Università degli studi di Firenze
Dottorato in dinamica non lineare e sistemi complessi XXI ciclo

Reconstruction of the free energy landscape of proteins via mechanical manipulation

Stefano Luccioli

Tutore

dott. Alessandro Torcini

Coordinatore

prof. Roberto Genesio

Settore disciplinare FIS/03

Contents

1 Proteins and their manipulation 9

1.1 Proteins . . . 9

1.2 Fundamental interactions within proteins . . . 10

1.3 Primary, secondary and tertiary structure . . . 12

1.4 The folding problem . . . 13

1.5 All-atom models and minimalistic models . . . 13

1.6 Manipulation and mechanical unfolding of proteins . . . 14

2 Models and simulation methods 19

2.1 Simulation methods for the minimalistic model . . . 19

2.1.1 The model . . . 19

2.1.2 Simulation protocol: equilibrium Langevin dynamics . . . 23

2.1.3 Simulation protocol: out-of-equilibrium mechanical unfolding . 24

2.2 Simulation methods for the all-atom model . . . 24

2.2.1 The model . . . 24

2.2.2 Pulling protocols . . . 26

2.2.3 Dynamics . . . 27

3 Free energy reconstruction techniques 29

3.1 Umbrella sampling and weighted histogram analysis method . . . 29

3.2 Jarzynski equality . . . 31

3.3 Extended Jarzynski equality . . . 33

3.4 Inherent structures formalism . . . 37

4 Free energy landscape of mechanically unfolded model proteins 41

4.1 Thermodynamical properties . . . 41

4.2 Extended Jarzynski equality reconstruction . . . 44

4.2.1 Good folder . . . 45

4.2.2 Bad folder . . . 48

4.3 Inherent structures landscape . . . 51

4.4 Tail-pulling versus head-pulling . . . 55

4.5 Concluding remarks . . . 56

5 Mechanical unfolding of FNIII$_{10}$ 59
5.1 Introduction 59
5.2 Analysis of pathways and intermediates 61
5.3 Description of the calculated unfolding traces 62
5.4 Identifying pathways and intermediates 65
5.5 Worm-like chain model analysis 67
5.6 EJE reconstruction 70
5.7 Role of pulling strenght in mechanical unfolding 73
5.8 Concluding remarks 75

Conclusions and perspectives 81

Introduction

Proteins are involved in every function that characterizes a living organism: from the control of the gene expression to the transmission of information between cells and organs (hormones); from the defense against intruders in the immune system (antibodies) to simple structural functions [1]. Moreover many diseases are due to mutations in proteins or to structural modifications causing the loss of their functionality.

Proteins are heteropolymers composed by an ordered sequence of amino acids that forms the so called *primary structure*. The peculiar feature of proteins is that under physiological conditions they *fold* in a unique compact three dimensional structure, called *native configuration*. A complete understanding of the mechanisms involved in the folding process, still lacking, is crucial because it's due to the three dimensional structure that proteins are able to perform their biological functions. Indeed the main aim in protein science is the prediction of the folded configuration starting from primary structure.

In the last fifteen years modern experimental techniques, as atomic force microscopy (AFM) or optical tweezers, have been developed and have allowed the microscopic manipulation of single biomolecules. With these techniques it's possible to induce transitions, as unfolding or dissociation, by exerting mechanical forces and to measure binding forces responsible for the stability of biomolecules with a resolution up to the order of piconewtons. Such a kind of experiments represent therefore a powerful tool to extract information on the internal structure of biomolecules, as well as on unfolding and refolding pathways followed by proteins.

The mechanical manipulation of biomolecules (proteins and nucleic acids RNA and DNA) became important also in conjunction with fluctuation relations (sometimes called *non-equilibrium work relations* [2]), recently developed, that describe the behavior of systems driven out-of-equilibrium. In fact, most of the manipulation's experiments are actually performed switching the system faster than its slowest relaxation rate and therefore in non equilibrium conditions; hysteresis effects are typical markers of irreversible mechanical stretching. Working in out-of-equilibrium regime precluded usually to obtain equilibrium information from experimental data. The novelty introduced by relations as Crooks equality [3] and Jarzynski equality [4, 5], is represented by the amazing feature that it's possible to extract *equilibrium* properties of the system, as thermodynamic free energy differences, from *non* *equilibrium* experiments based on the measure of the work done on the system. The first application of Jarzynski equality to stretching of a single RNA molecule provided for example an estimate of the free energy variation associated to an unfolding reaction [6].

The main aim of this thesis is the numerical study of the mechanical unfolding of two different protein models. The first one is a simplified (*minimalistic*) off-lattice model originally introduced by Honeycutt-Thirumalai [7] and successively modified by Berry *et al.* [8]; it consists of point-like monomers, mimicking the amino acids of a polypeptide chain, where for the sake of simplicity, only three types of amino acids are considered (hydrophobic, polar and neutral ones). We used this model with two different monomers sequences: one previously identified as *good folder* and a sequence randomly chosen (*bad folder*). The aim was to find specific features of protein-like behavior from comparison of the properties of the two sequences.
The second one is an *all-atom* model, developed by Irbäck and coworkers [9, 10] and it has been used in this thesis to investigate the mechanical unfolding in different pulling conditions (at constant force and at constant pulling velocity) of a real protein, the tenth type III domain from fibronectin, FnIII_{10}. The relevance of this study relies on the fact that it was possible to compare the results of the simulations with findings coming from AFM experiments performed by Li *et al.* [11] on FnIII_{10}. From constant velocity pulling simulations and applying the Jarzynski equality in an *extended version* [12, 13, 14, 15] (EJE) we reconstructed the *free energy landscape* (FEL) for both the models as a function of an internal collective coordinate, namely the polypeptide chain extension, which is the natural coordinate reaction in the mechanical unfolding process. The reconstructed FEL, although only *one-dimensional*, was able nevertheless to explain many of the properties of the proteins studied. The EJE reconstruction on the all-atom model extends previous work done on simplified protein models [16, 17, 18, 19]. In particular, pulling simulations on all-atom models may be needed to facilitate comparisons with EJE reconstructions based on experimental data. Indeed quite recently this approach has been successfully applied for the first time to data obtained from manipulation of a real protein, titin I27 domain, with atomic force microscopy [20, 21].
The relevance of this new method, that is able to recontruct the FEL from simple pulling experiments, is crucial because proteins, as well as several states of matter as supercooled liquids, glasses, atomic clusters are typical examples of systems whose thermodynamic behavior can be traced back to the topological properties of the underlying free energy landscape [22].

