a.). Their folding was extensively characterized by Mallam and Jackson [182, 186, 193–200], in the following we particularly refer to their latest work [186]. Here, the authors studied the folding and knotting process by building a cell-free transcription-translation system, that is an *in vitro* set-up containing all the essential components to synthesize the proteins, but nothing more. In this way, considering newly translated chains, they were sure to follow the folding starting from an unknotted initial unfolded configuration. Mallam and Jackson could therefore conclude that YibK and YebA can spontaneously fold to their native knotted configuration in an efficient way (i.e., with no trace of misfolded species), and with no aid from any cellular machinery. Knotting is hence a post-translational event.

In a previous work [193], the same authors found that in urea-denatured conformations of YibK and YebA, although all secondary structures were disrupted, the backbone was still self-tied. Since unknotting has never been experimentally observed, they suggest that knotted configurations could be kinetically trapped and knotting itself irreversible.

They also measured the folding rate starting from a newly translated chain and the refolding time of a chemically denatured one [186]. Results are reported here in table 4.1. It can be seen that folding from an unknotted configuration is a slow process, displaying folding times of about $10-20$ min.

To investigate the role of chaperones in the folding of YibK and YebA, Mallam and Jackson added the GroEL-GroES complex to the cell-free transcription-translation system [186]. They found that chaperones enhance the rates and thus accelerate the folding and knotting process. Moreover, the accelerated rates are compatible with those measured considering the refolding of configurations denatured in urea, which are conversely not significantly accelerated by chaperones. Mallam and Jackson were hence able to conclude that, for

[1] YibK is 160 residues long (PDB 1MXI) and YbeA is 155 residues long (PDB 1NS5)

	YibK	YbeA
newly translated	0.05	0.09
urea denatured	1.8	0.3

Table 4.1: Folding rates for YibK and YebA using newly translated (unknotted) configurations, and chemically denatured (knotted) ones. Rates values are in min^{-1}. Data are taken from Ref. [186].

this pair of proteins, knotting is the rate-limiting step, which is specifically enhanced by the presence of chaperones [186].

4.1.3 Computational approaches

Experimental techniques are not able to resolve the knotting mechanism yet. This latter is the specific sequence of conformational changes leading a terminus to thread one or more loops, consequently tying the backbone. In order to unravel it, the insight offered by computational approaches becomes fundamental [201].

Several studies appeared during the last years, ranging from lattice models of proteins [191, 202, 203], CG and AA representations employed with Gō-type energy functions [168, 172, 187, 204–207], to AA simulations using a realistic force-field in implicit [169] and explicit solvent [188, 192, 208].

Despite the many methodological differences, a rather homogeneous scenario emerges. Indeed, all the results generally agree pointing out that a protein knots by threading a terminus through a native loop. This process can occur following two possible pathways. The first one is characterized by the threading of a straight terminus through a native loop, as a string in the eye of a needle. In the second one the terminus threads the loop in a hairpin-like configuration (i.e., partially bent backwards), thus by proceeding as a temporary slipknot, and only at the end by forming the physical knot folding to its native straight configuration. We shall refer to the former mechanism as *direct threading* and to the latter as *slipknotting* (see Fig. 4.2).These two mechanisms are not only related to proteins but have proven to be quite general [185].

Remarkably, by using Gō-type potentials alone, thus just considering

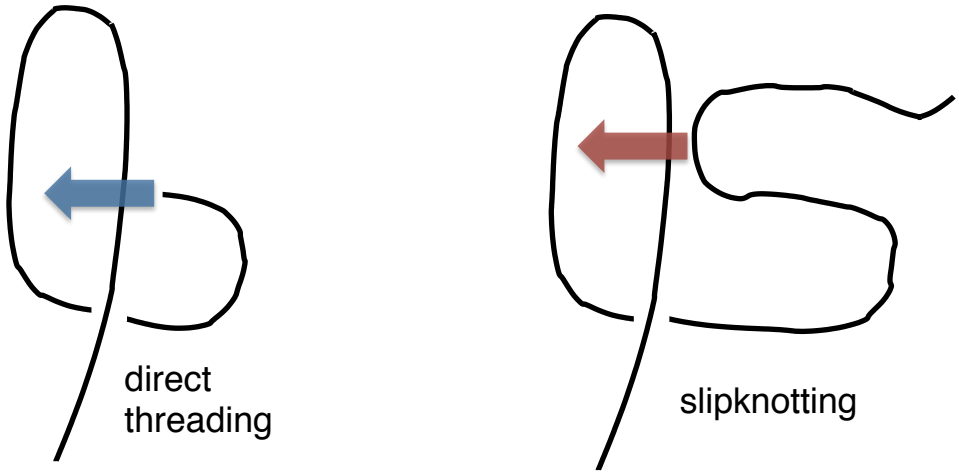


Figure 4.2: Sketches representing the two general knotting mechanisms, direct threading and slipknotting. Fig. reproduced with permission from Ref. [170].

native-like attractive interactions, it is possible to obtain the knotting of AA models of several proteins starting from fully extended configurations [205–207]. Hence, we can conclude that the funneled energy landscape picture still holds even for a complicated topology, or in other words, that the fraction of native contacts is a reasonable reaction coordinate also for the folding of knotted proteins.

Slipknotting is more frequent when Gō-models are used to simulate the folding. In particular, Noel *et al.* have shown that MJ0366, the smallest known knotted protein, could fold by following both pathways [205]. However, raising temperature or using a longer threading terminus dramatically favoured the slipknotting mechanism, which became the dominant one, occurring in all the cases. Since Gō-type FF only promote the formation of native interactions, with the same favorable contact energy, it is plausible to deduce that slipknotting is entropically favored.

Non-native interactions could be nevertheless important. This effect was first explored in a seminal study by Wallin *et al.*, who simulated the folding of an AA representation of YibK in a Gō-type force field [204]. The authors obtained correctly knotted native conformations only by promoting a specific

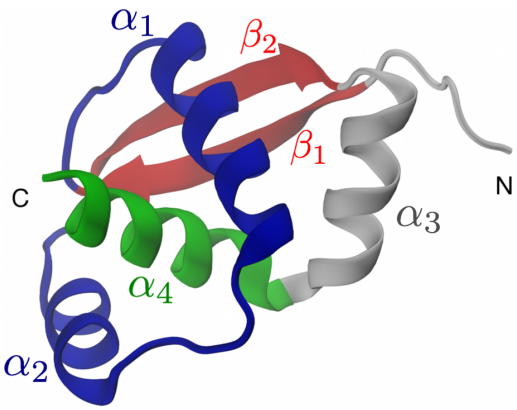


Figure 4.3: MJ0366, the smallest known knotted protein [209]. The PDB structure (2efv) is only 82-residues long, and displays a shallow trefoil knot at its C-terminus. Figure reproduced with permission from Ref. [169].

subset of non-native contacts, effectively mimicking an hydrophobic attractive interaction between a native loop and the threading terminus. They concluded that non-native contacts made the knotted configuration more kinetically accessible, presumably by stabilizing the TS.

4.2 Folding the smallest knotted protein with a realistic force field

In 2010 MJ0366 the smallest known knotted protein was reported in Ref. [209]. This protein, isolated in *M.jannaschii*, is only 92-residues long, but only 82 have been resolved and are available on the PDB (code: 2efv). We specifically used the latter reduced structure, reported here in Fig. 4.3, which displays a shallow trefoil knot located at the C-terminus, with a size of about 5 amino-acids. The knot is composed by the C-terminal α_4–helix (green) which protrudes a native loop formed by α_1 and α_2 helices (blue) and two unstructured coil segments of the chain. A highly non-local β–sheet (red) is formed by β–strands β_1 and β_2 , which are separated in the sequence by 38 amino-acids, and is responsible for locking the native blue loop.

Since we considered the investigation on the folding dynamics of the WW domain to be a successful benchmark, this little protein displaying a non-

trivial topology is a natural test case to further validate and "stretch" the applicability domain of the DRP computational approach. We therefore simulated the folding by considering almost 100 unfolded initial configurations and a total of almost 4000 rMD trial trajectories. Eventually, we selected the most probable ones according to the DRP reweighting criterion and ended up with 31 folding trajectories in which the knot was correctly formed (for the computational details refer to section 4.2.5). To our best knowledge, the results shown and discussed in this section represent the first instance where a realistic FF is employed to follow the folding of initially unfolded and unknotted AA conformations into a knotted native state, albeit in implicit solvent and in presence of a native bias.

The experimental results have determined that knotting is a post-translational process [186]. It is hence a well-posed endeavor to use our folding trajectories to investigate some of the open questions that rise from the computational studies already done on the spontaneous knotting of proteins. We therefore particularly focused in characterizing the knotting event, trying to understand when it happens, whether the protein can fold following different pathways and by which particular mechanism the terminus threads through the loop. Furthermore, using a realistic force field, which naturally takes into account not only native attractive interactions but also non-native ones, is a valuable opportunity to probe the role of the latter in the folding of this particular class of proteins.

4.2.1 Characterizing the folding trajectories

Most of the attempted rMD folding simulations did not converge to the correct knotted native conformation. Eventually, we obtained 31 folding trajectories each starting from a different initial configuration. In all cases, the knotting event corresponded to the formation of the native trefoil knot (i.e., we did not see any non-specific knot), thus indicating that incorrect knot formation is not a major source of kinetic trapping for MJ0366.

We carried out several different analyses in order to characterize the mechanism leading to the formation of the knot.

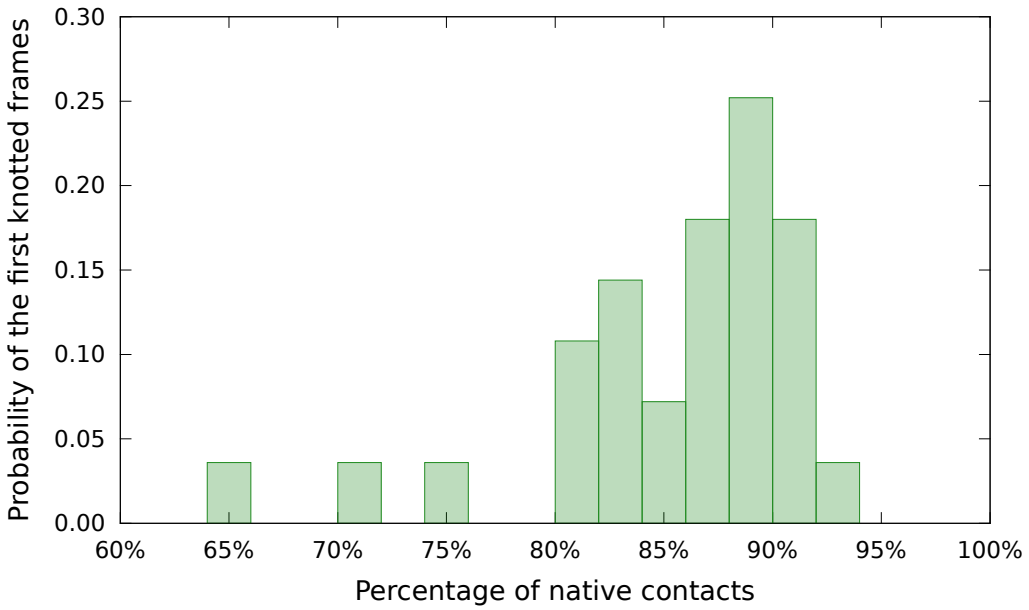


Figure 4.4: Probability of detecting a knot for the first time in a frame vs. fraction of native contacts in that frame. All the selected DRP folding trajectories were considered. The first knotting event occurs at a rather late stage of the folding. Fig. reproduce with permission from Ref. [169].

4.2.1.1 When the knot forms

As a first step we identified the folding stage at which the backbone self-ties into a knot. Accordingly, for each trajectory we calculated the percentage of native contacts that were formed when the first knotting event had occurred (a backtracking event spontaneously unknotting the molecule is almost impossible in the rMD scheme). The distribution of these overlaps for the considered trajectories is shown in Fig. 4.4. The distribution is peaked at about 90% overlap. This indicates that the knot is typically formed at a rather late stage of the folding process.

4.2.1.2 Measuring the pathway heterogeneity

To quantitatively measure the folding pathways diversity we implemented the analysis described by Camilloni *et al.* [134], that will be shortly summa-

rized in the following. Here we consider a folding mechanism as the specific sequence of formation of native contacts. Hence, for each path we measured the time of formation of each native contact as the frame of the trajectory where the contact is firstly formed. Given t_{ik} as the time of formation of the i^{th} native contact in the k^{th} trajectory, we computed for each path k the matrix $M_{ij}(k)$ defined as:

$$M_{ij}(k) = \begin{cases} 1 & t_{ik} < t_{jk} \\ 0 & t_{ik} > t_{jk} \\ \frac{1}{2} & t_{ik} = t_{jk} \end{cases} \tag{4.1}$$

containing all the information regarding the folding mechanism as defined above. For each pair of pathways k, k' it is possible to compute a *similarity* s defined as

$$s(k,k') = \frac{1}{N_c(N_c-1)} \sum_{i<j} \delta\left(M_{ij}(k) - M_{ij}(k')\right) \tag{4.2}$$

N_c being the total number of native contacts. This similarity ranges from 0, i.e., a completely different mechanism, to 1, which means exactly the same mechanism. This quantity measures the consistency of the temporal succession in which the native contacts are formed in two given pathways.

Finally, it is possible to compute the (un-normalized) distribution

$$p(s) = \sum_{k<k'} \delta\left(s - s(k,k')\right) \tag{4.3}$$

of the similarity parameter for all the folding pathways and plot it. A single narrow peak in this plot can be interpreted as a homogeneous folding mechanism, meaning that for any couple of trajectories the s parameter is equivalent. On the contrary, a broad distribution (or, as a limit case, two distinct peaks) is likely to represent a heterogeneous folding. In this latter case, the similarity analysis is not able to yield the number of different mechanisms.

The value of s depends only on the time order of native contact formation events, and not on their exact timing. Thus this scheme can be properly applied to analyze our DRP trajectories.

To have a robust indication of the degree of heterogeneity of the successfully knotting trajectories, we computed the distribution of s over all possible

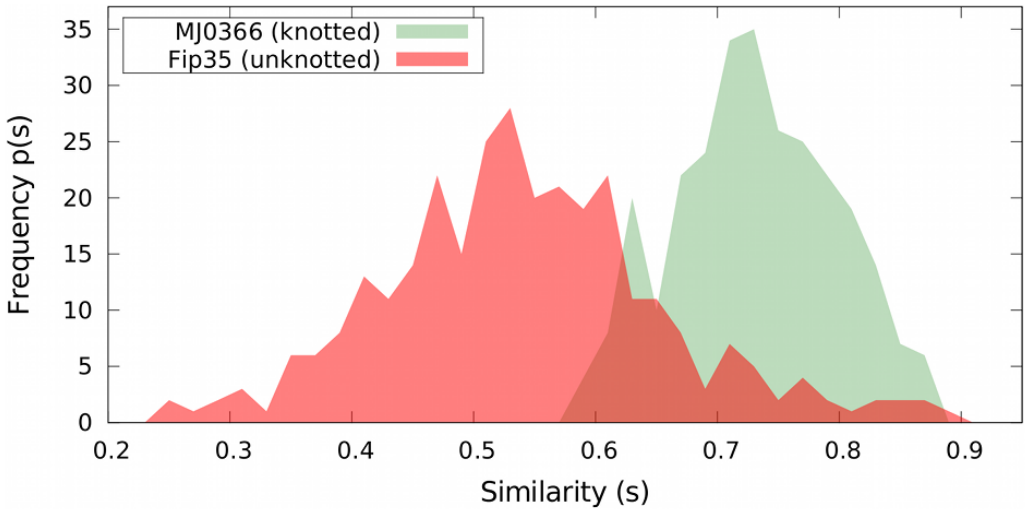


Figure 4.5: In red the un-normalized distribution of the similarity parameter calculated for the successful folding trajectories of the Fip35 domain. In green the same calculation for the knotted MJ0366 protein. Fig. reproduced with permission from Ref. [169].

pairs of trajectories, see Fig. 4.5. As a term of reference term, the same Fig. shows the s distribution computed over previously studied folding trajectories of the unknotted WW domain FIP35, that we have already discussed in Chapter 3. It can be seen that the distribution of MJ0366 is narrower and shifted towards significantly higher values of s than for the unknotted protein. Indeed, the former has a peak at $s \approx 0.75$ whereas the latter has it at $s \approx 0.5$. This means that in the pairwise taken folding trajectories of the knotted protein 75% of native contacts are formed with the same order. This overlap reduces to 50% in the unknotted protein folding trajectories. Moreover, the spread of the distribution is rather different in the two cases. Regarding the unknotted protein, we can find trajectory pairs in which only 25% of native contacts forms with the same order, or other pairs sharing 90%. This variance is significantly reduced for the knotted protein case, since the distribution ranges from $\approx 65\%$ to $\approx 90\%$.

The relatively low value of s and the distribution broadness is typical of folding processes that proceed by multiple pathways, as we have shown Fip35 does. The different characteristics of the s distribution for MJ0366 therefore strongly suggest the existence of one dominant folding pathway.

4.2.2 How the knot forms

To gain further insight in the knotting mechanism, we profiled the folding trajectories along two relevant order parameters: the RMSD to the native structure and the RMSD to the native β–sheet. The first collective variable monitors the overall progress towards the native geometry. The second one, instead, carries information about one of the expected entropic bottlenecks of the folding process. In fact to form the native anti parallel β–sheet amino-acid pairs wit a sequence separation as large as 38 have to meet. Since in the native MJ0366 structure the C-terminal helix protrudes through a native loop that is locked by the two paired β–strands, monitoring the formation of the β–sheet is relevant to understand whether the sheet is formed before or after the knot.

The results shown in the left panel of Fig. 4.6, where we plotted the negative of the logarithm of the probability of each bin. Following Ref. [207], we will call this plot a *"kinetic* free energy surface", since it is calculated as a thermodynamic free energy but lacking the equilibrium condition.