Proteins are dynamical systems: they vibrate within a given conformation and also jump from a conformation to the other, each corresponding to minima, or *inherent structures* (ISs), in the complex and rough *potential energy landscape*. The *fluctuations* that drive protein dynamics depend on external parameters such as the solvent properties, among which the major role is played by the temperature: at

low temperature protein remain trapped in particular conformations while at room temperature they are able to visit all accessible structures.
The importance of investigating the stationary points of the potential energy landscape goes back to the pioneering work by Stillinger and Weber on inherent structures of liquids [23]. Similar approaches have been proposed and successfully applied, in glasses [24] and supercooled liquids [25]. More recently, this kind of analysis has been applied to the study of protein models [22, 26, 27, 28, 29, 30, 31, 32], which share with the previous systems a rough and intricate energy landscape with a large number of local minima. Moreover, detailed analysis of the thermodynamic and dynamical features, characteristics of proteins, have been quite recently carried out in terms of ISs [30, 31, 32, 33]. These analysis suggest that the folding process of a protein towards its native configuration depends crucially on the structure and topological properties of its (free) energy landscape. Confirming somehow the conjecture that the FEL of a protein has a funnel–like shape: the native configuration being located inside the so–called native valley at the bottom of the funnel itself [34].

Within this framework, in this thesis we use the minimalistic model to compare the free energy landscape reconstruction obtained in terms of two independent methods: the extended Jarzynski equality and an approach based on thermodynamics of inherent structures. Both methods are then compared with the reconstruction performed using a standard equilibrium technique (i.e. the umbrella sampling used in conjunction with the weighted histogram analysis method). Although the model used is relatively simple it shows the main thermodynamic features of a protein-like beahavior and a not trivial FEL with barriers and local minima.

Proteins are not isolated systems but embedded in cells and membranes. Cell-generated forces can extend to several times proteins unstretched lenght. The strenght and the conditions of mechanical pulling can in principle affect the so called *unfolding pathway*, meaning the order of rupture of the sub-structures (β sheets and α helices) inside proteins. So the process of mechanical unfolding can occur through *intermediates states*, partially unfolded, between the native and completely stretched configuration. The mechanical unfolding can also expose critical binding sites, otherwise ”hidden” in the folded state, for activating interaction inside the cell. It is in this context that becomes relevant to study the mechanical unfolding of real protein as FnIII$_{10}$, that is a modul of fibronectin, a giant multidomain protein existing in both soluble (dimeric) and fibrillar forms. In its fibrillar form, fibronectin plays a central role in cell adhesion to the extracellular matrix. Increasing evidence indicates that mechanical forces exerted by cells are a key player in initiation of fibronectin fibrillogenesis as well as in modulation of cell-fibronectin adhesion, and thus may regulate the form and function of fibronectin [35, 36].

The plan of this thesis is the following. The first Chapter is devoted to a brief introduction about the structure of proteins and the relevance of mechanical ma-

nipulation techniques in this context. In the second Chapter we describe the two employed protein's models, as well as the simulation protocols and methods. The third Chapter is devoted to the techniques used for reconstructing the free energy profile of a protein as a function of an internal coordinate, namely the chain extension. The original results about the simplified and all-atom model are reported respectively in Chapters four and five.
In the forth Chapter, after a description of the main thermodynamic properties of the two studied sequences (with bad and good folding properties), we compare the free energy landscape reconstruction obtained in terms of the extended Jarzynski equality, weighted histogram analysis method and inherent structures approach.
In Chapter five we report the analysis of mechanical unfolding of FnIII_{10}, both at constant force and at constant pulling velocity, and the comparison with the experimental data. Finally in the Conclusions the main results obtained in this thesis and future perspectives are discussed and summarized.

The present work of thesis, supported by the European Community via the STREP project EMBIO NEST (contract n.12835), led to the publication of the following three papers:

- A. Imparato, S. Luccioli, and A. Torcini, "Reconstructing the free energy landscape of a mechanically unfolded model protein", Phys. Rev. Lett. 99, 168101 (2007) and addendum in Phys. Rev. Lett. 100, 159903(E) (2008);
- S. Luccioli, A. Imparato, and A. Torcini, "Free energy landscape of mechanically unfolded model proteins: extended Jarzynski versus inherent structure reconstruction", Phys. Rev. E 78, 031907 (2008);
- S. Mitternacht, S. Luccioli, A. Torcini, A. Imparato and A. Irbäck, "Changing the mechanical unfolding pathway of FnIII_{10} by tuning the pulling strength", to appear in 2009 in Biophysical Journal.

Chapter 1

Proteins and their manipulation

In this Chapter we briefly introduce the structure of proteins and their main dynamical features: under physiological conditions proteins fold in a unique compact three dimensional structure thermodynamically stable. Then we describe the relevance of mechanical manipulation techniques in this context. From one side such kind of techniques represent a powerful tool to extract information on the internal structure as well as on the unfolding and refolding pathways of proteins. On the other side, the mechanical manipulation of proteins, and more generally of biomolecules, became important in conjunction with non equilibrium fluctuation relations, recently theoretically developed. In fact using this type of relations, involving the distribution of work done on a system driven out-of-equilibrium (as the process of mechanical unfolding actually is) it is possible to extract equilibrium properties from non equilibrium experiments. In particular, we apply in this thesis an extended form of Jarzynski equality that, from pulling experiments, allows to recover the equilibrium free energy profile of the unconstrained protein.

1.1 Proteins

A protein consists of a chain of amino acids; only 20 kinds of amino acids are present in proteins [1]. All the amino acids (except proline [2]) share the common structure NH_2-C_αRH-COOH where NH_2 is the amino group, COOH the carboxyl group and the group R, called *side chain*, is what makes each amino acid different from each other [3].

Usually amino acids are classified into three main groups according to the na-

[1] In proteins are usually involved only the 20 amino acids of the L-α series (for a complete list see for example Table 1.1 in [37]); nevertheless they may be covalently modified after biosynthesis of the polypeptide chain.

[2] In proline the side chain is also connected covalently to the N atom.

[3] C_α indicates the central carbon atom of the amino acid and C the carbon atom belonging to the carboxyl group. The other carbon atoms eventually belonging to the side chain are indicated with the symbols C_β, C_γ, C_δ and so on.

ture of the side chain: *polar*, *hydrophobic* (or *non-polar*) and *charged*. The amino acid glycine, whose side chain consists of one hydrogen atom only, forms a group by itself. The polar amino acids have an own electric dipole moment that makes them participate in the hydrogen bond network of water. The charged amino acids have a net electric charge and are subject to Coulomb interactions.
In proteins amino acids are connected by the the so called *peptide bond*. When the carboxyl group of the amino acid *i* reacts with the amino-group of the amino acid *i+1*, the peptide bond between $(C)^i$ and $(N)^{i+1}$ is formed and a water molecule (from one H of the amino group and OH of the carboxyl group) is released. Sometimes the amino acids belonging to protein chain (and so missing H and OH) are called *residues*. Therefore a protein consists of a *backbone* formed from the repetition of the elementary unit NH-C_αH-CO and of the side chains attached to it (see Fig. 1.1). The structure of the backbone is identified [4] if for every C_α of the chain the *Ramachandran angles*, ψ and ϕ, are given; ψ and ϕ are respectively the torsion angles between the axes C_α-R and C=O, and C_α-R and N-H (see Fig. 1.2).

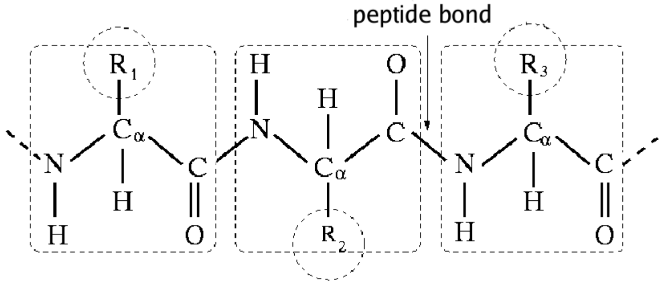

Figure 1.1: Portion of protein chain. The letters R_i stand for the side chain *i* and the peptide bond is indicated with an arrow.

The main feature of proteins is that under physiological conditions (aqueous solvent, temperature 37°C, pH 7, atmospheric pressure) they *fold* in a unique compact three dimensional structure thermodynamically stable, called *native state*. Such a three dimensional structure it's what makes proteins able to accomplish their biological function.

1.2 Fundamental interactions within proteins

To take in account the interactions relevant for stabilizing the protein structure it's necessary, in principle, to include all the non covalent interactions between the

[4] Due to the specific structure of the peptide bond the atoms on its two ends cannot rotate around the bond. Hence the atoms of the group O=C-N-H are fixed on the same plane (called the *peptide plane*); the whole plane may rotate around the N-C_α bond (ϕ angle) or C-C_α bond (ψ angle).

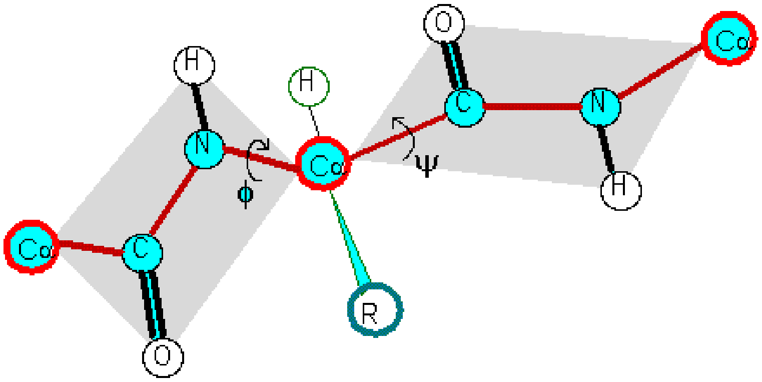

Figure 1.2: Illustration of Ramachandran angles, ϕ and ψ.

atoms and the interactions between the atoms and the molecules of the solvent. In fact, the forces involved in the (covalent) peptide bond are actually three order of magnitude higher than the non covalent forces; nevertheless such energy contribution (of about 2.5 eV) can be neglected because it is essentially constant for every protein conformation and it is not thermally excited at room temperature.

The fundamental non covalent interactions, thermally excited at room temperature, are the hydrogen bonds, the electrostatic forces and the van der Waals forces.

The hydrogen bond, which is a dipole-dipole interaction, has a bond energy of about 0.1-0.3 eV and a bond lenght of around 3 Å. The electrostatic forces, involving groups of atoms with a partial charge, have bond energy and lenght of the same order. The van der Waals forces are an order of magnitude weaker than the previous ones and are long range interactions [5]. One of the most important effect due to the presence of the solvent (typically an aqueous solution) is to influence the spatial distribution of the hydrophobic amino acids, which are not provided of a permanent electric dipole moment and can't participate in the hydrogen bond network of water. In fact they are packaged inside the protein (*hydrophobic core*) while the polar amino acids are in the external part exposed to the water molecules. This process can be looked at as an effective attraction between the hydrophobic residues, called *hydrophobic interaction* (with an energy scale of the order of 0.08 eV), and it is considered one of the key interactions ruling the folding process [38, 39].

[5] The total effective interaction coming from the van der Waals attractive forces is usually represented by a Lennard-Jones potential 6-12, $V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$, where r is the distance between the two molecules, $\sigma \simeq$ 7-9 Å and the repulsion term $\propto 1/r^{12}$ is meant to avoid overlapping of pairs of atoms.

1.3 Primary, secondary and tertiary structure

The *primary* structure of a protein is simply the ordered sequence of the amino acids along the polypeptide chain. The residues are numerated starting from the extremum belonging to the amino group not involved in the peptide bond (*N-terminus*), and then going on until the opposite extremum (*C-terminus*).
Within the typical three dimensional shape of the protein (called *tertiary* structure) it's possible to identify regions of the sequence that form local regular substructures, called *secondary* structures. The pieces of backbone connecting different secondary structures are called *loops*. The most common secondary structures are α *helices* and β *sheets* (see Fig. 1.3). The former are formed when the principal chain atoms belonging to consecutive residues are arranged according to an helix shape. The α helices are stabilized by hydrogen bonds between the group CO of the residue i and the group NH of the residue $i+4$. The side chains are outside the α helix and don't interact with the structure.
The β sheet is formed by an assembly of aligned strands (called β *strands*) composed by typically 5-10 amino acids with an almost fully extended conformation [6]. The β strands are linked by hydrogen bonds between the group CO of one strand and the group NH of the close strand. Moreover, there are proteins where polypeptide chains with an own tertiary structure are assembled in a larger structure called *quaternary* structure.

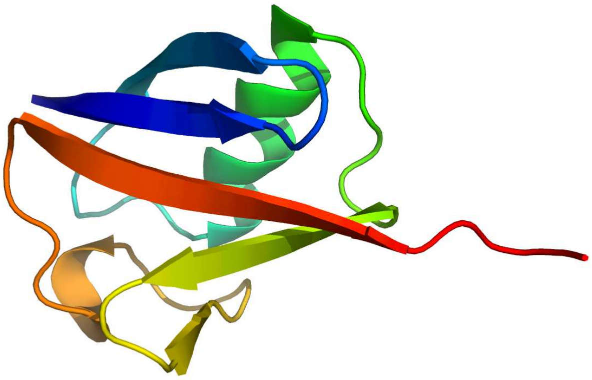

Figure 1.3: Illustration of the native structure of ubiquitin (Protein Data Bank 1UBQ). It's possible to recognize in such a protein the most common secondary structures: α helix and β strands (arrow shaped).

[6] The β sheet is said *parallel* if all the strands are aligned in the same biochemical direction, i.e. from the amino to the carboxyl group (N→C). The β sheet is said *antiparallel* if amino acids in close strands have opposite direction (N→C and C→N).