Inspection of the plot indicates that the β–sheet is fully formed quite early, when the total RMSD to native of the chain is about 15 Å. At this stage the fraction of formed native contacts is about 40-50%. The self-tying of the molecule into a trefoil knot typically occurs after the formation of the β–sheet. This is evident from the placement of the diamond symbols in Fig. 4.6, which mark the first occurrence of knots for each of the 31 trajectories. It is seen that all first-knotting events occur when the β–sheet is fully formed, with only two exceptions that will be discussed later.

A detailed inspection of the trajectories highlights the particular mechanism by which the knot forms (Fig. 4.7) . We found that this happens almost invariably through the direct threading mechanism. Indeed, in 26 DRP trajectories over a total of 31, the C-terminal α–helix (residues 74-87) directly enters, without bending, the open region between amino acids 17-54 involving helices α_1 and α_2 and the intertwining loop. In this case, the threaded region and the β–sheet (respectively shown in blue and red in Fig. 4.7) establish a tertiary contact before the terminal helix penetrates into the open region in between the helices α_1 and α_2.

In three other cases, the folding was found to occur through the slipknotting. In all the three instances the C terminus entered the loose α_1–α_2 region in a hairpin–like conformation, as shown in the central panel of Fig. 4.7.

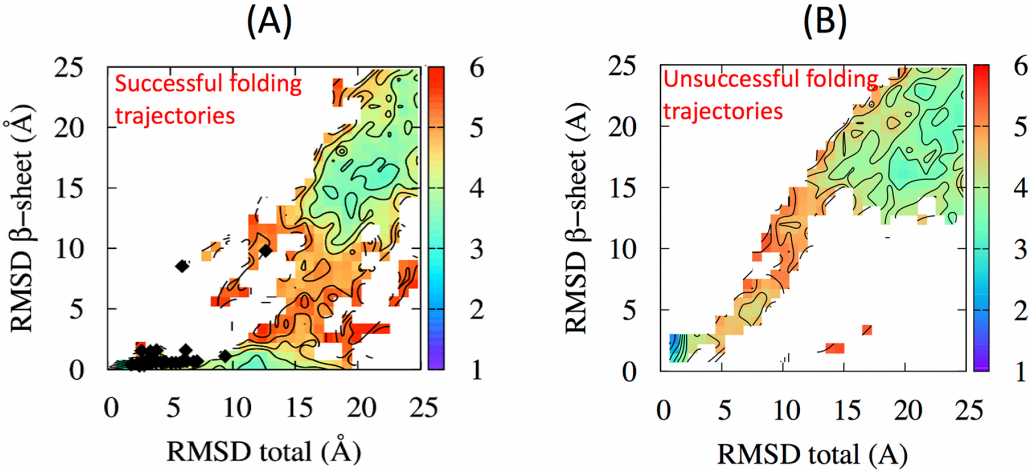


Figure 4.6: Kinetic free energy surface calculated for the RMSD to native of the non-local β–sheet and the RMSD of the global structure. The scale on the left corresponds to the logarithm of the number of times a given spot is visited by the DRP trajectories. Panel (A) shows the surface obtained using successful knotting trajectories. The diamonds mark the collective coordinates at the time of knot formation. Panel (B) was obtained using unsuccessful trajectories. Figure adapted from Ref. [169].

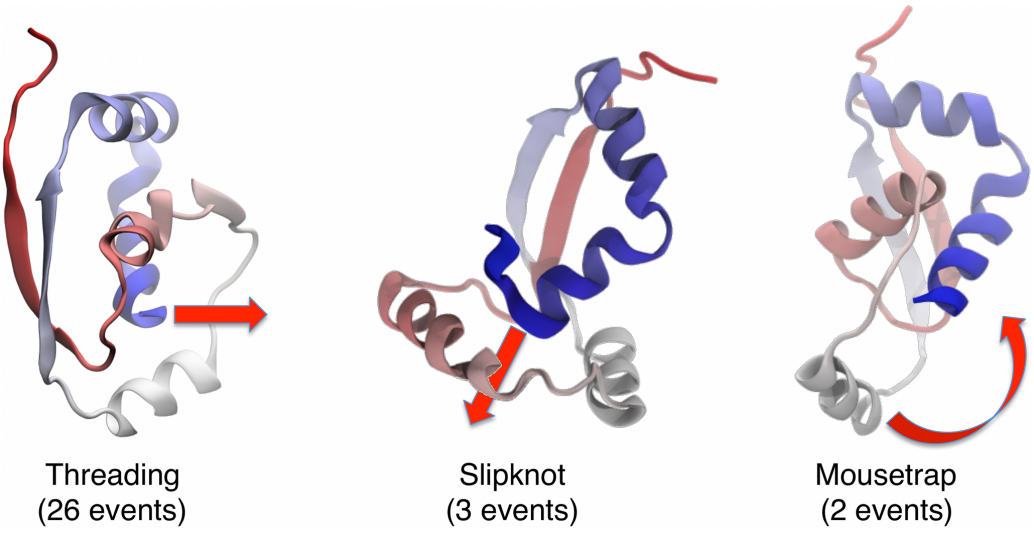


Figure 4.7: The three different types of knotting mechanisms observed in our atomistic DRP simulations. Fig. reproduced with permission from Ref. [169].

Finally, in two further cases we observed another knotting mechanism which involves a concerted backbone movement that had not been previously reported for MJ0366. Namely, in these two trajectories when the β–sheet and the terminal α–helix are already formed and closed in an unknotted configuration, the loop performs a "mousetrap-like" movement establishing the native knotted topology. This movement is schematically represented in the right panel Fig. 4.7. Mousetrap knotting events correspond to the two outlying diamonds reported in Fig. 4.6, with collective coordinates (6 $\mathring{A}$, 8 $\mathring{A}$) and (12 $\mathring{A}$, 10 $\mathring{A}$).

4.2.3 What happens when knotting fails

To investigate what happens when knotting does not succeed, we have carried out a comparative analysis of the reaction mechanism in successful and unsuccessful folding trajectories. Trajectories can be unsuccessful either when the final structure is not similar to the native one (high RMSD to native values), or when it is similar (low RMSD values) but the knot is not properly formed. Specifically, we considered the successful set of the 26 trajectories displaying the dominant direct threading knotting mechanism. The unsuccessful set included an equal number of trajectories that reached an unknotted configuration and nevertheless had a good native similarity (namely an RMSD to the crystal structure less than 5 Å).

The projection of the unsuccessful trajectories along the two collective coordinates considered before is shown in Fig. 4.6. The qualitative difference with respect to the analogous plot for the successful ones shown in panel (A) is striking. In particular, it is seen that in successful trajectories the formation of the sheet involving strands β_1 and β_2 occurs rather early and prior to the establishment of the overall tertiary organization of MJ0366. In fact, the total RMSD to native decreases appreciably only after the β–sheet is established. By converse, for unsuccessful trajectories, this hierarchy of contact formation is not observed, and the β–sheet formation proceeds in parallel with the acquiring of the overall native structure. Therefore it is possible to conclude that the early formation of the β–sheet provides the most appropriate conditions for knotting.

This conclusion is supported by the detailed inspection of the unsuccessful trajectories, which are exemplified in the sequence of snapshots shown in Fig. 4.8. As it is visible in this figure, the C–terminal helix threads the correct

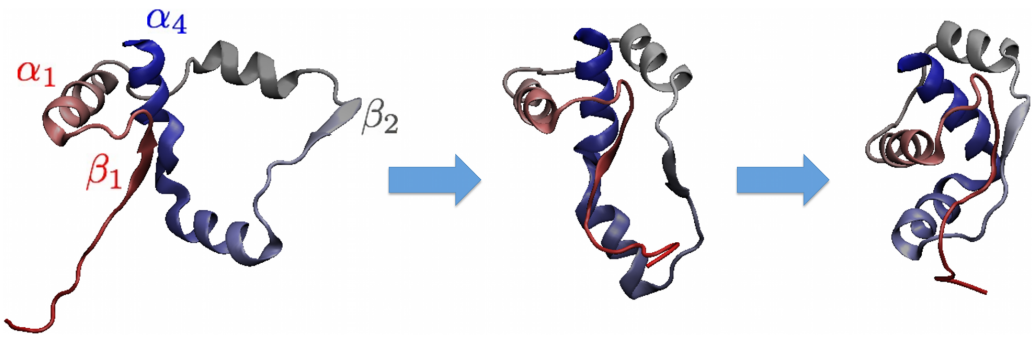


Figure 4.8: Typical example of unsuccessful trajectory. The α_4-terminus threads the loop at an early stage, whereas the late formation of the β–sheet traps the N-terminus on the "wrong" side of the loop and prevents attaining the (native) knotted topology. Fig. taken from Ref. [169].

region between strands β_1 and β_2 prior to the formation of the β–sheet. When the latter is finally established, the N–terminal segment remains trapped on the wrong side of the loop bridging β_2 and α_3, and for steric reasons cannot go past it and attain the native knotted topology. The relevance of this mechanism for misfolding is highlighted by the fact that all unsuccessful trajectories displays a late formation of the β–sheet.

We emphasize again that, according to our simulations, the correct knotting of the chain is not promoted by the formation of specific contacts which fail to form in misfolding events. Instead, for the chain to acquire the native topology, it is essential that the native contacts form in the correct order.

4.2.4 Discussion: the role of non-native interactions

So far we have shown that the DRP algorithm is able to fold the MJ0366 protein to the correct knotted configuration. Analyzing the 31 trajectories in atomistic detail, we found that knotting occurs at a very late stage of the folding and that native contacts form in a rather similar order. Moreover, we characterized the knotting mechanism, and although three possible pathways are detected, only one (for instance direct threading) is far more likely than the others.

Unfortunately, the folding of MJ0366 has not been experimentally characterized yet. It is thus natural to discuss our results by comparing them

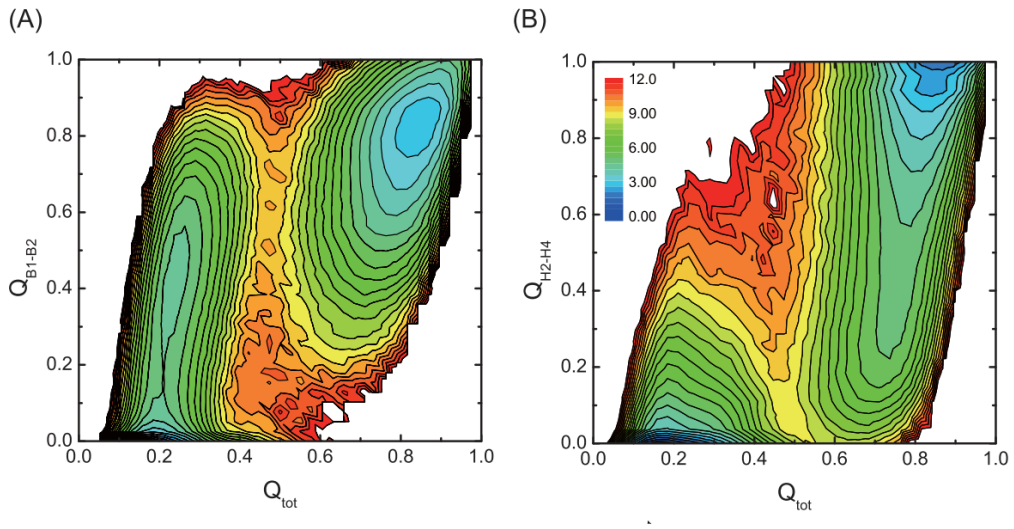


Figure 4.9: Free energies plots calculated from folding simulations MJ0366 using a native-centric potential. (A) Free energy calculated for the fraction of native contacts of the non-local β–sheet vs. the overall one. (B) Fraction of native contacts between the threading C-terminal α_4–helix and the α_2–helix from the native loop. Fig. reproduced with permission from Ref. [207].

with those obtained by means of Gō-type potentials [205, 207] and unbiased MD in explicit solvent [208].

Li *et al.* have investigated the folding of MJ0366 using an advanced Gō-type potential displaying residue specific interactions and flexibility based on a statistical potential [207]. They reversibly folded and knotted the protein, and calculate two free energy landscapes, reported here for convenience of the reader in Fig. 4.9.

In panel A the authors show the free energy landscape calculated using the fraction of native contacts formed between the two β strands and those formed by the whole structure. They conclude that the β–sheet has to form to its native conformation before the overall structure can increase the fraction of native contacts. It can be seen that the basin of the intermediate state where the β–sheet is formed is separated from the native one by the TS. Hence, they conclude that knotting is the rate limiting step [207].

In panel B Li *et al.* show the free energy landscape calculated using the fraction of native contacts formed by the threading helix and a helix compos-

ing the native loop, and again the overall value of Q. Since no intermediate state is detected, the authors draw the conclusion that knotting is the last event of folding and happens at a late stage [207]. Ref. [205] claims similar conclusions.

The scenario emerging from these results is in fairly good agreement with our findings, as already discussed and shown in Fig.'s 4.4 and 4.6.

4.2.4.1 Slipknotting vs. direct threading

As already mentioned, Noel *et al.* [205] simulated the reversible folding of MJ0366 using a Gō force field both in CG and AA resolution. They obtained knots with a very low efficiency in both the representations of the molecules. Using the AA resolution enhanced the yield of knotting events and, most importantly, avoided the formation of knots in non-native locations of the amino-acid chain. They concluded that taking into account side-chains lowers the probability of the protein to be locked in a topological trap and dramatically enhances the specificity of the knot.

Noel *et al.* also characterized the knotting mechanism. Notably, they found two possible pathways, direct threading[2] and slipknotting. At the melting temperature and using the crystal structure of MJ0366 with a "shorter" C-terminal α–helix, the pathways were followed with approximately the same weight (direct threading 55%, slipknotting 45%) . But by lowering the temperature and by using the extended conformation of the knotted protein (which displays 5 more residues on the C-terminus), direct threading was not detected anymore and slipknotting accounted for all the knotting events [205]. This competition between the two mechanisms might be a hallmark of a subtle enthalpic-entropic balance. In fact, a threading terminus has to severely reduce its conformational entropy to enter a loop. The subsequent increase in free energy can be mitigated by stabilizing interactions between the terminus and the loop (native or non-native), which would favour a direct threading mechanism. Conversely, as stated in [205], temporary contacts could form also in a slipknot conformation of the terminus, balancing the cost of the entropic reduction and creating an effectively shorter terminus. Since the entropy loss required for threading increases with the length of the terminus, a shorter terminus would enhance the probability to thread the loop.

[2] In paper [205] they refer to the direct threading mechanism as "plugging".

4.2.4.2 Turning non-native interactions on and off

One of the major difference between our approach and the one exploited by Noel *et al.* is the presence of non-native interactions. In fact, in the Gō-type force field employed in Ref. [205] these are absent, since only native attractive interactions are taken into account. Whereas attractive non-native interactions are naturally considered in a realistic force field as AMBER99SB-ILDN [56], which was used in our study.

In order to investigate a possible role for non-native interactions, we generated several folding trajectories for MJ0366 using simplified CG models and two Gō-type energy functions where the effect of non-native interactions could be easily turned on or off, which we have already described in Chapter 3. Specifically, we considered a first model with only native-centric interactions, and a second one additionally incorporating non-native interactions. The latter included quasi-chemical and screened electrostatic pairwise interactions, effectively mimicking non-native and hydrophobic interactions. These two force fields have already been applied to a knotted protein in the recent study of the early folding stages of a trefoil-knotted carbamoyltransferases [168].

The folding process presents major differences in the two models. First, they differ significantly in terms of knotting probability. In particular, for each model we considered an extensive set of 10,000 uncorrelated configurations, equilibrated at the nominal temperature of 300 K. In the native-only case, 12% of the sampled configurations were knotted, while this number had a sixfold increase, up to 75%, in presence of non-native interactions. This result aptly complements the atomistic DRP simulations, for it highlight the helping role of non-native interactions in the formation of the native knotted topology of MJ0366.

Second, productive trajectories obtained by a dynamical MC simulation follow different dominant mechanisms in the two models. Namely, when the pure native-centric force field is used, 8 out of 10 trajectories involved the slipknotting mechanism, while the threading one is observed in all trajectories (10 out of 10) with the additional non-native interactions. The latter result, which is in full accord with the atomistic DRP simulations, reinforces the concept that non-native interactions can promote the correct order of contact formation following the direct threading mechanism.

This point is further supported by the inspection of the density plots in

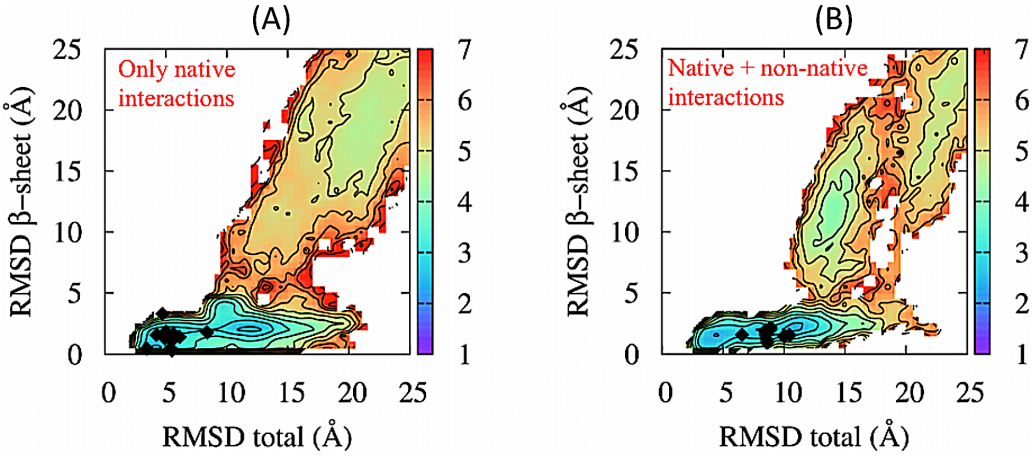


Figure 4.10: Kinetic free energy surfaces calculated using folding pathways obtained from CG MC simulations with local crankshaft moves which mimic the chain dynamics, projected on the plane selected by the total RMSD to native and by the RMSD to native of the β–sheet. Panel (A) refers to the model with only native interactions, while panel (B) refers to the model with both native and non-native interactions. The diamonds denote the values of the collective coordinates at the time of knot formation. The scale on the left is the logarithm of the number of times the point is visited by folding trajectories, in analogy with free-energy landscape plots. Fig. reproduced with permission from Ref. [169]

Fig. 4.10. Indeed, non-native interactions are more clearly associated to the early formation of the β–sheet than in the native-only case.

Notably, we shall remember that, as already shown in Chapter 3, the same analysis yielded no differences in the folding of the Fip35 WW domain.

4.2.4.3 The role of non-native interactions

The enhanced knot formation discussed above can be explained by observing that in the simplified model the early formation of the β–sheet is promoted by the fact that the non-native quasi-chemical interaction generates an overall attractive interaction between the residues in β_1 and those in β_2. Consistently with the misfolding events previously discussed, it is possible to argue that the weaker drive of the native-centric model to form the β–sheet early on is

also responsible for its lower knotting propensity.

According to these results, we can conclude that mutations in the β–sheet regions with residues characterized by a weaker effective attraction would delay the formation of the β–sheet in the folding process, and make the chain more prone to reach the unknotted misfolded state. This prediction may be verified experimentally.

The impact of non-native interactions in driving the knotting has also been proved by Škrbić *et al.* in [168]. Here the authors considered two homologous evolutionary-related proteins, similar in sequence and structure but differing by the presence in one of them of a knot[3]. The initial stage of the folding of these big proteins was investigated employing the same two simplified Go-type potentials described above. The authors' striking conclusion is that non-native interactions have a relevant role only in proteins displaying a native knot, where they dramatically enhance the knot's probability to form in an early stage of the folding. This happens because they effectively introduce an attractive interaction between the threading terminus and the hydrophobic core where the native loop is located [168]. A similar effect of non-native interactions was also suggested by Wallin *et al.* in their already mentioned study [204].

Noel *el al.* recently published the result of several folding simulations of MJ0366, by means of AA unbiased MD in explicit solvent, carried on the ANTON supercomputer [208]. They considered the extended crystal structure of the protein, characterized by a knot size of 15 amino-acid. Simulations were initiated from 15 almost folded configurations, where a slipknot was formed and had already threaded 10 over 15 residues through the loop. Using 2 μs for each configuration the authors reported several cases, namely completion of the knotting process, backtracking of the slipknot, and trajectories where the slipknot was trapped during all the time. In all this instances a multitude of temporary salt-bridges was detected, which either stabilized temporary conformations or trapped them [208].

To conclude, our findings suggest a crucial role for non-native interactions in enhancing the probability to correctly form a knot and determine the actual self-tying mechanism.

[3]Specifically the two proteins are the trefoil-knotted N-acetylornithine carbamoyltransferase (AOTCase, PDB 2g68, 332- residues long), whereas the other is an unknotted ornithine carbamoyltransferase (OTCase, PDB 1pvv, 313-residues long)

4.2.5 Computational details

4.2.5.1 DRP algorithm

We used the crystal structure of the monomeric unit of the natively-knotted MJ0366 protein that can be found in the PDB (2evf), which is 82-residues long.

We first generated an ensemble of 100 denatured configurations by unfolding the crystal structure using 100 ps of atomistic MD simulations at high temperature (1600 K) followed by 100 ps of thermalization at 300 K. All unfolded configurations were unknotted.

The folding and knotting dynamics of MJ0366 was then studied by carrying out 48 folding attempts for each of the 100 denatured configurations by means of the rMD algorithm, for a total of about 4000 attempted folding trajectories. Notably, the biased rMD evolution promotes only the overall geometrical similarity with the native state and does not reward the formation of specific contacts that could lead to knotting.

All the simulations were carried on in atomistic detail using the AMBER99SB-ILDN [56] force field in implicit solvent within the Generalized Born formalism implemented in GROMACS 4.5.2 [167]. The Born radii are calculated according to the Onufriev-Bashford-Case algorithm [54]. The hydrophobic tendency of non-polar residues is taken into account through an interaction term proportional to the solvent-accessible-surface-area [210].

All the trajectories associated to the various knotting modes do not present significant quantitative differences regarding the overall solvent accessibility of polar and non-polar residues during the folding process [169]. This result provides a quantitative basis for expecting that the relative weight of the knotting mechanisms should not depend critically on the specific model adopted to describe the solvent-induced interactions.

Next, we applied the DRP approach and retained only one productive pathway per initial condition, namely the one with the highest statistical weight. This weight corresponds to the probability that each trial trajectory is generated by an overdamped Langevin dynamics. Because the weights are calculated with reference to an unbiased diffuse dynamics, the DRP selection criterion lessens a posteriori the rMD steering effects.

4.2.5.2 Coarse grained simulations

We used the schemes already described in Chapter 3, hence the CG folding simulations were based on the model developed in Ref.'s [95, 156, 159], and the simulations for protein MJ0366 were performed using a MC algorithm described in detail in Ref. [168].

The folding dynamics of CG model with native and non-native interactions was simulated by generating 200 MC trajectories, while the dynamics of the model with only native interactions was investigated by generating 500 MC trajectories. For both CG models, trajectories consist of 1.5×10^8 attempted MC moves, corresponding to 1.5×10^4 saved frames. The employed MC moves were the local crank-shaft and Cartesian moves, whose boldness was chosen such that the acceptance rate was nearly constant and approximately equal to 50%. In both cases, we have collected a total of 10 folding transitions, leading to native configurations with the correct knotted topology.

In order to compute the frequency of knotted configurations at thermal equilibrium, we performed MC simulations which combine local moves and global pivot moves.

4.2.5.3 Knot detection

The conformations visited during the MC dynamics were analyzed to establish their global and local knotted state. The global topological state was established and assigned by computing the Alexander determinants after suitable closure of the whole protein chain into a ring. For each configuration, this entailed 100 alternative closures where each terminus is prolonged far out of the protein along a stochastically chosen direction, and the end of the prolonged segments are closed by an arc "at infinity" (i.e., not intersecting the protein). As in Ref. [168], in order to avoid considering back-turning closures, stochastic exit directions are picked uniformly among those which form an angle of more than 90° with the oriented segment going from each terminus to the C-α at a sequence distance of 10. If the majority of the 100 stochastic closures return non-trivial Alexander determinants, then the whole conformation can be considered as globally knotted. Because protein knotting can occur through slipknot formation [206], the global topology investigation was complemented by a local one. In deed, a slipknot can be

detected by identifying a non-trivially knotted portion of the chain that has a different global topology, in our case the unknotted one. To this purpose, we repeated the above-mentioned statistical closure scheme for all possible sub-portions of length 20, 30, 40, ... of the protein chain so as to identify the smallest knotted, or pseudo-knotted, chain portion [171].

Chapter 5

Projecting a complex dynamics on a simple network: Milestoning

In this last chapter we will show some partial and preliminary results of an attempt to analyze in a more sophisticated way the folding trajectories produced by the DRP algorithm. In particular, we will use the Milestoning method to better assess in a quantitative way what is the effect of the bias applied by the rMD on the folding kinetics.

Milestoning and Markov-State-Models have been originally developed to overcome the timescale sampling limitation of MD simulations ([74, 81, 149, 211–215] and references therein). Both methods are able to use short MD trajectories to obtain important information on much longer timescales, as the MFPT of a conformational transition, by "gluing" the short pieces together. The most famous and peculiar application in this sense is for sure Folding@Home[1], a massively parallel world-wide community effort, which uses the idle resources of personal computers owned by volunteers.

During the last years though, Milestoning and Markov-State-Models have been also used as an effective way to extract the dynamical content of an MD trajectory, thus as a sophisticated analysis tool [74, 146, 147, 154, 216–218]. Through a coarse graining, both approaches are able to reduce a trajectory to a network which gives a quantitative description of the transitions occurring in the trajectory. This network can be simple enough to be human-readable, thus offering the invaluable opportunity of an intuitive insight.

[1] http://folding.stanford.edu/home/

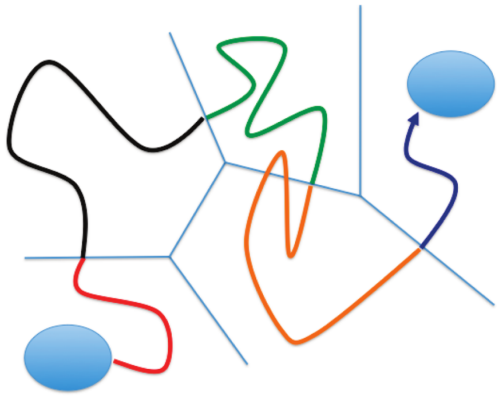


Figure 5.1: A pictorial representation of the Milestoning coarse graining procedure. A trajectory connecting initial and final configurations (blue areas) is separated in many pieces according to a tessellation of the configuration space. The tiles are called anchors, and the borders that separate a pair of tiles are the milestones. For the sake of clarity, in this Fig. there is no distinction between incoming and outgoing milestones. Fig. reproduced with permission from Ref. [216].

The work shown in this chapter was partially done during a visiting period at Prof. Ron Elber's group at University of Austin,. I deeply thank him, his group, and Steven Kreuzer for being so friendly, and for the interesting discussions we had during my staying there.

Unfortunately, due to technical difficulties, the results of our investigation are only partial and no clear and robust conclusion can be drawn. Results shown in this chapter are unpublished.

5.1 The Milestoning algorithm

Let us consider a very long MD trajectory, or equivalently a set of shorter trajectories, which display the evolution of a system. MD trajectories are complex objects, and it is not trivial to extract all their dynamical content. Analyzing molecular conformational transitions by tracking the value of few geometrical order parameters gives a first overview of the dynamics, but a more sophisticated scheme is needed to gain further and deeper insights. This

can be achieved by using the Milestoning algorithm, and in this section we will follow [216] and briefly review the formalism and theory. For a more rigorous and complete overview of the method we suggest Ref. [213].

First of all, we have to coarse-grain the conformational space Ω by defining several subsets $\Omega_i \subset \Omega$ which will be called *anchors*. Intuitively, an anchor is a set of conformations which we can consider to be equivalent at a resolution lesser than the atomic one. A possible way to define and find anchors is to cluster together conformations which are close according to some metric, e.g., when the RMSD between them is under a given threshold. Following this choice anchors correspond to meta-stable states. Another way is to use one's chemical intuition and define anchors based on the particular structure of the investigated molecule. We will use this second approach, which will be thus clearer in the following.

The coarse-graining procedure is a sort of tessellation of Ω, and we define the boundary between two anchors, i.e., the border separating two tiles of this tessellation, as milestone. An evolving system in an MD simulation travels across Ω, spending some time in an anchor, then leaving it to go to another one, and in this way it hits the milestone separating the two anchors. In the following, we will use Latin letters for anchors and Greek letters for milestones. Milestoning analysis consists in neglecting all the details of the microscopic trajectory but just two: at which milestone the trajectory was seen last time, and when. We will use a specific flavor of the algorithm, that is the Directional Milestoning, where crossing the same boundary between two anchors in opposite directions defines two different milestones, the incoming one and the outgoing one. In symbols, given two connected anchors i and j, we define milestone $\alpha : i \to j$, and milestone $\beta : j \to i$. We will use the notation $\mathbf{x}_\alpha$ to say that a given configuration $\mathbf{x} \in \mathbf{\Omega}$ belongs to the set of configurations where milestone α is located.

We shall define the following quantities:

$p_\alpha(\mathbf{x}_\alpha, t)$	the probability that at time t the last milestone to be crossed was α at the configuration $\mathbf{x}_\alpha$.
$q_\alpha(\mathbf{x}_\alpha, t)$	the probability density that exactly at time t a given trajectory hits milestone α at configuration $\mathbf{x}_\alpha$.
$K_{\gamma\alpha}(\mathbf{x}_\alpha, t\|\mathbf{x}_\gamma, t')$	the propagator between milestones, that is the conditional probability density that a trajectory hits milestone

α at time t, given that it last hit milestone γ at time t'.

The above quantities have to satisfy probability conservation, which yields to three relations. The first one reads

$$p_\alpha(\mathbf{x}_\alpha, t) = \int_0^t \mathrm{d}t' q_\alpha(\mathbf{x}_\alpha, t') \left\{ 1 - \sum_{\gamma \in \mathcal{N}(\alpha)} \int_0^{t-t'} \mathrm{d}\tau \int_{\mathcal{V}(\gamma)} \mathrm{d}\mathbf{x}_\gamma K_{\alpha\gamma}(\mathbf{x}_\gamma, t' + \tau | \mathbf{x}_\alpha, t') \right\}, \tag{5.1}$$

where $\mathcal{N}(\alpha)$ is the "neighborhood" of milestone α, that is the set of milestones directly connected with milestone α; $\mathcal{V}(\gamma) \in \Omega$ is the hyper-volume in configuration space which defines milestone γ. Eq. (5.1) states that the probability that the last crossed milestone was α is given by the probability that α was first crossed at a previous time t', minus the probability that in the time interval $t - t'$ the trajectory moved away and crossed a contiguous milestone γ.

The probability density q_α instead satisfies the following Eq.

$$q_\alpha(\mathbf{x}_\alpha, t) = p_\alpha(\mathbf{x}_\alpha, 0)\, \delta(t) + \sum_{\gamma \in \mathcal{N}(\alpha)} \int_o^t \mathrm{d}t' \int_{\mathcal{V}(\gamma)} \mathrm{d}\mathbf{x}_\gamma q_\gamma(\mathbf{x}_\gamma, t')\, K_{\gamma\alpha}(\mathbf{x}_\alpha, t | \mathbf{x}_\gamma, t') \ . \tag{5.2}$$

This Eq. states that to touch milestone α exactly at time t, a trajectory can be already on milestone α at time $t = 0$, or it can transit from a milestone γ directly connected to α that was touched at a previous time t'.

Eq.'s (5.1), (5.2) are rigorous but practically not solvable in cases of interest, thus they have to be simplified. We will consider a stationary process, s.t. the time dependence of the propagator depends only on time differences

$$K_{\alpha\gamma}(\mathbf{x}_\gamma, t' | \mathbf{x}_\alpha, t) = K_{\alpha\gamma}(\mathbf{x}_\gamma, t' - t | \mathbf{x}_\alpha) \ .$$

Furthermore, we will suppose that milestones are distant enough, s.t. when a trajectory crosses a milestone γ it does not remember the specific point where it started on milestone α, i.e.,

$$K_{\alpha\gamma}(\mathbf{x}_\gamma, t' - t | \mathbf{x}_\alpha) \approx K_{\alpha\gamma}(\mathbf{x}_\gamma, t' - t) \ .$$

We now want to obtain a series of relations that depends only on the

milestone indexes, and therefore we introduce the following quantities

$$
\begin{aligned}
p_\alpha(t) &= \int_{\mathcal{V}(\alpha)} \mathrm{d}\mathbf{x}_\alpha p_\alpha(\mathbf{x}_\alpha, t) \\
q_\alpha(t) &= \int_{\mathcal{V}(\alpha)} \mathrm{d}\mathbf{x}_\alpha q_\alpha(\mathbf{x}_\alpha, t) \\
K_{\alpha\gamma}(t) &= \int_{\mathcal{V}(\alpha)} \mathrm{d}\mathbf{x}_\alpha K_{\alpha\gamma}(\mathbf{x}_\alpha, t)
\end{aligned}
$$

where integration removes the explicit dependence on coordinates. We can thus integrate over Eq.'s (5.1) and (5.2), and get

$$
\begin{aligned}
p_\alpha(t) &= \int_0^t \mathrm{d}t' q_\alpha(t') \left\{ 1 - \sum_{\gamma \in \mathcal{N}(\alpha)} \int_0^{t-t'} \mathrm{d}\tau K_{\alpha\gamma}(\tau) \right\} \\
q_\alpha(t) &= p_\alpha(0)\,\delta(t) + \sum_{\gamma \in \mathcal{N}(\alpha)} \int_0^t \mathrm{d}t' q_\gamma(t') K_{\alpha\gamma}(t - t')
\end{aligned}
\tag{5.3}
$$

which are the fundamental equations of the Milestoning method. They still express a probability balance, but now only in terms of milestones. Note that $p_\alpha(t)$ is a probability, whereas $q_\alpha(t)$ is a probability flux in time, and thus has the dimension of an inverse time.

We now assume that, for long times, a stationary probability distribution describing the system exists. It is possible to solve Eq.'s (5.3) looking for the time independent probability and flux, i.e.,

$$
\begin{aligned}
p_\alpha &= \lim_{t\to\infty} p_\alpha(t) \\
q_\alpha &= \lim_{t\to\infty} q_\alpha(t)
\end{aligned}
$$

by using the stationary kernel

$$
K_{\alpha\gamma} = \int_0^\infty \mathrm{d}t' K_{\alpha\gamma}(t') \tag{5.4}
$$

and in this way we get the very simple relation [216]

$$
\mathbf{q}(1 - \mathbf{K}) = 0 \tag{5.5}
$$

where bold characters correspond to matrices and vectors. This remarkable result permits to practically compute the stationary probability flux distribution. The meaning of the stationary probability distribution is the following: if we observe the stationary system for a time interval Δt, the (unnormalized) probability to see a trajectory hitting milestone α is $\propto q_\alpha \Delta t$. Once the stationary flux is calculated, one can define the net probability flux flowing between two anchors i and j as the difference $q_\alpha - q_\beta$, where α and β are incoming and outgoing milestones through the same surface. Note that in Eq. (5.4) $K_{\alpha\gamma}$ is now a probability, thus has to obey the normalization condition $\sum_\gamma K_{\alpha\gamma} = 1$, since the system has to hit a milestone once it leaves α.