1.4 The folding problem

A complete understanding of the mechanisms involved in the *folding process* is still lacking. The Levinthal's "paradox" [40] rules out that folding process occurs by a random sampling of the huge number of all possible conformations, because it would be necessary a time greater of the age of the universe. Therefore it seems that to reach the native state (that according to Anfinsen's [41] thermodynamic hypothesis is the global minimum of the free energy) in the observed time scales proteins follow well defined pathways. This means that the native state has to be *kinetically accessible* starting from a generic initial condition: the pathway in the free energy surface from the unfolded to the native state has to be "smooth", proceeding through "small" successive conformational rearrangements. In [42, 43, 44] it was conjectured that the potential energy landscape of protein, though it is very rough [7] with a large number of local minima, shows a global slope towards the native configuration with a *funnel*-like shape. Usually in thermal folding/unfolding dynamics three *transition temperatures* are identified: *glassy*, *folding* and *hydrophobic collapse* temperature (see also Section 4.1). The glassy temperature represents the critical value above which the protein can still reach the native state jumping from a minimum to the other; otherwise, if the temperature is below such a value an arrest of the jumping dynamics occurs, and the protein can remain trapped in a local minimum without reaching the native state in a finite time. At the folding temperature by definition the configurations visited by the protein in its dynamics belong predominantly to the native basin. The collapse temperature discriminates between phases dominated by open (*random-coil*) and compact configurations.
A common and schematic representation of folding-unfolding dynamics is shown in Fig. 1.4: the native state (N) and the unfolded state (U) correspond to free energy minima with respect to some reaction coordinate and are separated from a transition state (T).

1.5 All-atom models and minimalistic models

The *ab initio* approach to the folding problem relies on the fact that it's possible, at least in principle, to get the three dimensional structure from the sequence of the amino acids using a molecular dynamics code that integrates numerically the motion's laws of all atoms of the protein using a model as realistic as possible, including all the intermolecular interactions and the interaction solvent-protein (*all-atom models*). The smallest time scale in proteins corresponds to vibrational motions of atoms, which are of the order of 10^{13}-10^{14} Hz; therefore for the numerical integration it's necessary a time step of at least 10^{-15} s. But considering the large number of

[7] During the folding process different parts of protein come close and there are parts of the energy potential that compete. So proteins are systems characterized by the presence of *frustration* and the configurations corresponding to local energy minima are connected each other in a complex way.

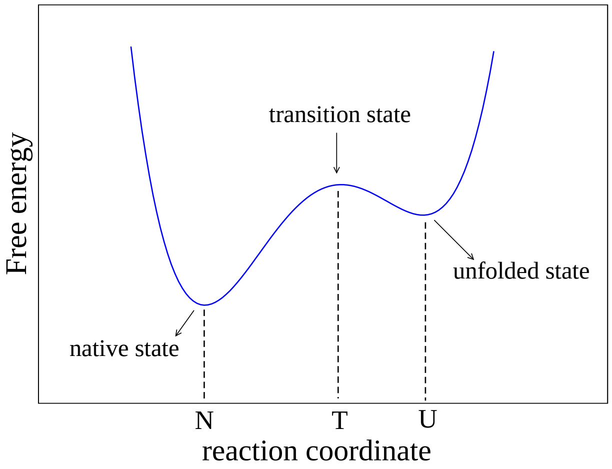

Figure 1.4: Sketch of folding-unfolding dynamics.

atoms belonging to protein with the latest computers it's possible to simulate dynamics of proteins only up to a time of order of microseconds which is much lesser of the experimentally observed folding times (of order 10^{-3}-10 s).
To overcome the computational impracticability of an *ab initio* approach, a first approximation is to replace the effect of the interaction between the protein and the solvent by an effective potential (*implicit solvent*). On the other side, skipping more and more details of the intermolecular interactions, it's possible with *minimalistic models* to follow the evolution of the system until the folding times. On the other hand all-atom models can be appropriate for studying the process of *mechanical unfolding* (see Section 1.6) also for large proteins, and actually we did for FnIII_{10}, tenth type III domain from fibronectin (see Chapter 5). In fact, in order to study mechanical unfolding, it is sufficient to verify that the native state is a long-lived state corresponding to a local free energy minimum.

1.6 Manipulation and mechanical unfolding of proteins

In the last fifteen years modern experimental techniques, as atomic force microscopy (AFM) or optical tweezers, have been developed and have allowed the microscopic manipulation of single biomolecules (proteins and nucleic acids RNA and DNA). With these techniques it's possible to induce transitions, as unfolding or dissociation, by exerting mechanical forces and to measure binding forces responsible for the stability of biomolecules. Therefore, mechanical unfolding of single biomolecules represents a powerful technique to extract information on the internal structure of these microsystems as well as on the unfolding and refolding pathways of proteins [53, 57, 60, 58, 61].

With atomic force microscopy it's possible to measure forces in the pN-nN range

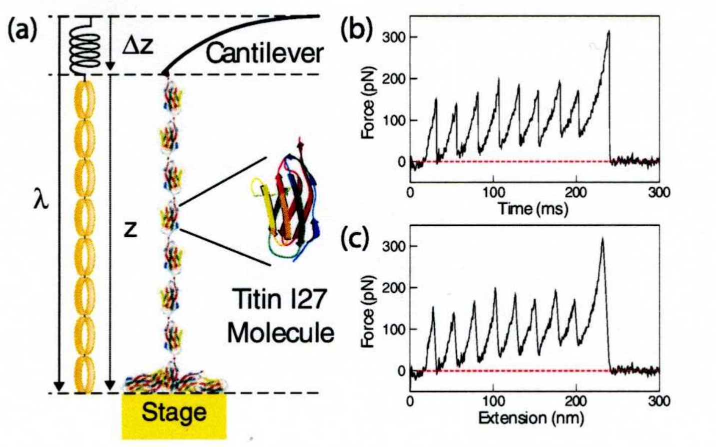

Figure 1.5: (a) Sketch of a constant velocity pulling experiment with AFM of an engineered protein composed of eight domains of titin I27. The cantilever acts as a spring obeying Hooke's law. (b) and (c) Typical sawtooth pattern in the force-time and force-extension curve (the figure is taken from [20] and the data refers to a pulling velocity of 1.00 μm/s).

with high resolution in a realistic solvent [45]. The main component is a microfabricated cantilever (with a lenght in between 20 and 300μm) at the end of which a sharpened tip (with a radius of about 10nm) is sited. The interaction between the tip and the sample gives rise to a force (obeying Hooke's law) that can be measured from the cantilever detection. When the AFM is used for mechanical unfolding one end of the protein is immobilized on the gold substrate of a stage and the other end is attached to the tip cantilever. By moving vertically or the cantilever or the stage (depending on the type of AFM) at a constant velocity (*constant velocity protocol)* an increasing force is applied to the protein until it unfolds and the force abruptly drops down. Generally this is done with an engineered protein composed of several domains; in such a case the outcome of the experiment results in a typical *sawtooth* pattern in the force-extension curve (see Fig. 1.5). Another technique for manipulation of biomolecules is the optical tweezers, which is a special kind of optical trap. It relies on the fact that a strongly focused laser beam can be used to catch and hold particles of dielectric material in a size range from nanometers to micrometers. The biomolecules are attached to these dielectric beads; one of these is held in the trap and the other one is moved. Also in this case the interaction between the bead and the trap follows Hooke's law.