Furthermore, stationary probability and flux are related through

$$p_\alpha = q_\alpha \langle t_\alpha \rangle$$

where $\langle t_\alpha \rangle$ is the milestone average lifetime. This is the time interval that a trajectory needs on average to hit a new milestone γ after having left milestone α, i.e.,

$$\langle t_\alpha \rangle = \sum_\gamma \int_0^\infty \mathrm{d}t \cdot t \cdot K_{\alpha\gamma}(t) \ . \tag{5.6}$$

The MFPT to hit a final milestone is written as follows

$$\langle \tau \rangle = \int_0^\infty \mathrm{d}\tau \cdot \tau \cdot q_f(\tau)$$

that is the probability to hit this final milestone exactly at time τ, integrated over all times. This expression greatly simplifies and can be cast in a form that can be used in practical cases,

$$\langle \tau \rangle = \mathbf{p} \cdot (\mathbb{I} - \mathbf{K})^{-1} \langle t \rangle$$

where $\mathbb{I}$ is the identity matrix, and $\langle t \rangle$ the vector made of elements in Eq. (5.6).

The outcome of a Milestoning analysis applied on a set of atomistic trajectories is a directed weighted network. The nodes represent the anchors, whereas the edges are the observed transitions between two anchors weighted by the stationary probability fluxes hitting the corresponding milestone. Once the trajectories have been reduced to such a network, if the

number of anchors is limited, then we get a sort of human-readable representation of the complex dynamics contained in the microscopic trajectories, that permits to have an overall qualitatively and intuitive picture. Furthermore, we can use the numerical value of the probability fluxes across the milestones and build with them many observables to have a more quantitative assessment of the dynamical content in the analyzed trajectories.

In particular, since we have in mind to use the Milestoning algorithm to analyze complex conformational changes in a molecule, we can look for all the connected paths on the network that lead from the initial configuration to the final one. Each of this paths is defined by a set of stationary probability fluxes, and has a global weight given by the sum of all these fluxes. The Maximum Flux Path (MFP) is the path connecting the initial and final configurations along which the highest amount of probability flux flows [219, 220]. If in a given molecular transition the MFP accounts for most of the total probability flux, then it can be identified as a reaction coordinate of the system [216].

In order to practically use the Milestoning algorithm, the two objects that have to be populated from a long MD trajectory or a set of short ones, are the stationary transition matrix $\mathbf{K}$ and the vector of milestone lifetimes $\langle t \rangle$. This can be done simply by projecting the atomistic trajectories on the anchor space, thus obtaining a discrete trajectory $(i_{t_1}, i_{t_2}, \ldots, i_{t_n}, \ldots)$, i.e., the sequence of visited anchors i in time. One can count how many times milestone α is hit, n_α, and how many times a transition $\alpha \to \beta$ is seen, $n_{\alpha\beta}$, and in this way estimate

$$K_{\alpha\beta} = \frac{n_{\alpha\beta}}{n_\alpha} \, .$$

The average milestone lifetime, instead, is simply estimated by the average number of frames that have to pass in order to see the trajectory hitting a new milestone, that is

$$\ldots i \, , i \, , \overbrace{i, \, j}^{\alpha} \, , j \, , j \, , \ldots , \overbrace{j \, , k}^{\beta} , k \, , k \, , \ldots$$

with n_α frames before milestone β is hit. Once matrix $\mathbf{K}$ has been populated with the trajectories, Eq. (5.5) can be solved by means of standard techniques and the stationary probability flux obtained. With this and $\langle t \rangle$, we can further calculate all other quantities.

If we use the Milestoning algorithm to analyze an ideal infinitely long equilibrium MD simulation, then stationarity is natural. But if we use it to get an insight from one or several short trajectories that are out of equilibrium, as for instance the ones produced by the DRP algorithm, then stationarity has to be imposed. This can be done by imposing cyclic boundary conditions on $\mathbf{K}$ once an emitting and an absorbing milestones have been defined. For example, we can be interested in simulating the formation of three h-bonds in an $\alpha-$helix, starting from a completely unfolded configuration where none are present. The latter configuration would define the emitting milestone, the former the absorbing one. We can use very short simulations that show the folding to the final configuration to populate matrix $\mathbf{K}$. Then we artificially impose a cyclic boundary condition by demanding that all the trajectories hitting the absorbing milestone disappear and immediately reappear with probability one at the emitting milestone.

On the other hand, if we want to calculate the MFPT to transit from an initial milestone to a final one, then the latter has to be an absorbing boundary condition, $K_{\alpha\beta} = 0, \forall\, \beta$. In other words, the trajectory dies as soon as it hits the final milestone α.

We reported this brief overview of the Milestoning method for the sake of completeness. For a self contained and rigorous illustration we shall recommend Ref. [213], whereas for a more exhaustive operative implementation we shall refer to Ref. [216].

5.2 Refolding a long myosin chain

We investigate the refolding by means of rMD simulations of chain A of the human cardiac β-myosin S2 structure, which is in its native state a 126-residue long $\alpha-$helix (PDB: 2fxm). We took advantage of several unfolding simulations of the same model performed by using unbiased MD simulations [217, 218]. These studies focus on the first unfolding event, which is the first residue that looses its native configuration. Kreuzer *et al.* showed that unfolding is characterized by a complex diffusive behavior, and a residue visits different conformations before it unfolds. In order to describe this diffusive process, the author introduced a set of anchors, upon which the dynamics of a given residue has to be projected [217]. In particular, they defined anchors depending on the number and kind of h-bonds spanning a given residue, and

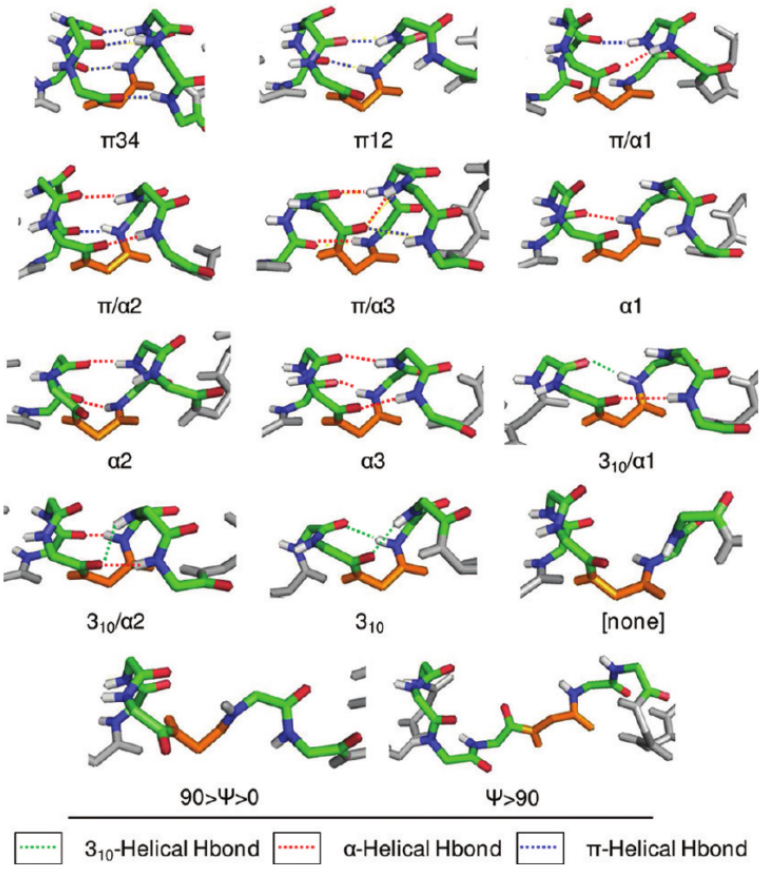


Figure 5.2: Anchors definitions. A given residue (orange) is projected to an anchor depending on the number and kind of h-bonds spanning it, and the value of its dihedral ψ angle. The residue is folded when $\psi < 0$ and three α h-bonds are formed (anchor α_3), whereas it is unfolded when the dihedral angle is open and h-bonds are completely absent (anchor $\psi > 0$). Fig. reproduced with permission from Ref. [217], which we refer to also for all the details of the anchors definition.

the value of the ψ dihedral angle of that residue (Fig. 5.2). Three different types of h-bonds are considered: the usual α types, π h-bonds, which are slightly longer, and 3_{10}, which on the contrary are slightly shorter. H-bonds are defined according to an exclusively geometrical criterion. A residue is folded when $\psi < 0$ and three α h-bonds span it, and we say it is in anchor α_3; on the contrary, it is unfolded when the dihedral angle is wide open and h-bonds are completely absent, case in which it is in the anchor $\psi > 90°$. As we said, while unfolding a residue diffuses through all the remaining anchors represented in Fig. 5.2. Kreuzer *et al.* estimated an unfolding MFPT (in a FF different than the one we employed) of ≈ 5 ns.

A milestone is hit when the residue we are projecting on the anchor space changes its conformation going from one anchor to another. This means that the number of h-bonds spanning it can change, or the type, or the value of ψ.

With this anchor space at hand, we can investigate the effect of simulating the refolding of the myosin chain with the rMD algorithm.

5.2.1 Marginally thermally activated transitions

In simulations performed by Kreuzer *et al.,* the rate-limiting-step of the unfolding process is the transition $0° < \psi < 90° \rightarrow \psi > 90°$, i.e., the final step that completely opens the dihedral angle. The authors produced also a free energy landscape by using long unbiased unfolding simulations, and concluded that the rate-limiting-step just described is the only thermally activated transition [217]. All the other milestones, h-bond forming and breaking and partial opening of ψ, are thus considered only marginally thermally activated. Although we use a different FF in implicit solvent, and moreover we simulate refolding instead that unfolding, we shall assume the same separation of free energies to exists also in our model.

First of all, we want to assess the effect of the rMD simulation on a marginally thermally activated transition. We thus simulated the refolding of the myosin protein starting from a configuration where a single residue is partially unfolded. We considered two residues:

case A The initial configuration is characterized by parameters $\psi = -75.74°$, $310 = 0$, $\alpha = 0$, $\pi = 2$, corresponding to anchor π_{12}, which is partially folded.

residue number	temperature (K)	% of folded traj.'s	MFPT (ps)
Case B	290	72	8.9±0.6
	310	69	9.1±0.6
Case A	290	60	13.7±0.7
	310	71	12.6±0.6

(a) Temperature dependence. Data showed are calculated on 98 unbiased MD trajectories for each row.

Table 5.1

case B The initial configuration is characterized by parameters $\psi = 87.72°$, $310 = \alpha = \pi = 0$, and the corresponding anchor is $0° < \psi < 90°$, which is partially unfolded.

To investigate the extend of temperature effects on the refolding from this two initial configurations, we simulated the refolding trajectories with a value of the rMD constant $k = 0$ (thus an unbiased MD simulation in implicit solvent) of case A and B at the nominal temperatures of 290 and 310 K (table 5.1a). There is a rather weak scaling with temperature, suggesting that indeed we are dealing with marginally thermally activated transitions.

In order to study the effect of the bias in our algorithm, we performed the refolding simulation varying in a wide range the rMD free parameter k, which sets the strength of the biasing force. In particular, we considered 30 values lying in the interval $[4 \times 10^{-7} \div 2]$ kcal/mol. Moreover, a simulation at $k = 0$ (unbiased MD) was performed. For each value of k 96 folding trajectories were attempted. Each of this trial trajectories share the same initial configuration but has different initial velocities, that are generated randomly changing the seed. Hence, if n of these 96 trial trajectories fold in the given simulation time, then we have n independent realizations of the same refolding. A further independent set of simulations was performed for the values of k ranging from 4×10^{-5} to 1.6×10^{-2} kcal/mol. Thus, a total amount of 3168 trajectories was simulated and analyzed.

5.2.1.1 Case A (partially folded conformation)

Percentage of folded trajectories

It is important to check that most of the trajectories hit the final anchor, in order to have a sufficiently broad sample. In Fig. 5.3, red data, we can see that for all values of k, more than half of the 96 trial trajectories fold in the given time. Data seems to have a logarithmic scaling with k, and is consistent with the case $k = 0$, in which the percentage of folded trajectories is $\approx 61\%$. Almost all the trajectories fold for $k > 2 \times 10^{-2}$ kcal/mol.

Mean First Passage Time

MFPT was calculated for all values of k using both the straightforward definition and the milestone approach. In the first case, we averaged the time to reach the absorbing anchor (corresponding to the folded residue) amid the set of folded trial trajectories for each value of k. Error bars are given by standard deviation over $\sqrt{N}$, being N the number of correctly folded trial trajectories. In the second we applied the milestoning protocol to the same set of trajectories. Results can be seen in Fig. 5.4.

First of all, we can say that there is an excellent agreement between the two calculations, which witnesses that milestoning is effective and that its working hypotheses can be considered to be satisfied in rMD simulations. The latter observation is surely non trivial.

In the plot we have represented the unbiased MFPT as a blue line. The value of 13.3 ± 0.7 ps refers to a refolding from a partially folded initial configuration, and hence corresponds to the formation of several h-bonds. It is not too far from the one found in the unbiased simulation in explicit solvent reported in Ref. [217], in which the timescale for h-bond forming/breaking is ~ 10 ps. One has the biggest acceleration using $k = 2$ kcal/mol, although this gain consist at most only in a factor equal to 3. The scaling of the MFPT with k is compatible with a logarithmic one.

Max Flux Path

The MFP was calculated for all the considered 30 values of k. In 25 cases the path is $\pi_{12} \to \pi_{34} \to \pi/\alpha_1 \to \alpha_1 \to \alpha_2 \to \alpha_3$, while in the remaining 5 cases

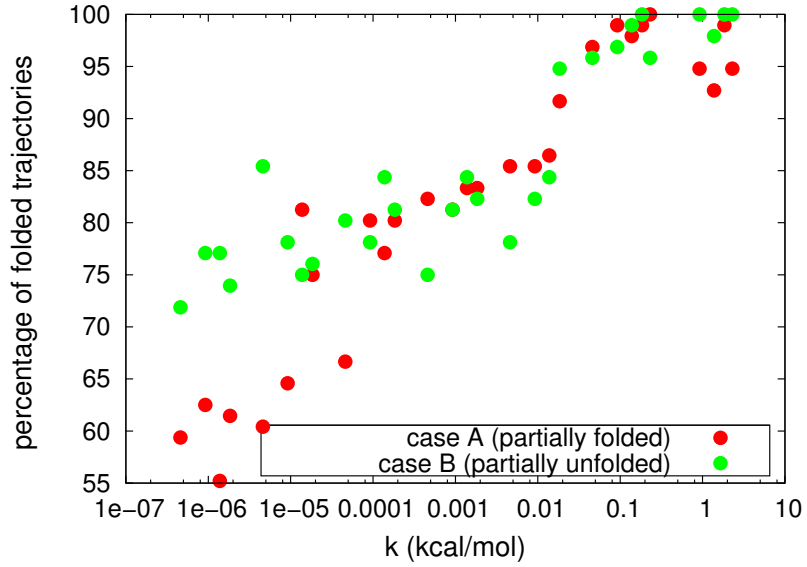


Figure 5.3: Percentage of folded trajectories in the simulation time ($\sim$ 50 ps).

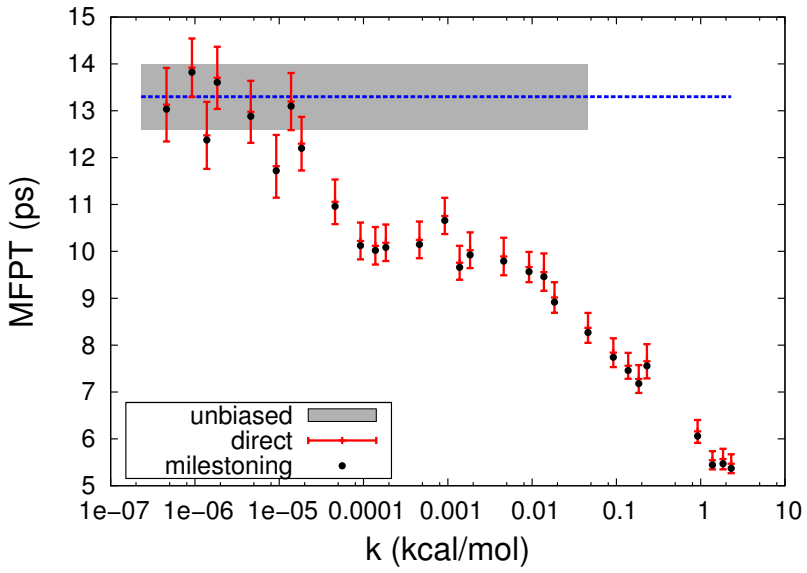
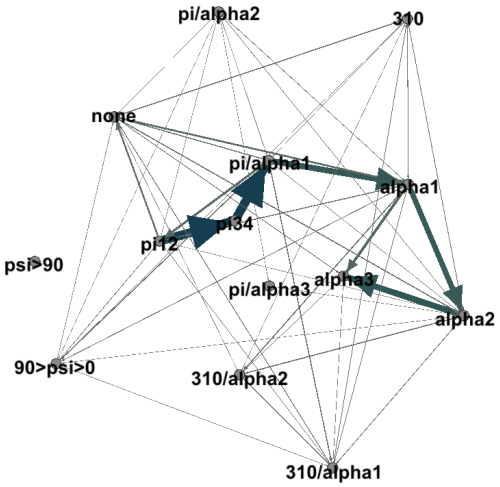
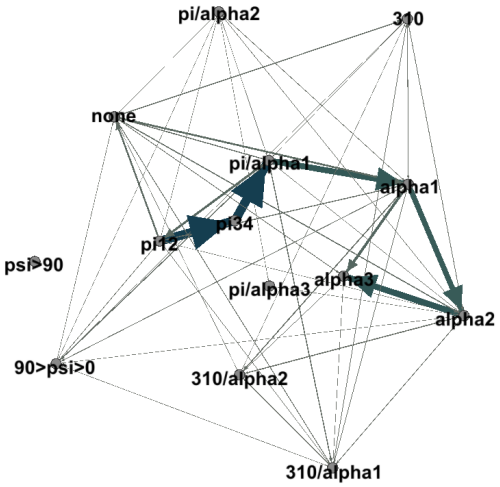


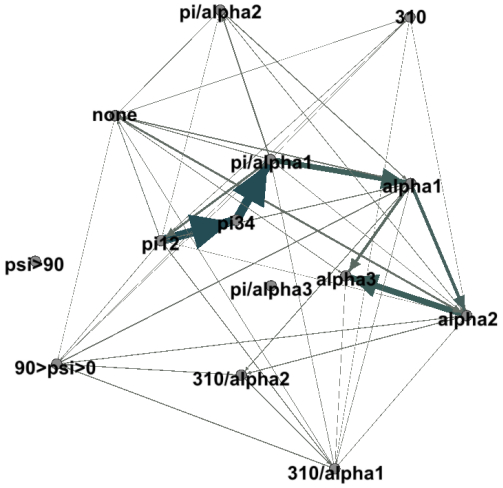
Figure 5.4: MFPT vs. k. Milestoning data (black dots) have unknown uncertainties. The blue line and grey shadowed area represent the MFPT for the unbiased case ($k = 0$) with its uncertainty.



(a) $k = 2\,\mathrm{kcal/mol}$.



(b) $k = 0.04\,\mathrm{kcal/mol}$.



(c) Unbiased simulation.

Figure 5.5: Weighted and directed networks obtained by applying the milestoning algorithm on rMD simulations. The simulated transition describes refolding of the myosin chain in the case A, i.e., with starting anchor π_{12}. The nodes of the graphs represent anchors, while the edges represent the observed connections between visited anchors. The thickness of the arrows is proportional to the net flux conveyed on that edge. The three networks describe simulations obtained by using three different values of the biasing constant k. Although minor differences exist, the networks are remarkably similar both qualitatively and quantitatively.

it is $\pi_{12} \to \pi_{34} \to \pi/\alpha_1 \to \alpha_1 \to \alpha_3$. In both cases the first transition is fixed since it is used as emitting milestone. It seems there is no trivial correlation between the two slightly different paths and the value of k. Hence, we can say that the MFP is not affected by the rMD. Fig. 5.5 shows the transition networks calculated for three values of k. It appears that the MFP and the general structure of the network are highly conserved varying the value of k.

It is interesting now to understand why the MFPT is raising with the lowering of k. It might be that the bigger k, the bigger the fraction of net flux that is conveyed by the MFP instead of wandering around thoroughly the possible states, and/or that the number of reversible transitions between anchors is suppressed by an increasing k.

Regarding the first hypothesis it can be seen in Fig. 5.6 (upper panel) that a correlation exists, although a non dramatic one. In fact, the plotted fraction depends little on the value of k, and taking into account average and standard deviation, one has 0.61 ± 0.08, which is fully compatible with the unbiased value 0.59.

Regarding the second hypothesis, we introduced the r parameter, calculated to have a measure of how much the dynamics is reversible. Given two anchors i and j and the stationary fluxes connecting them, $q_{i\to j}$ and $q_{j\to i}$, one calculates

$$r = \left| \frac{q_{i\to j} - q_{j\to i}}{q_{i\to j} + q_{j\to i}} \right|$$

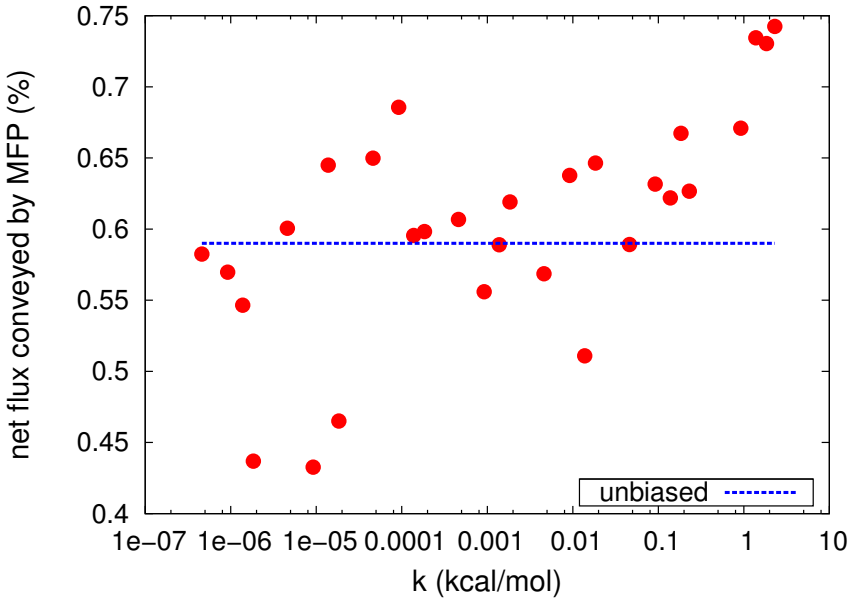
that is the absolute value of the ratio of the net stationary flux and the sum of the (gross) fluxes. This quantity ranges from 0, when the dynamics is completely reversible, to 1, when it is completely directional.

We calculated the weighted average of the r parameter over the folded trial trajectories for each k. Since we want this average to reflect the behavior of the most probable transitions, we used as weights the net stationary fluxes connecting anchors. In Fig. 5.7 (upper panel) it can be seen that again a correlation exists. Noteworthy, the value for the unbiased simulation is relatively high, i.e. ≈ 0.60.

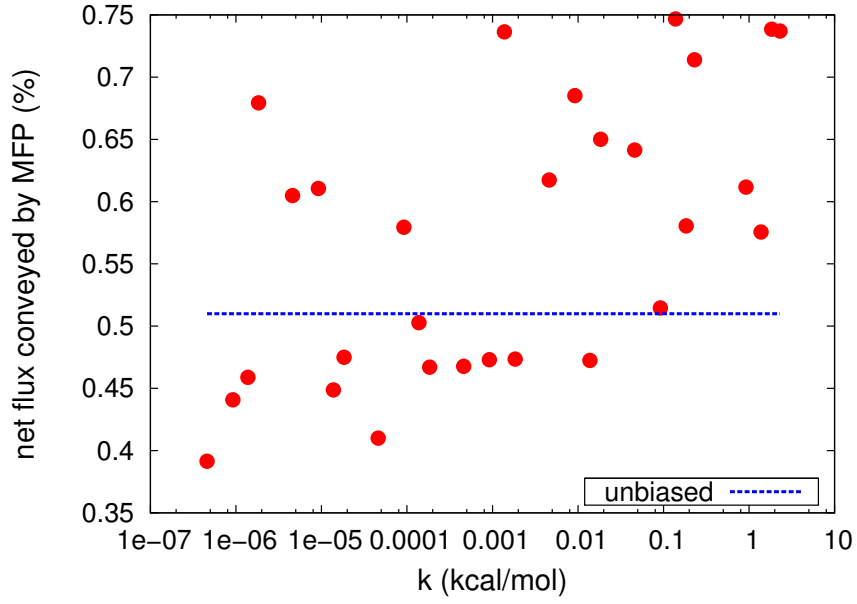
Using instead as weights the sum of the (gross) stationary flux connecting two anchors, one is actually calculating (indexes range over all anchors)

$$\langle r \rangle_2 = \frac{\sum_{i\neq j} |q_{i\to j} - q_{j\to i}|}{\sum_{i\neq j} (q_{i\to j} + q_{j\to i})},$$

which has the property to range from 0, if *all* the transitions are completely

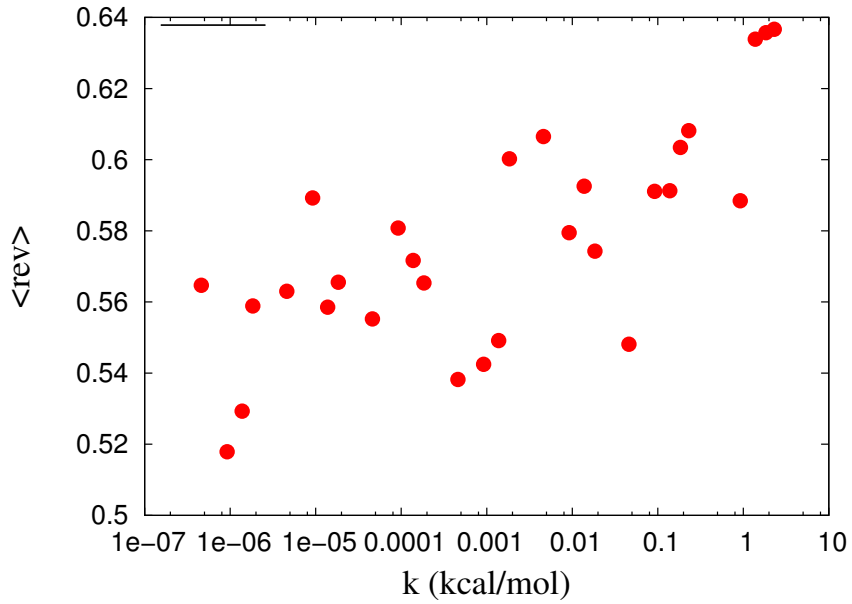


(a) Case A

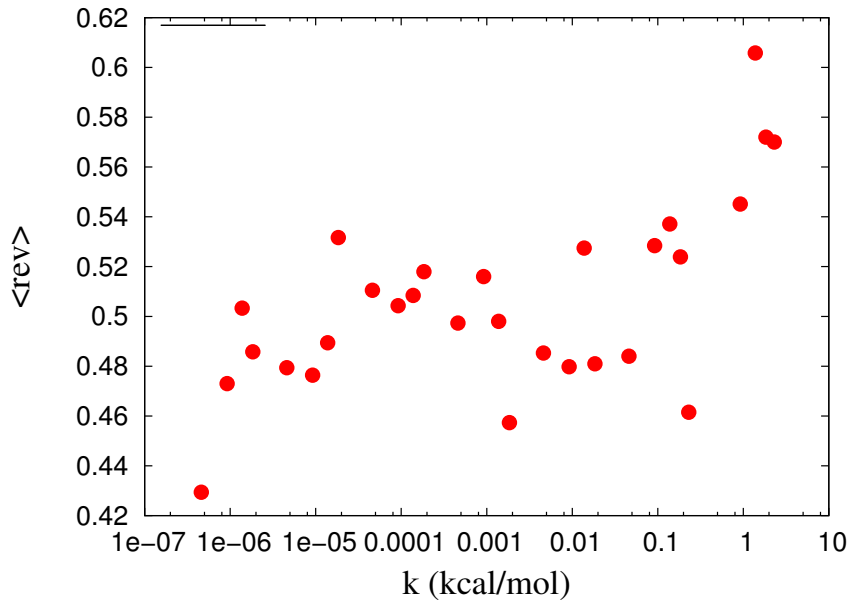


(b) Case B

Figure 5.6: Fraction of net flux conveyed by the MFP over total.



(a) Case A. Value in the unbiased case is $\langle r \rangle = 0.60$.



(b) Case B. Value in the unbiased case is $\langle r \rangle = 0.46$.

Figure 5.7: Average of r reversibility parameter vs. k. r is 0 for completely reversible transitions and 1 in the opposite case.

reversible, to 1, if *all* transitions have only one direction. The results of this second calculation is shown in Fig. 5.10 (upper panel), and it can be seen that this is more correlated to k than the previous weighted average.

Entropy

Once we have calculated the stationary probability distribution of each milestone, $\mathbf{p}$, we can calculate a Gibbs entropy of the milestones network, defined as

$$S = -\sum_{\alpha} \mathbf{p}_\alpha \ln \mathbf{p}_\alpha$$

In Fig. 5.8 the scaling of entropy vs. k can be seen. It is logarithmic up to $k = 2 \times 10^{-2}$ kcal/mol, after which it converges to a value slightly smaller than in the case $k = 0$, which acts as a sort of upper bound.

5.2.1.2 Case B (almost unfolded initial configuration)

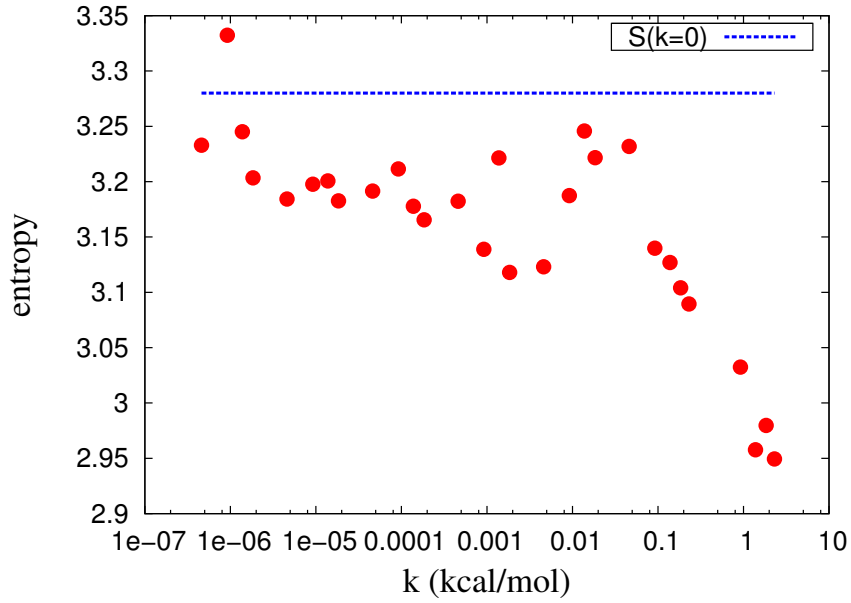
Used procedures are the same compared to the analysis on Case A, and we will report only results.

Percentage of folded trajectories

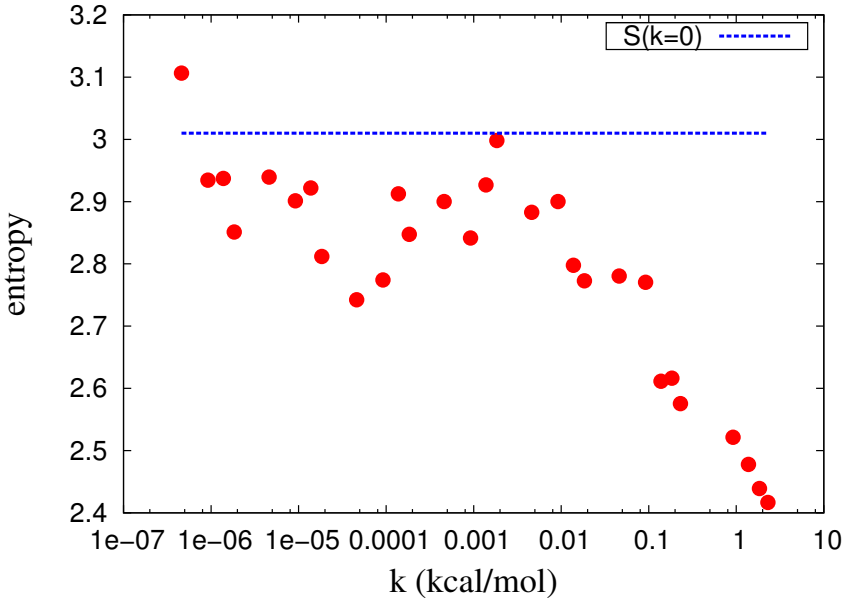
Also in this case (Fig. 5.3 green data) for all values of k in the trial trajectories majority the cracked residue refolds correctly. Again the scaling seems to be logarithmic and all trajectories fold for $k > 2{\times}10^{-2}$ kcal/mol. The percentage for $k = 0$ is $\approx 79\%$, which would indicate that the different values for the two sets of data in the region $k < 2 \times 10^{-5}$ kcal/mol is due to the different initial configurations used and not to a different behavior of the rMD.

Mean First Passage Time

Also in this case we have an excellent agreement between direct determination of MFPT and the one using milestoning (Fig. 5.9). Regarding the scaling, same considerations done for case A hold.



(a) Case A.



(b) Case B.

Figure 5.8: Gibbs entropy of the network (in k_B units).

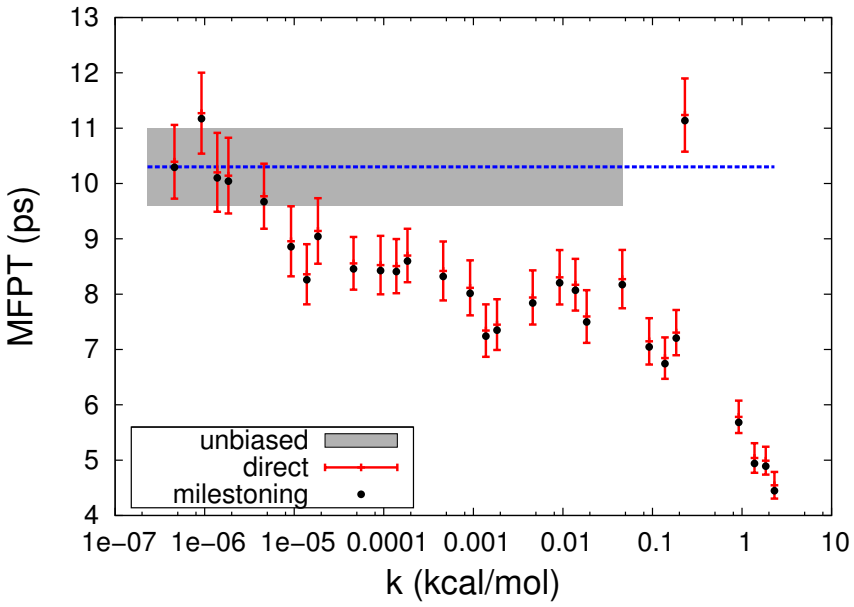


Figure 5.9: MFPT vs. k. Milestoning data (black dots) has unknown uncertainties. The blue line and grey shadowed area represent the MFPT for the unbiased case ($k = 0$) with its uncertainty.