The atomic force microscope and the optical tweezers can be used also with a different pulling protocol, called *constant force protocol.* In this kind of experiments a force-clamp technique, based on a feedback system, is used to control the magnitude of the force acting on the protein. Also in this experimental setup a polyprotein made of several modules is generally used; in this case the unfolding trajectories result in a typical *stepwise* pattern in the lenght-time curve (see Fig. 1.6).

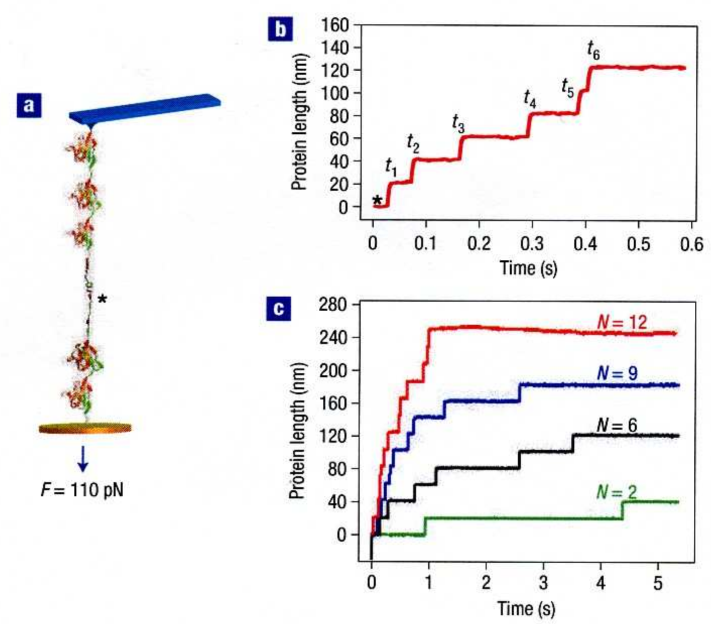

Figure 1.6: AFM constant force pulling experiments on a polyprotein composed of several modules of ubiquitin. Figures (a) and (b) refer to a chain of six modules and figure (c) refers to ubiquitin chains with a varying number of modules, from N=2 to N=12 (the figure is taken from [46] and data refers to a constant force of 110 pN).

The mechanical manipulation of biomolecules became important also in conjunction with fluctuation relations (sometimes referred to as *non-equilibrium work relations* [2]), recently developed, that describe the behavior of systems driven out-of-equilibrium. In fact, most of the manipulation's experiments are actually performed switching the system faster than its slowest relaxation rate and therefore in out of equilibrium conditions. Working in out-of-equilibrium regime precluded usually to obtain equilibrium information from experimental data. The novelty introduced by relations as Crooks equality [3] and Jarzynski equality [4, 5] (see Chapter 3) is represented by the amazing feature that it's possible to extract *equilibrium* properties of the system, as thermodynamic free energy differences, from *non equilibrium* experiments based on the measure of the work done on the system. In particular, in this thesis we extensively used Jarzynski equality in an *extended form* [12, 13, 14, 15] that we will introduce in Chapter 3. The Jarzynski equality was applied for the first time by Liphardt *et al.* [6] in a famous experiment regarding the stretching with optical tweezers of a single RNA molecule derived from the P5abc domain of the *Tetrahy-*

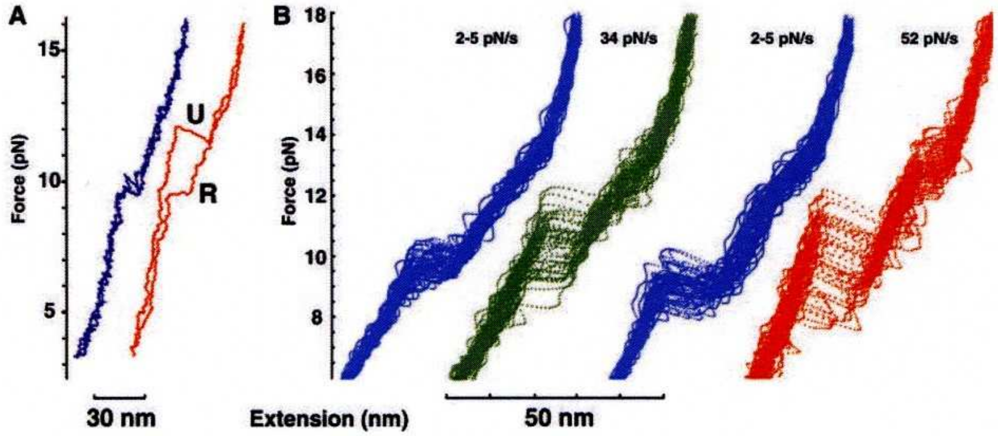

Figure 1.7: Force-extension curves during mechanical unfolding (U) - refolding (R) cycles of single RNA molecule at different switching rates (in pN/s). In A) two examples of trajectories are presented in reversible (blue, 2 to 5 pN/s) and irreversible (red, 52 pN/s) switching conditions (please note the hysteresis effect). In B) a superposition of about 40 curves per experiment are presented (figure taken from [6]).

mena thermophyla group 1 intron; the experiment provided an estimate of the free energy variation associated to the unfolding reaction. Liphardt *et al.* showed that with this molecule both the equilibrium and out-of-equilibrium regime were experimentally accessible: the molecule unfolds *reversibly* when stretched slowly compared to its typical relaxation time and *irreversibly* when stretched more rapidly. Typical marker of irreversible mechanical stretching was hysteresis effect recorded in the force-extension curves during a cycle of unfolding-refolding (see Fig. 1.7).

Due to effect of the collisions between the system and the molecules of the solvent, the dynamics of systems of size of proteins (from a few to some hundreds nanometers) in a thermal bath is essentially stochastic. So the trajectory followed by the system is different for every repetitions of the *same* out-of-equilibrium experiment. Let x be the variable that identifies the system microscopic state, e.g. the collection of the positions and momenta of all the particles in the system $x = \{\boldsymbol{r}_i, \boldsymbol{p}_i\}$. Let $\lambda(x)$ be a macroscopic observable of the system whose value is varied during the process according to a well defined *non-equilibrium protocol.* For example in a pulling experiment on a protein with an AFM (see Fig 1.5) the variable λ is the distance between the cantilever and the surface where the protein is blocked. Due to the manipulation process the variation of inner energy of the system will be $\Delta U = Q + W$, where Q is the exchanged heat and W is the total work done on the system when