Max Flux Path

In 11 cases the MFP is $0°<\psi<90° \rightarrow none \rightarrow \alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$, in 18 $0° < \psi < 90° \rightarrow none \rightarrow \alpha_1 \rightarrow \alpha_3$ and in only one $0° < \psi < 90° \rightarrow none \rightarrow \alpha_2 \rightarrow \alpha_3$. Furthermore, in this case only the anchor $0° < \psi < 90°$ is given as input, while the anchor *none* is found to be the second one in all the simulations. So, the MFP is not depending on k and almost constant up to small variations.

For case B the fraction of net flux conveyed by the MFP scales in a similar way as for case A (Fig. 5.6). Similar resemblance can be found in the plot for $\langle r \rangle$ calculated in both ways (Fig. 5.7 and Fig. 5.10).

Entropy

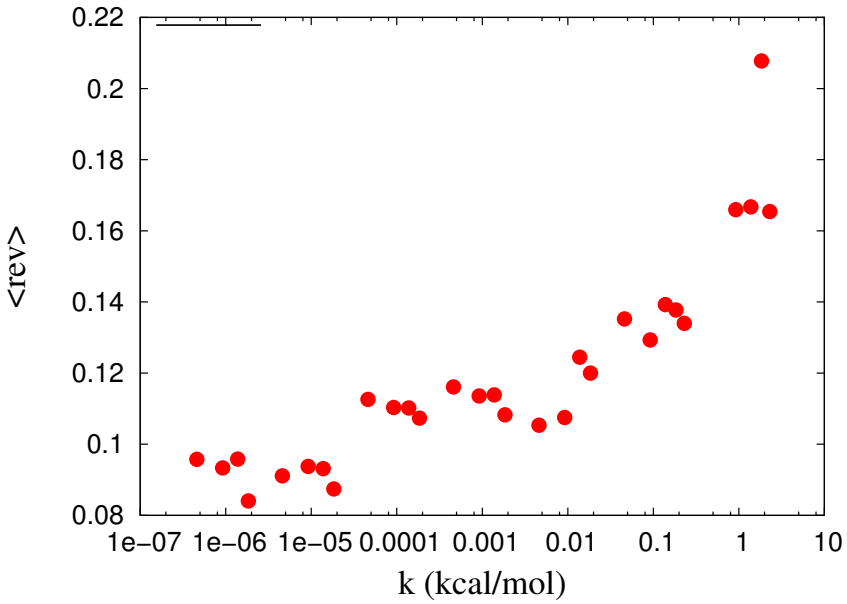
A striking similarity is relevant also for the calculation of entropy as a function of k (Fig. 5.8).

5.2.2 Thermally activated transition

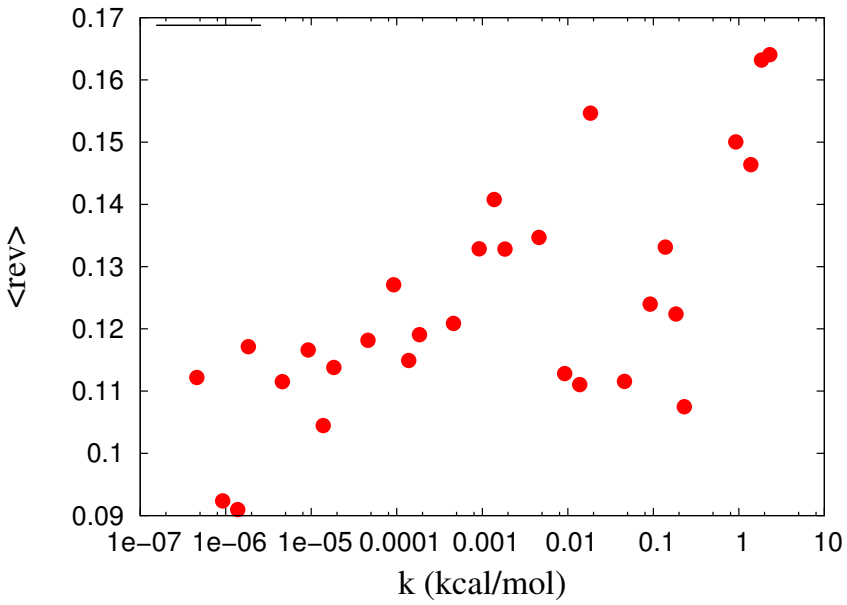
We consider now three configurations in which there is clearly an unfolded residue (i.e., the initial anchor is $\psi > 90°$). For each of them we simulated 98 refolding trajectories varying the value of k, for a total amount of several thousands of simulated and analyzed trajectories.

The important difference with respect to the previous set of simulations is represented in Fig. 5.11a, showing the percentage of folded trajectories. For values $k < 2 \times 10^{-2}$ kcal/mol the number of folded trajectories drops significantly, reaching 20%, whereas in Fig. 5.3 it is never lower than $\approx 60\%$. This means that the MFPT is much higher than in the previous case, and that the value calculated from this simulations (Fig. 5.11b) is not reliable for $k < 2 \times 10^{-2}$ kcal/mol due to the low amount of statistics. It also presumably means that rMD is not able to significantly enhance the folding efficiency with this free energy barrier if the strenght of the biasing force is too low. Nonetheless, we can repeat the analysis performed for the previous initial configurations, bearing in mind that it will be statistically significant only for values $k > 2 \times 10^{-2}$ kcal/mol .

We tried to perform an refolding unbiased MD simulations, but we did not see a significant increase in the percentage of the successful refolding

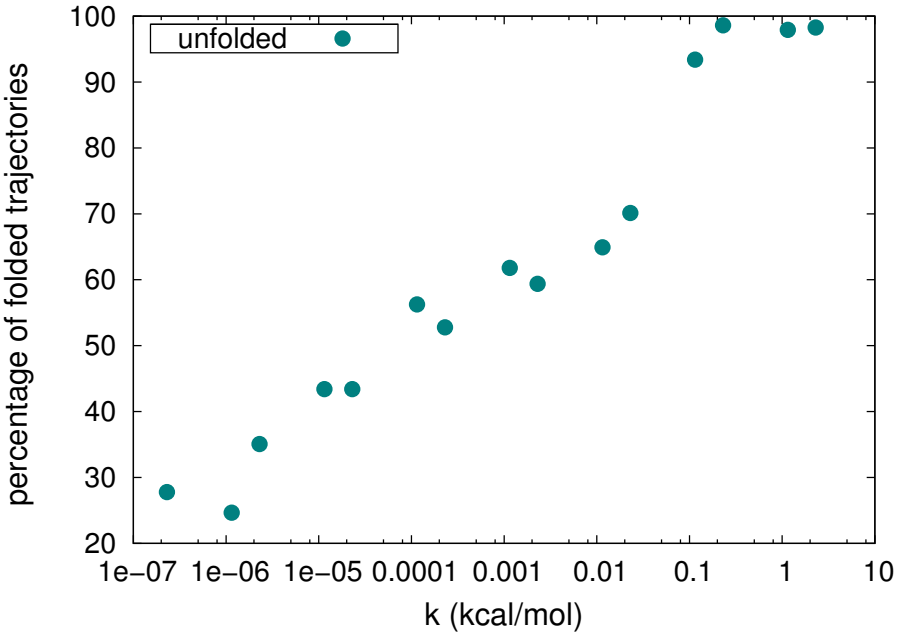


(a) Case A. For the unbiased simulation the value is ~ 0.08

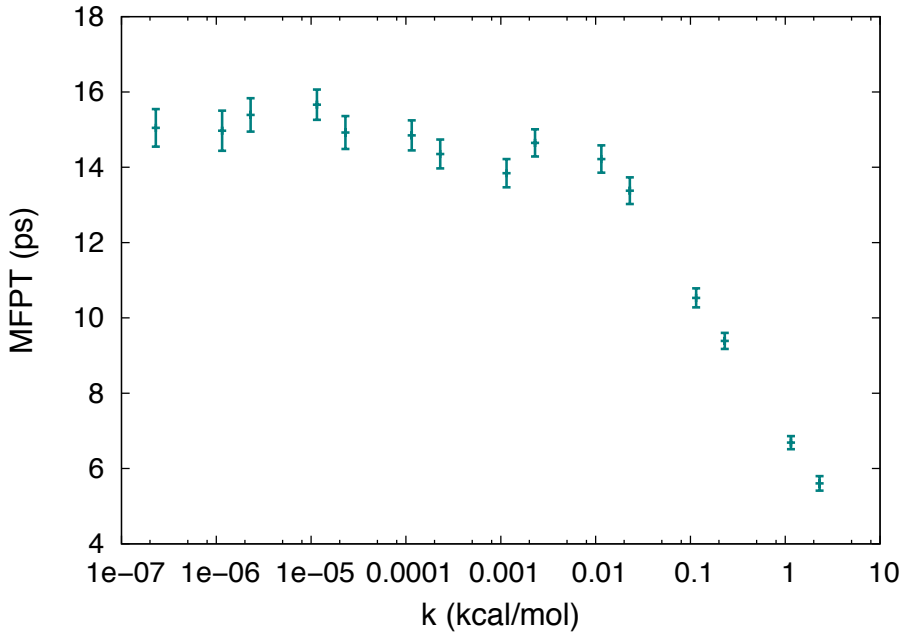


(b) Case B. For the unbiased simulation the value is ~ 0.11

Figure 5.10: Weighted average of the r reversibility measure, with used weights the sum of the (gross) flux connecting anchors.



(a) Percentage of folded trajectories.



(b) Mean first passage time for all the folded trajectories. Note that we did not succeed to calculate the unbiased MFPT value. Kreuzer *et al.* estimated this value to be ~ 100 ps, in unbiased MD with explicit solvent [217].

Figure 5.11

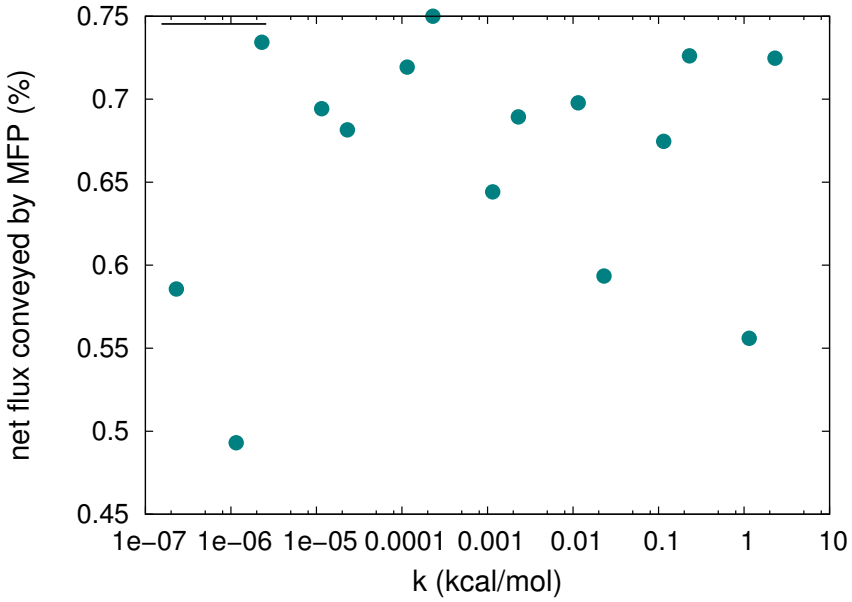
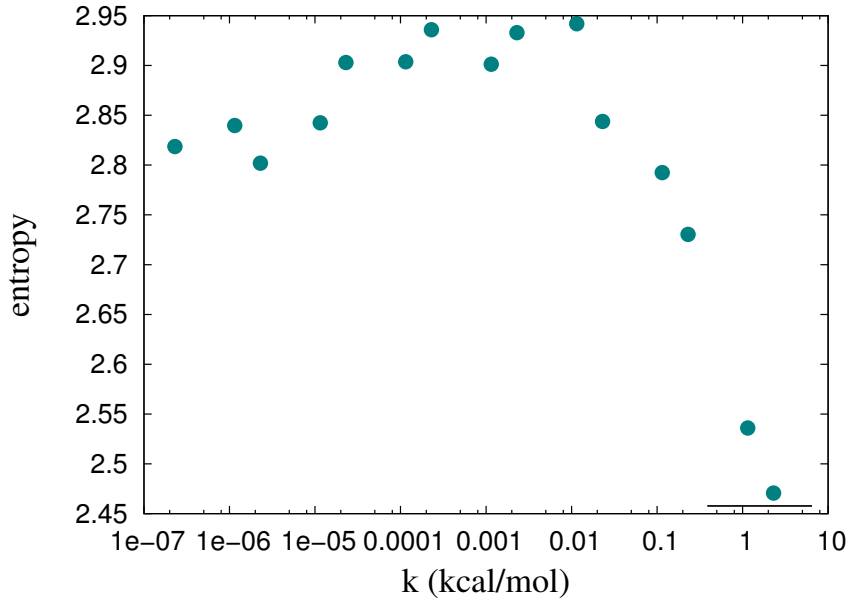


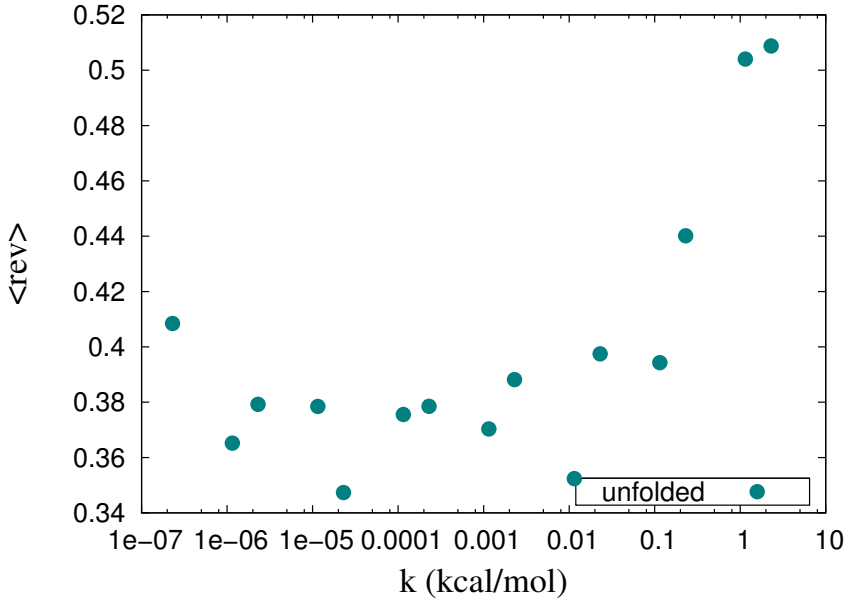
Figure 5.12: Percentage of the net probability flux conveyed by the MFP.

trajectories in several simulations up to the ns scale. This is presumably due to the FF we employed, and in particular the implicit solvent, since we checked that the native structure of the myosin chain is not stable at normal temperature.

The MFP is again very stable, since in most cases it is $\psi > 90° \rightarrow 0° < \psi < 90° \rightarrow none \rightarrow \alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$, with few variations where anchor α_2 is not visited. Excluding the different starting anchor, this MFP is identical to the one studied in marginally thermally activated transition described in case B. The percentage of net flux carried by the MFP is reported in Fig. 5.12. Notably, values are not significantly higher than those referred to the unbiased cases reported in Fig.'s 5.6, where presumably no big free energy barrier is crossed. Gibbs entropy of the milestones and the weighted average of the reversibility parameter can be seen in Fig. 5.13. Again, the statistically meaningful values (those for $k > 2 \times 10^{-2}$ eV) are not far from values referred to unbiased simulations for case A and B, reported above.



(a)



(b)

Figure 5.13

5.2.3 Discussion

We conducted an extensive study on the refolding of an unfolded or partially folded residue in a long $\alpha-$helix by means of the rMD algorithm. In order to assess the effects of the bias on the dynamics in our trajectories, we varied the coupling constant, which determines the strength of the biasing force, on seven orders of magnitude.

The milestoning algorithm yields a quantitative description of the probability fluxes flowing across a coarse grained representation of the system. In particular, we can calculate the path connecting the unfolded and folded configurations, which conveys the highest amount of net probability flux, i.e., the MFP. If the net flux in the MFP is an important fraction of the overall flux, then the MFP represents the most probable mechanism for refolding, or, as claimed in [216], a reaction coordinate of the system. Notably, in the system we analyzed the MFP is not depending on k, but is determined by the FF and the initial configuration. We can thus conclude that, in this system, the refolding mechanism is not changed by the effect of the rMD algorithm, but is exclusively determined by the FF. This conclusion is clear in the case where simulations are started from partially folded and unfolded configurations, where the refolding mechanism is not depending on k and equivalent in unbiased and biased simulations. Unfortunately, since due to problems with the FF we did not obtain unbiased simulations starting from fully unfolded configurations, the same conclusion is not as robust in simulating the thermally activated process.

The MFPT is the one of the most important observables in kinetics investigations, and the main advantage of using the rMD biasing scheme is that simulated folding times are dramatically reduced. As can be seen by the values of the MFPT in the unbiased simulations of case A and B, simulations in implicit solvent are extremely fast. Moreover, in these cases the reduction of the MFPT is weak and amounts at most to a factor 3. The scenario is rather different in the simulations that start from a fully unfolded configuration. There is a steady acceleration for relatively high values of k, but after a given threshold, which is $k < 2 \times 10^{-2}$ kcal/mol, no further gain is measured, and the estimated MFPT remains constant. At the same time, below the same threshold the fraction of successfully folded trajectories in the simulation time drops significantly. Both observations can be explained assuming that rMD stops to efficiently accelerate refolding in the $k < 2 \times 10^{-2}$ kcal/mol

regime. It is likely that, in the regime in which rMD works, the MFPT is accelerated exponentially.

For the sake of understanding why the system is lowering its MFPT as k is raised, we studied the fraction of net flux conveyed by the MFP over its total amount and the degree of reversibility of the simulations. Conclusions can be drawn only qualitatively, but both quantities are weakly dependent on k, hence it would seem that both are involved in the behavior of kinetics. In other words, the effect of the bias is to slightly concentrate the probability flux along the MFP and reduce the reversibility of transitions between anchors.

As a final remark, we found that in the FF and implicit solvent we employed, conformations displaying π h-bonds are not stable.

5.2.4 Computational details

Simulations were performed using DOLOMIT, the software written by our lab that calls GROMACS 4.5.2 [167] as an external library to calculate forces. The all atom force field Amber ff99SB [56] was used, along with the generalized Born OBC implicit solvent model [54]. Nominal temperature and time step of respectively 300 K and 1 fs were used. Frames in the output pdb file have been saved every 100 fs. Simulation time was 9 hours for the biggest values of k and 12 for all the others on 48 CPUs, corresponding to ~ 50 ps. All the simulations were performed on the Aurora cluster. We gratefully thank Steven Kreuzer, who provided us with the unfolded myosin conformations obtained by unbiased MD simulations [217, 218].

Conclusions

We introduced and validated the DRP method, which is an approximate numerical method that yields an atomistic detailed description of the protein folding mechanism by employing computer simulations. In order to identify statistically significant folding paths, provided that the unfolded and folded configurations are known, the method consists of two steps: first, we can efficiently produce many microscopic folding trajectories thanks to a biased sampling protocol; second, we rank them according to their statistical significance in a unbiased diffusive dynamics.

We investigated a small WW domain, which displays a protein-like folding widely characterized during the last years. We produced a set of statistically significant folding trajectories at atomistic resolution by means of the DRP algorithm in a realistic force field. The analysis of the simulated trajectories reveals the existence of two dominant folding pathways, defined by the sequential formation of two $\beta-$hairpins, which are the structural sub-motifs of the WW domain. Indeed, the most probable mechanism displays a formation of the first hairpin followed by the folding of the second one, whereas in the less probable pathway the hairpin formation is reversed. This result is compatible with both experimental investigations and unbiased MD simulations on the millisecond scale, showing spontaneous folding and unfolding. We then calculated the free energy for the same molecule by employing simplified native-centric models, which confirm the existence of the same two folding mechanisms. The free energy landscape remains qualitatively unaltered by further considering attractive non-native interactions, thus suggesting the folding mechanisms to be completely determined by native interactions only. We stress the high computational efficiency of our approach, since producing the set of statistically significant trajectories took only two days of calculations on 48 CPU's.

As second study we investigated the folding of a protein which displays a knot in its native conformation by applying the DRP computational scheme. We obtained the first set of atomistic folding trajectories of a knotted protein in a realistic force field ever reported in literature. These trajectories permit to study the concerted conformational changes leading to the formation of the knot starting from a fully unfolded configuration. By analyzing them, we showed that folding is rather homogeneous, mostly following just one pathway defined by a specific sequence of secondary structure formations. The vast majority of knotting events occur by directly threading the molecule C-terminus through a native loop, whereas a far less frequent mechanism is due to slipknotting, i.e., the terminus threading the same loop in a hooked conformation. Both mechanisms were already reported in several studies by different authors. A comparison with the folding described in the same simplified models mentioned above shows that slipknotting is the favored knotting mechanism when only native interactions are taken into account. On the other hand, turning non-native interactions on not only strongly promotes a direct threading of the loop, but also enhances the probability to correctly form the knot.

Although general conclusions about folding cannot be drawn from few examples, we can sketch the picture emerging from the two cases studied in this thesis. In both proteins, folding is a homogeneous process, since most microscopic trajectories effectively describe a well defined sequence of events. Non-native attractive interactions play a constructive role in folding to a non-trivial topology, as in the case of the knotted protein, by facilitating knotting and determining the specific mechanism. On the contrary, frustration has a vanishing role in the folding pathways of the simple WW domain.

The trajectories we used are very short and far from equilibrium, nonetheless we showed that the folding mechanisms we found are plausible and, in the case of the WW domain, compatible with experiments and long unbiased MD simulations. We interpret this as an evidence that the coordinate along which we bias our simulations, i.e., the contact map distance to the native conformation, is a satisfactory reaction coordinate for folding. This seems reasonable, since this coordinate is clearly related to the fraction of native contacts. A preliminary investigation in a simplified native-centric model has indeed shown that the mapping is linear [221], but it would be useful to further look into this point.

The DRP algorithm we discussed in this thesis displays several open issues

and important drawbacks, and we will briefly outline the most significant ones in the following.

When sampling the probability to fold to the native configuration in a given time, the initial configuration should be selected according to the equilibrium Boltzmann distribution. However, calculating the latter in the unfolded basin of a protein is not a trivial task since a huge number of possible conformations has to be sampled and the result could be biased by inaccuracies of the force fields. In our studies, we used initial unfolded configurations obtained by short MD simulations at high temperature, followed by a relaxation at normal conditions. In general, for very short relaxation times, such a set of configurations is not expected to be distributed according to the correct equilibrium probability density. A first qualitative assessment of the systematic error introduced by this choice of unfolded configurations can be found in a paper in preparation by Cazzolli *et al.,* who investigated the folding of immunity proteins IM7 and IM9 within the DRP approach [222]. They employed denatured initial configurations obtained both by high-temperature unfolding and by constructing random-coils with knowledge-based parameters. The atomistic trajectories connecting these two sets of denatured configurations to the native one describe qualitatively overlapping folding mechanisms.

All the results discussed in this thesis were obtained performing MD simulations in implicit solvent. This is a natural choice, since the stochastic action formalism, which we employed to rank and select the most significant trajectories, is based on the assumption that the solvent degrees of freedom can be averaged out. On the other hand, it is well known that explicit solvent models represents more accurately the interactions between water and the protein and returns more reliable simulated time scales. Although the rMD sampling protocol can be easily used with explicit water, it is not obvious how to adapt the stochastic action functional to take into account an explicit solvent representation. Two possible solutions are trying to "remove on average" (i.e., renormalize) the effect of simulated water from the Onsager-Machlup action functional, and using a hybrid solvent model with an explicit description of the solvation shell and an implicit one for the bulk.

The most severe drawback is for sure the loss of physical time durations measured along the folding trajectories, due to the biasing potential of the rMD algorithm. Restoring a physical time scale would be a fundamental improvement of our method, indispensable to obtain more quantitative in-

formation from the statistically significant folding trajectories. For example, it would be possible to get insights in folding kinetics, that is one of the most important observables in experimental studies. Therefore, we have conducted some attempts to re-weight the biased trajectories and estimate the needed time rescaling factor. Unfortunately, the problem has proved to be technically very hard, and although we have found some interesting results, it stands still unsolved.

In this thesis we used the DRP scheme to specifically investigate the folding of proteins, but this technique is valid to characterize any kind of molecular conformational transition. In particular, it can be a valuable choice to study rare transitions in big molecules, when an unbiased MD simulation is unfeasible even on the most powerful computers. Following this direction, Cazzolli *et al.* have undertaken an atomic-level characterization of a rare conformational transition in serpins, an enzyme family of great biological interest [223].

We discussed the first application of the DRP method to realistic models of proteins and validated the results against more standard and experimental techniques. Much work has still to be done in order to gain more solid theoretical control and validation. Nevertheless, we think that these first encouraging results suggest that the DRP method can be a valid alternative to qualitatively characterize protein folding and rare conformational transitions in biological molecules at a rather low computational cost.

Credits

My daily work and what shown in this thesis would hardly be possible if I could not take advantage of free and open software, and shared knowledge instruments on the web. Thus, a special thanks and credits go to the Community and all those diverse people who created, maintain and make freely available these precious instruments: the molecular visual software VMD; the Molecular Dynamics simulation suite GROMACS; the wonderful Python and Scipy programming language. Moreover, this thesis was written using LyX, which is based on LaTeX. Images were processed with Gimp and Inkscape, and all the job was done on Ubuntu systems. And finally, a sincere thanks for their great and passionate job to the Wikipedia Community and the StackOverflow.com people.

Acknowledgments

My first and most sincere thanks goes with no hesitation to my supervisor, Pietro. I came to him with a completely different (and maybe useless) background than biological physics, and he taught me so much in such a short time. Most importantly, he instilled in me the love for physics and science "che se magna", which makes doing research so much funnier. Working with him is always a sort of thrill, sometimes maybe more than necessary. The two of us are rather different persons, but we found a good way get along. I am so grateful for his passion, enthusiasm and great humanity, for they fostered me as a scientist and human being. He and Lidia are to me a life example to follow, and I hope to be able to share their friendship also in future.

I often use to say that a highly developed skill is basically indistinguishable from magic. This is exactly what I thought the first time I saw Silvio a Beccara and Enrico Tagliavini in front a computer. With patience and generosity, they both taught me how to use a computer not just to read emails and surf the web. I have learned much more by reverse engineering their codes than by reading any manual.

I had the luck to collaborate with Prof. Ron Elber and Prof. Cristian Micheletti, and I want to thank them for being so kind with me and share a little part of their insights.

Tatjana Skrbic is to me a genuine example of unbounded dedication to physics. Besides, on a more funny side, she is an example of an apparently unlimited power to screw up any computer or technological device.

The results presented in this thesis is an inherently group effort, thus I really want to thank Tatjana, Silvio, Cristian and Pietro for the work done together and the fun in doing it.

I cannot thank enough my PhD mates, for their infinite patience and

generosity, and because we had the utmost luck to become real friends. I will always remember the stimulating discussions on science and life, which enriched the long hours spent closed in our office. Together with many more friends and mates met during this three years, we had the funniest days and wildest nights. I do hope that for all of them the time we spent together has been as much as happy as for me. Our common path is now over, and we will soon part and spread over Europe and the World. Our friendship will change, but no nostalgia will ever be able to suppress the consciousness of the privilege of having met, and the confidence that whenever we will meet again, it will be like we never parted.

At last, my endless gratitude goes to Nicolò, who extensively proofread the draft of the thesis and adjusted my English. And besides, because he makes my life so much happier. We made it so far, we will make in future.

Meglio aggiungere vita ai giorni che non giorni alla vita.

Rita Levi Montalcini

Bibliography

[1] P. W. Anderson, Science **177**, 393 (1972).

[2] Finkelstein and O. B. Ptitsyn, *Protein Physics* (Academic Press, 2002).

[3] K. Huang, *Lectures On Statistical Physics And Protein Folding*, illustrate ed. (World Scientific Pub Co Inc, 2005).

[4] P. Echenique, Contemporary Physics **48**, 53 (2007), arXiv:0705.1845 .

[5] L. Brocchieri and S. Karlin, Nucleic acids research **33**, 3390 (2005).

[6] X.-Z. Li, B. Walker, and a. Michaelides, Proceedings of the National Academy of Sciences **108**, 6369 (2011).

[7] R. E. Hubbard and M. K. Haider, eLS **1**, 1 (2001).

[8] F. Crick, *What Mad Pursuit: A Personal View of Scientific Discovery* (Basic Books, New York, 1988).

[9] F. Crick, Nature **227**, 561 (1970).

[10] D. V. Fedyukina and S. Cavagnero, Annual review of biophysics **40**, 337 (2011).

[11] H. Krobath, E. I. Shakhnovich, and P. F. N. Faísca, The Journal of Chemical Physics **138**, 215101 (2013).

[12] C. B. Anfinsen, Science (New York, N.Y.) **181**, 223 (1973).

[13] A. H. Elcock, Current opinion in structural biology **20**, 196 (2010).

[14] F. U. Hartl and M. Hayer-Hartl, Nature structural & molecular biology **16**, 574 (2009).

[15] A. R. Fersht and V. Daggett, Cell **108**, 573 (2002).

[16] M. Karplus, Folding and Design , 69 (1997).

[17] K. a. Dill and J. L. MacCallum, Science (New York, N.Y.) **338**, 1042 (2012).

[18] T. R. Sosnick and D. Barrick, Current opinion in structural biology **21**, 12 (2011).

[19] Y. Arinaminpathy, E. Khurana, D. M. Engelman, and M. B. Gerstein, Drug discovery today **14**, 1130 (2009).

[20] J. L. MacCallum and D. P. Tieleman, Trends in biochemical sciences **36**, 653 (2011).

[21] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, Annual review of biophysics **37**, 215 (2008).

[22] S. E. Jackson and a. R. Fersht, Biochemistry **30**, 10428 (1991).

[23] B. Gillespie and K. W. Plaxco, Annual review of biochemistry **73**, 837 (2004).

[24] T. Lazaridis and M. Karplus, Biophysical chemistry **100**, 367 (2003).

[25] R. A. Goldstein, Proteins **79**, 1396 (2011).

[26] J. M. Sanchez-Ruiz, Biophysical chemistry **148**, 1 (2010).

[27] R. B. Dyer, Current opinion in structural biology **17**, 38 (2007).

[28] K. W. Plaxco, K. T. Simons, and D. Baker, Journal of molecular biology **277**, 985 (1998).

[29] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, Biochemistry **39**, 11177 (2000).

[30] V. Daggett and A. Fersht, Nature reviews. Molecular cell biology **4**, 497 (2003).

[31] C. Levinthal, in *Mössbaun Spectroscopy in Biological Systems Proceedings*, Vol. 24 (1969) pp. 22–24.

[32] R. Zwanzig, A. Szabo, and B. Bagchi, Proceedings of the National Academy of Sciences **89**, 20 (1992).

[33] R. Zwanzig, Proceedings of the National Academy of Sciences of the United States of America **92**, 9801 (1995).

[34] P. G. Wolynes, J. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[35] J. D. Bryngelson and P. G. Wolynes, The Journal of Physical Chemistry **93**, 6902 (1989).

[36] C. Levinthal, Journal de Chimie Physique et de Physico-Chimie Biologique **65**, 44 (1968).

[37] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Structure, Function, and Bioinformatics **21**, 53 (1994), arXiv:9411008 [chem-ph] .

[38] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annual review of physical chemistry **48**, 545 (1997).

[39] S. S. Plotkin and J. N. Onuchic, *Quarterly reviews of biophysics*, Vol. 35 (2002) pp. 205–86.

[40] J. N. Onuchic and P. G. Wolynes, Current opinion in structural biology **14**, 70 (2004).

[41] M. Mezard, G. Parisi, M. A. Virasoro, and D. J. Thouless, Physics Today **41**, 109 (1988).

[42] K. H. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, Cambridge, 1991).

[43] P. W. Anderson, Journal of the Less Common Metals **62**, 291 (1978).

[44] E. Shakhnovich, G. Farztdinov, A. Gutin, and M. Karplus, Physical Review Letters **67**, 1665 (1991).

[45] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins **47**, 506 (2002).

[46] N. Go and H. Taketomi, Proceedings of the National Academy of Sciences of the United States of America **75**, 559 (1978).

[47] K. A. Dill and H. S. Chan, Nature Structural Biology **4**, 10 (1997).

[48] M. Karplus, Nature chemical biology **7**, 401 (2011).

[49] D. J. Bicout and A. Szabo, Protein science : a publication of the Protein Society **9**, 452 (2000).

[50] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).

[51] P. H. Hünenberger, Advanced Computer Simulation , 105 (2005).

[52] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of Chemical Physics **79**, 926 (1983).

[53] M. Feig and C. L. Brooks, Current opinion in structural biology **14**, 217 (2004).

[54] A. Onufriev, D. Bashford, and D. A. Case, Proteins **55**, 383 (2004).

[55] A. D. Mackerell, Journal of computational chemistry **25**, 1584 (2004).

[56] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins **78**, 1950 (2010).

[57] R. B. Best and J. Mittal, The journal of physical chemistry. B **114**, 8790 (2010).

[58] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, PloS one **7**, e32131 (2012).

[59] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Biophysical journal **100**, L47 (2011).

[60] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Nature physics **6**, 751 (2010).

[61] C. Clementi, Current opinion in structural biology **18**, 10 (2008).

[62] H. S. Chan, Z. Zhang, S. Wallin, and Z. Liu, Annual review of physical chemistry **62**, 301 (2011).

[63] S. Takada, Current opinion in structural biology **22**, 130 (2012).

[64] V. G. Contessoto, D. T. Lima, R. J. Oliveira, A. T. Bruni, J. Chahine, and V. B. P. Leite, Proteins **81**, 1727 (2013).

[65] A. N. Adhikari, K. F. Freed, and T. R. Sosnick, Physical Review Letters **111**, 028103 (2013).

[66] G. E. Moore, Proceedings of the IEEE **86**, 82 (1998).

[67] M. Vendruscolo and C. M. Dobson, Current biology : CB **21**, R68 (2011).

[68] J. A. McCammon, B. R. Gelin, and M. Karplus, Nature **267**, 585 (1977).

[69] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. a. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science (New York, N.Y.) **330**, 341 (2010).

[70] Y. Duan and P. a. Kollman, Science **282**, 740 (1998).

[71] R. B. Best, Current opinion in structural biology **22**, 52 (2012).

[72] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, Annual review of biophysics **41**, 429 (2012).

[73] D. E. Shaw, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, R. O. Dror, S. Piana, Y. Shan, B. Towles, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, and B. Batson, in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*, c (ACM Press, New York, New York, USA, 2009) p. 1.

[74] T. J. Lane, D. Shukla, K. a. Beauchamp, and V. S. Pande, Current Opinion in Structural Biology , 1 (2012).

[75] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science (New York, N.Y.) **334**, 517 (2011).

[76] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw, Journal of the American Chemical Society **134**, 3787 (2012).

[77] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Proceedings of the National Academy of Sciences of the United States of America **109**, 17845 (2012).