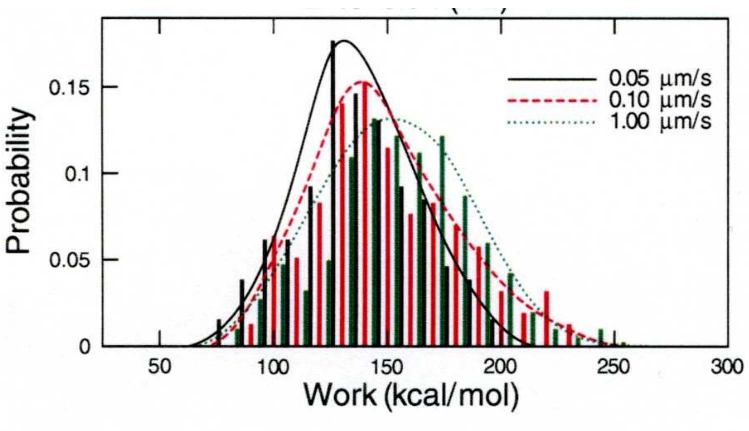

Figure 1.8: Distributions of work for stretching and unfolding one titin I27 domain in the same experimental setup as in Fig. 1.5 for three different pulling velocity 0.05, 0.10 and 1.00 μm/s (figure taken from [20]).

the variable λ is changed from the initial value λ_i to the final value λ_f:

$$W = \int_{\lambda_i}^{\lambda_f} F d\lambda \tag{1.1}$$

being F the force applied to the system. The only deterministic quantity in such experiments is the observable λ (*control parameter*); while the force F, the work W and the heat Q are random. In Fig. 1.8, for example, one can see the distributions of works obtained during many repetitions of the same constant velocity pulling experiment with AFM [20]; it is worth noting that decreasing the pulling velocity, and so approaching the reversible limit, the distributions become narrower. In such a framework the work and heat distributions are very important for characterizing the behavior of the system.

Chapter 2

Models and simulation methods

This Chapter is devoted to the introduction of the studied protein's models, as well as of the simulation protocols and methods. Two kind of models have been employed. The first one is a modified version of a minimalistic model, previously introduced by Thirumalai and coworkers [7], simulated via a Langevin dynamics and with a constant pulling velocity protocol. The minimalistic model has been used to study the mechanical unfolding both of a sequence known in literature as a *good folder* and of a random sequence, *bad folder*. The second one is an all-atom model, developed by Irbäck and coworkers, simulated via Monte Carlo dynamics both at constant force and at constant pulling velocity; with this model it has been investigated the mechanical unfolding of a real protein, the tenth type III domain from fibronectin (FnIII$_{10}$).

2.1 Simulation methods for the minimalistic model

The model we used is a modified version of the 3d off-lattice model introduced by Honeycutt-Thirumalai [7] and successively generalized by Berry *et al.* to include a harmonic interaction between next-neighbouring beads instead of rigid bonds [8]. This model has been widely studied in the context of thermally driven folding and unfolding [7, 64, 65, 8, 69, 29, 70, 32] and only more recently for what concerns mechanical folding and refolding [72, 71].

2.1.1 The model

The model consists of a chain of L point-like monomers mimicking the residues of a polypeptide chain. For the sake of simplicity, only three types of residues are considered: hydrophobic (B), polar (P) and neutral (N) ones.

The intramolecular potential is composed of four terms: a stiff nearest-neighbour harmonic potential, V_1, intended to maintain the bond distance almost constant, a three-body interaction V_2, which accounts for the energy associated to bond angles, a four-body interaction V_3 corresponding to the dihedral angle potential, and a long–range Lennard-Jones (LJ) interaction, V_4, acting on all pairs i, j such that $|i-j| > 2$, namely

$$V_1(r_{i,i+1}) = \alpha(r_{i,i+1} - r_0)^2, \tag{2.1}$$

$$V_2(\theta_i) = A\cos(\theta_i) + B\cos(2\theta_i) - V_0, \tag{2.2}$$

$$V_3(\varphi_i,\theta_i,\theta_{i+1}) = C_i[1 - S(\theta_i)S(\theta_{i+1})\cos(\varphi_i)] + D_i[1 - S(\theta_i)S(\theta_{i+1})\cos(3\varphi_i)] \tag{2.3}$$

$$V_4(r_{i,j}) = \varepsilon_{i,j}\left(\frac{1}{r_{i,j}^{12}} - \frac{c_{i,j}}{r_{i,j}^{6}}\right) \tag{2.4}$$

Here, $r_{i,j}$ is the distance between the i-th and the j-th monomer, θ_i and φ_i are the bond and dihedral angles at the i-th monomer, respectively. The parameters $\alpha = 50$ and $r_0 = 1$ (both expressed in adimensional units) fix the strength of the harmonic force and the equilibrium distance between subsequent monomers (which, in real proteins, is of the order of a few Å). The value of α is chosen to ensure a value for V_1 much larger than the other terms of potential in order to reproduce the stiffness of the protein backbone. The expression for the bond-angle potential term $V_2(\theta_i)$ (2.2) corresponds, up to the second order, to a harmonic interaction term $\sim (\theta_i - \theta_0)^2/2$, where

$$A = -k_\theta \frac{cos(\theta_0)}{\sin^2(\theta_0)}, \qquad B = \frac{k_\theta}{4\sin^2(\theta_0)}, \qquad V_0 = A\cos(\theta_0) + B\cos(2\theta_0) \quad , \tag{2.5}$$

with $k_\theta = 20\epsilon_h$, $\theta_0 = 5\pi/12$ *rad* or 75^o and where ϵ_h sets the energy scale. This formulation in terms of cosines allows to speed up the simulation, since it is sufficient to evaluate $\cos(\theta_i)$ and the value of bond-angle is not needed, and at the same time to avoid spurious divergences in the force expression due to the vanishing of $\sin(\theta_i)$ when three consecutive atoms become aligned [73].

The dihedral angle potential is characterized by three minima for $\varphi = 0$ (associated to a so-called *trans state*) and $\varphi = \pm 2\pi/3$ (corresponding to *gauche states*), this potential is mainly responsible for the formation of secondary structures. In particular large values of the parameters C_i, D_i favor the formation of trans state and therefore of β-sheets, while when gauche states prevail α-helices are formed. The parameters (C_i, D_i) have been chosen as in [69], i.e. if two or more beads among the four defining φ are neutral (N) then $C_i = 0$ and $D_i = 0.2\varepsilon_h$, in all the other cases $C_i = D_i = 1.2\varepsilon_h$ (see Fig. 2.1.1). The *tapering function* $S(\theta_i) = 1 - \cos^{32}(\theta_i)$ has been introduced in the expression of V_3 in order to cure a well known problem in the dihedral potentials [73]. This problem is encountered whenever $\theta_i = 0$ or π, i.e. when three consecutive beads are in the same line, in these situations the associated dihedral angle is no more defined and a discontinuity in V_3 arises. In contrast to what reported in [73] this situation is not improbable for the present model. The quantity $S(\theta_i)S(\theta_{i+1})$ entering in the definition of V_3 has a limited influence on the dynamics apart in proximity of the above mentioned extreme cases. Moreover, $S(\theta_i)S(\theta_{i+1})$ is C^∞, its value is essentially one almost for any θ_i, it does

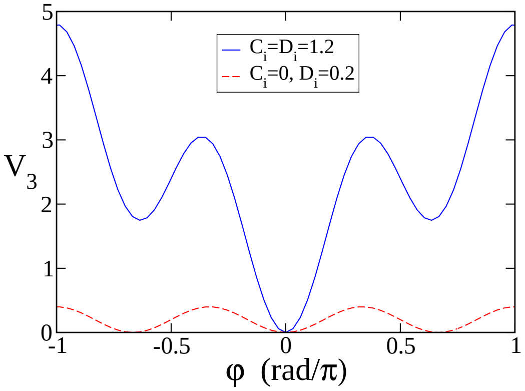

Figure 2.1: Dihedral angle potential, V_3, when two or more beads among the four defining φ are neutral (red dashed curve), and in all the other cases (blue solid curve). We fixed $\varepsilon_h = 1$ and $S(\theta_i) = S(\theta_{i+1}) \equiv 0$.

not introduce any extra minima in the potential and it vanishes smoothly for $\theta_i \longrightarrow 0$ or $\theta_i \longrightarrow \pi$ [74].

The last term V_4 has been introduced to mimic effectively the interactions with the solvent, it is a Lennard-Jones potential and it depends on the type of interacting residues as follows:

- if any of the two monomers is neutral the potential is repulsive $c_{N,X} = 0$ and its scale of energy is fixed by $\varepsilon_{N,X} = 4\varepsilon_h$;
- for interactions between hydrophobic residues $c_{B,B} = 1$ and $\varepsilon_{B,B} = 4\varepsilon_h$;
- for any polar-polar or polar-hydrophobic interaction $c_{P,P} \equiv c_{P,B} = -1$ and $\varepsilon_{P,P} \equiv \varepsilon_{P,B} = (8/3)\varepsilon_h$.

Accordingly, the Hamiltonian of the system reads

$$H = K + V = \sum_{i=1}^{L} \frac{p_{x,i}^2 + p_{y,i}^2 + p_{z,i}^2}{2} + \sum_{i=1}^{L-1} V_1(r_{i,i+1}) + \\ + \sum_{i=2}^{L-1} V_2(\theta_i) + \sum_{i=2}^{L-2} V_3(\varphi_i, \theta_i, \theta_{i+1}) + \sum_{i=1}^{L-3} \sum_{j=i+3}^{L} V_4(r_{ij}) \qquad (2.6)$$

where, for the sake of simplicity, all monomers are assumed to have the same unitary mass, the momenta are defined as $(p_{x,i}, p_{y,i}, p_{z,i}) \equiv (\dot{x}_i, \dot{y}_i, \dot{z}_i)$ and we fix $\varepsilon_h = 1$.

In this thesis we consider the two following sequences of 46 monomers,

- [GF]=$B_9N_3(PB)_4N_3B_9N_3(PB)_5P$ a sequence that has been widely analyzed in the past for spontaneous folding [7, 64, 65, 69, 8, 29, 70, 32] as well as for mechanical unfolding and refolding [72, 71];

- [BF]= $BNBPB_3NPB_4NBPB_2NP_2B_5N_2BPBNPB_2NBP_2BNB_2PB_2$ a randomly generated sequence, but with the same number of B, P and N monomers as the GF.

These two sequences have been chosen because GF has been previously identified as a reasonably fast folder [64] (see also [8] for a detailed and critical analysis of the basin-bottom structures observed for this model), while we expect that the sequence BF, being randomly chosen, cannot have the characteristic of a good folder. From now on we refer to the sequence GF (resp. BF) as the *good* (resp. *bad*) folder.

The 46-mer sequence GF exhibits a four stranded β-barrel Native Configuration (NC) (see Fig. 2.2a) with an associated potential energy $E_{NC} = -49.878$. Please note that the model is here analyzed by employing the same potential and parameter set reported in Ref. [69], but neglecting any diversity among the hydrophobic residues. The NC is stabilized by the attractive hydrophobic interactions among the B residues, in particular the first and third B_9 strands, forming the core of the NC, are parallel to each other and anti-parallel to the second and fourth strand, namely, $(PB)_4$ and $(PB)_5P$. The latter strands are exposed towards the exterior due to the presence of polar residues.

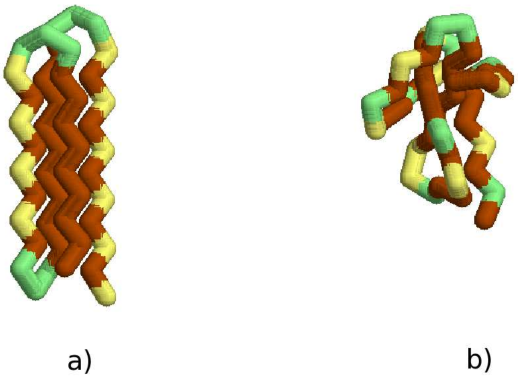

Figure 2.2: Native structure of the good (a) and bad (b) folder.

The native structure of the BF is quite different, it has a core constituted by the first three β-strands and a very long "tail" (made of 18 residues) wrapped around the core (see Fig. 2.2b). In particular, the first and second β-strand (namely, $BNBPB_3NP$ and B_4NBPB_2) are formed by 9 residues, and antiparallel to each other. For more clarity, we will term π_1 the plane containing the first 2 strands. The third strand (namely, P_2B_5) is made of 7 residues and it is located in a plane lying in-between the first and second strand, which is almost perpendicular to π_1. The chain rotates of almost 90 degrees in correspondence of the two consecutive neutral beads and then exhibits a short strand of 3 beads PBP before turning back

	GF	BF
V_{NC}	-49.878	-23.956
V_1	0.787	0.777
V_2	1.767	5.744
V_3	2.602	23.105
V_4	-55.035	-53.582

Table 2.1: Potential energy values associated to the NC of the GF and BF, the different contributions to the total potential energy V_{NC} are also reported.

with a parallel strand of 7 beads (PB_2NBP_2) that passes below π_1. Finally the chain turns once more back by passing this time above the plane π_1. In the final part of the tail of the chain a short strand of 5 residues, parallel to the 4-th and 5-th strands, can be identified as B_2PB_2. The potential energy of the NC of the BF is quite high with respect to the GF, namely $V_{NC} = -23.956$. Moreover, this difference, as reported in Table 2.1, is essentially due to the difference in the dihedral contributions, that is much higher in the NC of the BF with respect to the GF, while all the other contributions, in particular the LJ ones, have nearby values. The dihedral contribution that arises in the BF is essentially due to the configuration of the first 3 strands, since these are arranged over two almost orthogonal planes.