[78] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Proceedings of the National Academy of Sciences of the United States of America **110**, 5915 (2013).

[79] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, The journal of physical chemistry. B (2013), 10.1021/jp4020993.

[80] R. B. Best and J. Mittal, Proteins **79**, 1318 (2011).

[81] G. R. Bowman, V. a. Voelz, and V. S. Pande, Current opinion in structural biology **21**, 4 (2011).

[82] A. a. Nickson, B. G. Wensley, and J. Clarke, Current opinion in structural biology **23**, 66 (2013).

[83] E. Haglund, M. O. Lindberg, and M. Oliveberg, The Journal of biological chemistry **283**, 27904 (2008).

[84] B. G. Wensley, S. Batey, F. a. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke, Nature **463**, 685 (2010).

[85] E. Paci, M. Vendruscolo, and M. Karplus, Proteins: Structure, Function, and Bioinformatics **47**, 379 (2002).

[86] A. R. Viguera, C. Vega, and L. Serrano, Proceedings of the National Academy of Sciences of the United States of America **99**, 5349 (2002).

[87] C. Clementi and S. S. Plotkin, Protein Science **13**, 1750 (2004).

[88] A. Zarrine-Afsar, S. Wallin, a. M. Neculai, P. Neudecker, P. L. Howell, A. R. Davidson, and H. S. Chan, Proceedings of the National Academy of Sciences of the United States of America **105**, 9999 (2008).

[89] A. Azia and Y. Levy, Journal of molecular biology **393**, 527 (2009).

[90] B. C. Gin, J. P. Garrahan, and P. L. Geissler, Journal of molecular biology **392**, 1303 (2009).

[91] P. F. N. Faísca, A. Nunes, R. D. M. Travasso, and E. I. Shakhnovich, Protein science : a publication of the Protein Society **19**, 2196 (2010).

[92] R. J. Oliveira, P. C. Whitford, J. Chahine, J. Wang, J. N. Onuchic, and V. B. P. Leite, Biophysical journal **99**, 600 (2010).

[93] R. B. Best, G. Hummer, and W. A. Eaton, Proceedings of the National Academy of Sciences of the United States of America **110**, 17874 (2013).

[94] S. S. Plotkin, **345**, 337 (2001).

[95] R. B. Best and G. Hummer, Proceedings of the National Academy of Sciences of the United States of America **107**, 1088 (2010).

[96] Z. Zhang and H. S. Chan, Proceedings of the National Academy of Sciences of the United States of America **107**, 2920 (2010).

[97] B. G. Wensley, L. G. Kwa, S. L. Shammas, J. M. Rogers, S. Browning, Z. Yang, and J. Clarke, Proceedings of the National Academy of Sciences of the United States of America **109**, 17795 (2012).

[98] B. G. Wensley, L. G. Kwa, S. L. Shammas, J. M. Rogers, and J. Clarke, Journal of molecular biology **423**, 273 (2012).

[99] A. Borgia, B. G. Wensley, A. Soranno, D. Nettels, M. B. Borgia, A. Hoffmann, S. H. Pfeil, E. a. Lipman, J. Clarke, and B. Schuler, Nature communications **3**, 1195 (2012).

[100] R. B. Best, The journal of physical chemistry. B **117**, 13235 (2013).

[101] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, USA, 2001).

[102] D. Tong, "Lectures on Theoretical Physics," .

[103] M. Chaichian and A. Demichev, *Path Integrals in Physics: Volume I Stochastic Processes and Quantum Mechanics*, Institute of physics series in mathematical and computational physics (Taylor & Francis, 2001).

[104] L. Onsager and S. Machlup, Physical Review **91**, 1505 (1953).

[105] A. B. Adib, The journal of physical chemistry. B **112**, 5910 (2008).

[106] R. B. Best and G. Hummer, Proceedings of the National Academy of Sciences of the United States of America **102**, 6732 (2005).

[107] R. B. Best and G. Hummer, Physical Review Letters **96**, 228104 (2006).

[108] R. B. Best and G. Hummer, Physical chemistry chemical physics : PCCP **13**, 16902 (2011).

[109] H. A. Kramers, Physica **7**, 284 (1940).

[110] P. Hänggi, P. Talkner, and M. Borkovec, Reviews of Modern Physics **62** (1990).

[111] H. S. Chung, K. McHale, J. M. Louis, and W. a. Eaton, Science **335**, 981 (2012).

[112] A. Szabo, K. Schulten, and Z. Schulten, The Journal of Chemical Physics , 4350 (1980).

[113] H. S. Chung and W. a. Eaton, Nature (2013), 10.1038/nature12649.

[114] C. Hartmann, R. Banisch, M. Sarich, T. Badowski, K.-z. Zentrum, and C. Schütte, Submitted to Entropy , 1 (2013).

[115] G. Mazzola, S. a Beccara, P. Faccioli, and H. Orland, J Chem Phys **134**, 164109 (2011).

[116] S. a Beccara, G. Garberoglio, P. Faccioli, and F. Pederiva, The Journal of chemical physics **132**, 111102 (2010).

[117] S. a Beccara, P. Faccioli, M. Sega, F. Pederiva, G. Garberoglio, and H. Orland, The Journal of chemical physics **134**, 024501 (2011).

[118] E. Autieri, P. Faccioli, M. Sega, F. Pederiva, and H. Orland, The Journal of chemical physics **130**, 064106 (2009).

[119] O. Corradini, P. Faccioli, and H. Orland, Phys Rev E Stat Nonlin Soft Matter Phys **80**, 61112 (2009).

[120] P. Faccioli, M. Sega, F. Pederiva, and H. Orland, Phys Rev Lett **97**, 108101 (2006).

[121] M. Sega, P. Faccioli, F. Pederiva, G. Garberoglio, and H. Orland, Phys Rev Lett **99**, 118102 (2007).

[122] P. Faccioli, The journal of physical chemistry. B **112**, 13756 (2008).

[123] P. Faccioli, The Journal of chemical physics **133**, 164106 (2010).

[124] P. Faccioli, A. Lonardi, and H. Orland, The Journal of chemical physics **133**, 045104 (2010).

[125] P. Faccioli, Journal of Physics: Conference Series **336**, 012030 (2011), arXiv:1108.5074 .

[126] R. Elber and D. Shalloway, The Journal of Chemical Physics **112**, 5539 (2000).

[127] A. E. Cárdenas and R. Elber, Proteins **51**, 245 (2003).

[128] A. Ghosh, R. Elber, and H. a. Scheraga, Proceedings of the National Academy of Sciences of the United States of America **99**, 10394 (2002).

[129] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, Physical Review Letters **97**, 170201 (2006).

[130] M. Marchi and P. Ballone, The Journal of chemical physics **110**, 3697 (1999).

[131] E. Paci and M. Karplus, J Mol Biol **288**, 441 (1999).

[132] E. Paci and M. Karplus, Proceedings of the National Academy of Sciences of the United States of America **97**, 6521 (2000).

[133] M. Bonomi, F. L. Gervasio, G. Tiana, D. Provasi, R. a. Broglia, and M. Parrinello, Biophysical journal **93**, 2813 (2007).

[134] C. Camilloni, R. a. Broglia, and G. Tiana, The Journal of chemical physics **134**, 045105 (2011).

[135] P. O. Heidarsson, I. Valpapuram, C. Camilloni, A. Imparato, G. Tiana, F. M. Poulsen, B. B. Kragelund, and C. Cecconi, Journal of the American Chemical Society **134**, 17068 (2012).

[136] G. Tiana and C. Camilloni, The Journal of chemical physics **137**, 235101 (2012).

[137] B. Isralewitz, M. Gao, and K. Schulten, Curr Opin Struct Biol **11**, 224 (2001).

[138] S. A Beccara, T. Škrbić, R. Covino, and P. Faccioli, Proceedings of the National Academy of Sciences of the United States of America **109**, 2330 (2012).

[139] F. Liu, D. Du, A. a. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele, Proceedings of the National Academy of Sciences of the United States of America **105**, 2369 (2008).

[140] M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly, and M. Gruebele, Journal of molecular biology **311**, 373 (2001).

[141] F. Liu, M. Nakaema, and M. Gruebele, The Journal of chemical physics **131**, 195101 (2009).

[142] D. L. Ensign and V. S. Pande, Biophysical journal **96**, L53 (2009).

[143] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, Biophysical journal **94**, L75 (2008).

[144] J. Juraszek and P. G. Bolhuis, Biophysical journal **98**, 646 (2010).

[145] J. Karanicolas and C. L. Brooks, Proceedings of the National Academy of Sciences of the United States of America **100**, 3954 (2003).

[146] E. H. Kellogg, O. F. Lange, and D. Baker, The journal of physical chemistry. B **116**, 11405 (2012).

[147] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, Journal of the American Chemical Society **133**, 18413 (2011).

[148] G. G. Maisuradze, R. Zhou, A. Liwo, Y. Xiao, and H. a. Scheraga, Journal of molecular biology **420**, 350 (2012).

[149] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proceedings of the National Academy of Sciences of the United States of America **106**, 19011 (2009).

[150] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, and D. E. Shaw, Journal of molecular biology **405**, 43 (2011).

[151] T. R. Weikl, Biophysical journal **94**, 929 (2008).

[152] J. Xu, L. Huang, and E. I. Shakhnovich, Proteins **79**, 1704 (2011).

[153] S. V. Krivov, The journal of physical chemistry. B , 6 (2011).

[154] G. Berezovska, D. Prada-Gracia, and F. Rao, The Journal of chemical physics **139**, 035102 (2013).

[155] P. Ferrara and A. Caflisch, Proceedings of the National Academy of Sciences of the United States of America **97**, 10780 (2000).

[156] J. Karanicolas and C. L. Brooks, Proc Natl Acad Sci U S A **101**, 3432 (2004).

[157] D. K. Klimov and D. Thirumalai, Journal of molecular biology **353**, 1171 (2005).

[158] J. Karanicolas, C. L. Brooks, and C. L. Brooks III, Protein Sci **11**, 2351 (2002).

[159] Y. C. Kim and G. Hummer, Journal of molecular biology **375**, 1416 (2008).

[160] A. Matouschek, J. T. Kellis, L. Serrano, and A. R. Fersht, Nature (1989).

[161] A. Fersht, *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding* (Macmillan, 1999).

[162] C. Merlo, K. a. Dill, and T. R. Weikl, Proceedings of the National Academy of Sciences of the United States of America **102**, 10171 (2005).

[163] T. R. Weikl, Communications in Computational Physics **7**, 283 (2009).

[164] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, The Journal of chemical physics **122**, 184901 (2005).

[165] G. Settanni, F. Rao, and A. Caflisch, Proceedings of the National Academy of Sciences of the United States of America **102**, 628 (2005).

[166] W. Han and K. Schulten, The journal of physical chemistry. B (2013), 10.1021/jp404331d.

[167] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, Journal of Chemical Theory and Computation **4**, 435 (2008).

[168] T. Škrbić, C. Micheletti, and P. Faccioli, PLoS computational biology **8**, e1002504 (2012).

[169] S. a Beccara, T. Škrbić, R. Covino, C. Micheletti, and P. Faccioli, PLoS Computational Biology **9**, e1003002 (2013).

[170] R. Covino, T. Škrbić, S. a Beccara, P. Faccioli, and C. Micheletti, Biomolecules **4**, 1 (2013).

[171] L. Tubiana, E. Orlandini, and C. Micheletti, Physical Review Letters **107**, 188302 (2011).

[172] J. I. Sułkowska, P. Sułkowski, and J. Onuchic, Proceedings of the National Academy of Sciences of the United States of America **106**, 3119 (2009).

[173] N. P. King, A. W. Jacobitz, M. R. Sawaya, L. Goldschmidt, and T. O. Yeates, Proceedings of the National Academy of Sciences of the United States of America **107**, 20732 (2010).

[174] M. L. Mansfield, Nature Structural Biology **1**, 213 (1994).

[175] W. R. Taylor, Nature **406**, 916 (2000).

[176] P. Virnau, L. a. Mirny, and M. Kardar, PLoS Comput Biol **2**, e122. DOI: 10.1371/journal.pcbi.0020122 (2006).

[177] R. C. Lua and A. Y. Grosberg, PLoS computational biology **2**, e45 (2006).

[178] M. Kardar, The European Physical Journal B **64**, 519 (2007).

[179] N. P. King, E. O. Yeates, and T. O. Yeates, J Mol Biol **373**, 153 (2007).

[180] T. O. Yeates, T. S. Norcross, and N. P. King, Curr Opin Chem Biol **11**, 595 (2007).

[181] K. C. Millett, E. J. Rawdon, A. Stasiak, and J. I. Sułkowska, Biochemical Society transactions **41**, 533 (2013).

[182] P. Virnau, A. Mallam, and S. Jackson, J Phys Condens Matter **23**, 33101 (2011).

[183] R. Potestio, C. Micheletti, and H. Orland, PLoS Comput Biol **6**, e1000864 (2010).

[184] E. Shakhnovich, Nature materials **10**, 84 (2011).

[185] L. Tubiana, A. Rosa, F. Fragiacomo, and C. Micheletti, Macromolecules **46**, 3669 (2013).

[186] A. L. Mallam and S. E. Jackson, Nature chemical biology **8**, 147 (2012).

[187] J. I. Sułkowska, P. Sulkowski, P. Szymczak, and M. Cieplak, Proceedings of the National Academy of Sciences of the United States of America **105**, 19714 (2008).

[188] J. Dzubiella, Biophysical journal **96**, 831 (2009).

[189] T. Bornschlögl, D. M. Anstrom, E. Mey, J. Dzubiella, M. Rief, and K. T. Forest, Biophysical journal **96**, 1508 (2009).

[190] T. C. Sayre, T. M. Lee, N. P. King, and T. O. Yeates, Protein engineering, design & selection : PEDS **24**, 627 (2011).

[191] M. a. Soler and P. F. N. Faísca, PloS one **8**, e74755 (2013).

[192] J. Dzubiella, The Journal of Physical Chemistry Letters **4**, 1829 (2013).

[193] A. L. Mallam, J. M. Rogers, and S. E. Jackson, Proceedings of the National Academy of Sciences of the United States of America **107**, 8189 (2010).

[194] A. L. Mallam and S. E. Jackson, J Mol Biol **346**, 1409 (2005).

[195] A. L. Mallam and S. E. Jackson, Journal of molecular biology **359**, 1420 (2006).

[196] A. L. Mallam and S. E. Jackson, Journal of molecular biology **366**, 650 (2007).

[197] A. L. Mallam, S. C. Onuoha, J. G. Grossmann, and S. E. Jackson, Mol Cell **30**, 642 (2008).

[198] A. L. Mallam, E. R. Morris, and S. E. Jackson, Proceedings of the National Academy of Sciences of the United States of America **105**, 18740 (2008).

[199] A. L. Mallam, FEBS J **276**, 365 (2009).

[200] F. I. Andersson, D. G. Pina, A. L. Mallam, G. Blaser, and S. E. Jackson, The FEBS journal **276**, 2625 (2009).

[201] J. I. Sułkowska, J. K. Noel, C. a. Ramírez-Sarmiento, E. J. Rawdon, K. C. Millett, and J. N. Onuchic, Biochemical Society transactions **41**, 523 (2013).

[202] P. F. N. Faísca, R. D. M. Travasso, T. Charters, A. Nunes, and M. Cieplak, Physical biology **7**, 16009 (2010).

[203] M. a. Soler and P. F. N. Faísca, PloS one **7**, e52343 (2012).

[204] S. Wallin, K. B. Zeldovich, and E. I. Shakhnovich, Journal of molecular biology **368**, 884 (2007).

[205] J. K. Noel, J. I. Sułkowska, and J. N. Onuchic, Proceedings of the National Academy of Sciences of the United States of America **107**, 15403 (2010).

[206] J. I. Sułkowska, J. K. Noel, and J. N. Onuchic, Proceedings of the National Academy of Sciences of the United States of America **109**, 17783 (2012).

[207] W. Li, T. Terakawa, W. Wang, and S. Takada, Proceedings of the National Academy of Sciences of the United States of America **109**, 17789 (2012).

[208] J. K. Noel, J. N. Onuchic, and J. I. Sulkowska, The Journal of Physical Chemistry Letters **4**, 3570 (2013).

[209] D. Bölinger, J. I. Sułkowska, H.-p. Hsu, L. A. Mirny, M. Kardar, J. N. Onuchic, P. Virnau, and D. Bo, PLoS computational biology **6**, e1000731 (2010).

[210] M. Schaefer, C. Bartels, and M. Karplus, Journal of molecular biology **284**, 835 (1998).

[211] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods (San Diego, Calif.) **52**, 99 (2010).

[212] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, The Journal of chemical physics **134**, 174105 (2011).

[213] E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber, The Journal of chemical physics **129**, 174102 (2008).

[214] A. M. a. West, R. Elber, and D. Shalloway, The Journal of chemical physics **126**, 145104 (2007).

[215] P. Májek and R. Elber, Journal of chemical theory and computation **6**, 1805 (2010).

[216] S. Kirmizialtin and R. Elber, The journal of physical chemistry. A **115**, 6137 (2011).

[217] S. M. Kreuzer, R. Elber, and T. J. Moon, The journal of physical chemistry. B **116**, 8662 (2012).

[218] S. M. Kreuzer, T. J. Moon, and R. Elber, The Journal of chemical physics **139**, 121902 (2013).

[219] S. Huo and J. E. Straub, The Journal of Chemical Physics **107**, 5000 (1997).

[220] R. Zhao, J. Shen, and R. D. Skeel, Journal of chemical theory and computation **6**, 2411 (2010).

[221] P. Faccioli and F. Pederiva, Physical Review E **86**, 061916 (2012).

[222] G. Cazzolli, P. Faccioli, F. Wang, and P. Wintrode, in preparation ().

[223] G. Cazzolli, F. Wang, S. a Beccara, A. Gershenson, P. Faccioli, and P. Wintrode, submitted ().