2.1.2 Simulation protocol: equilibrium Langevin dynamics

Molecular dynamics (MD) canonical simulations at equilibrium temperature T have been performed by integrating the corresponding Langevin equation for each monomer of unitary mass (characterized by the position vector $\mathbf{r}_i$):

$$\ddot{\mathbf{r}}_i = \mathbf{F}(\mathbf{r}_i) - \gamma\dot{\mathbf{r}}_i + \eta(t) \qquad i = 1, L \tag{2.7}$$

where $\eta(t)$ is a zero average Gaussian noise term (mimicking the collisions of the molecules of the solvent and the monomers) with correlations given by $\langle \eta_\alpha(t)\eta_\beta(t')\rangle = 2T\gamma\delta(t-t')\delta_{\alpha,\beta}$; $\mathbf{F} = -\nabla V$, being V the intramolecular potential introduced in 2.1.1, γ the friction coefficient associated to the solvent and by assuming an unitary Boltzmann constant k_B.

Numerical integrations have been implemented via a standard Euler scheme with a time-step $\Delta t = 0.005$ and with a low friction coefficient $\gamma = 0.05$ [69]. Two different kinds of MD have been performed, namely unfolding simulations (US) and folding simulations (FS). In the first case the initial state of the system is taken equal to the native configuration (NC), that we assume to coincide with the minimal energy configuration. In the latter one the initial state is a completely unfolded configuration.

2.1.3 Simulation protocol: out-of-equilibrium mechanical unfolding

In order to mimic the mechanical pulling of the protein attached to a AFM cantilever, or analogously when trapped in an optical tweezer, one extremum of the chain was kept fixed and the last bead is attached to a pulling apparatus with a spring of elastic constant k. The external force is applied by moving the "cantilever" along a fixed direction with a certain protocol $z(t)$. Before pulling the protein, the coordinate system is always rigidly rotated, in order to have the z-axis aligned along the end-to-end direction connecting the first and last bead. Therefore by denoting with ζ the end-to-end distance the component of the external force along this direction reads as

$$F_{ext} = k(z - \zeta) \tag{2.8}$$

where $k = 10$ in order to suppress fast oscillations. As recently pointed out [16] it is extremely important to use a sample of thermally equilibrated initial configurations to correctly reproduce the equilibrium free energy landscape via the Jarzynski equality (see Chapter 3). Therefore, before pulling the protein, we have performed a thermalization procedure in two steps. At a fixed temperature T, initially the protein evolves freely starting from the NC for a time $t = 1,000$, then it is attached to the external apparatus, with the first bead blocked, and it equilibrates for a further time period $t = 500$. The system (at sufficiently low temperatures) quickly settles down to a "native-like" configuration. This configuration is then employed as the starting state for the forced folding. The protocol that we have used is a linear pulling protocol with a constant speed v_p, i.e. $z(t) = z(0) + v_p \times t$, by assuming that the pulling starts at $t = 0$. Usually we have employed velocities $v_P \in [5\times10^{-5} : 5\times10^{-2}]$ and set $z(0) = \zeta_0$, i.e. to the end-to-end distance associated to the native configuration.

2.2 Simulation methods for the all-atom model

This section is devoted to describe the simulation methods used for studying the mechanical unfolding of the tenth type III domain from fibronectin, FnIII$_{10}$ (see Chapter 5 where in Fig. 5.1 it is also reported the structure of FnIII$_{10}$).

2.2.1 The model

The model was developed by Irbäck *et al* [9, 10]. It is an all-atom model with implicit water where the only degrees of freedom are torsional (Ramachandran angles and side chain torsion angles). Bond lenghts, bond angles and peptide bond torsion angles are held constant. The interaction potential E is composed of four terms:

$$E = E_{loc} + E_{ev} + E_{hb} + E_{hp} \tag{2.9}$$

The term E_{loc} is local in sequence and represents an electrostatic interaction between adjacent peptide units along the chain. The other three terms are non-local in sequence. The excluded volume term E_{ev} represents a repulsion between pairs of atoms. The term E_{hb} represents two kinds of hydrogen bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The term E_{hp} represents an effective hydrophobic attraction between nonpolar side chains; it is a simple pairwise additive potential based on the degree of contact between two nonpolar side chains.

It has beeen shown [9] that this model provides a good description of the structure and folding thermodynamics of several peptides with about 20 residues. For larger proteins as FnIII_{10} it is computationally infeasible to verify that the native structure is the global free-energy minimum. However, in order to study unfolding, it is sufficient that the native state is a long-lived state corresponding to a local free energy minimum. Without any tuning of the parameters used in previous studies [9, 10], we found that the native state of FnIII_{10}, indeed, is a long-lived state corresponding to a free-energy minimum, as will be seen in Chapter 5.

Below, the most important features of the model are summarized, while the numerical values of all geometry parameters can be found in [9, 10]. The term E_{loc} is given by:

$$E_{loc} = k_{loc} \sum_{I} \Big(\sum_{\substack{i=N,H \\ j=C,O}} \frac{q_i q_j}{r_{ij}} \Big) \tag{2.10}$$

where each term of the inner sum represents the interaction of the partial charges q$_i$ of the backbone NH and CO groups for each amino acid I.

The term E_{ev} has the form:

$$E_{ev} = k_{ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12} \tag{2.11}$$

where the sum is over pairs of atom. The σ_i are the atomic radii. The parameters λ_{ij} compensate the fact that the bond lenghts and angles are held constant in the model: the decreased flexibility of a chain with only torsional degrees of freedom could in fact lead to artificial traps; to avoid that, for all atom pairs that are separated by more than three covalent bonds $\lambda_{ij} < 1$, otherwise $\lambda_{ij} = 1$.

The hydrogen bonds energy is given by:

$$E_{hb} = \epsilon_{hb}^{(1)} \sum_{bb-bb} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{hb}^{(2)} \sum_{bb-sc} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \tag{2.12}$$

where the first term of the sum represents the backbone-backbone interaction and the second one the interaction between the charged side chains and the backbone; the functions $u(r)$, $v(\alpha, \beta)$ are defined as:

$$u(r) = 5(\frac{\sigma_{hb}}{r})^{12} - 6(\frac{\sigma_{hb}}{r})^{10} \tag{2.13}$$

$$v(\alpha,\beta) = \begin{cases} (\cos(\alpha)\cos(\beta))^{1/2} & \text{if } \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

In the model only hydrogen bonds between NH and CO are considered, therefore r_{ij} is the HO distance, α_{ij} is the NHO angle and β_{ij} the HOC angle.

The hydrophobicity term E_{hp} is expressed as a pairwise additive form over non-polar (or hydrophobic) side chains:

$$E_{hp} = -\sum_{I<J} M_{IJ} C_{IJ} \tag{2.15}$$

where C_{IJ} is defined as:

$$C_{IJ} = \frac{1}{N_I + N_J}\Big[\sum_{i\in A_I} f(\min_{j\in A_J} r_{ij}^2) + \sum_{j\in A_J} f(\min_{i\in A_I} r_{ij}^2)\Big] \tag{2.16}$$

and the function $f(x)$ is (A and B are parameters):

$$f(x) = \begin{cases} 1 & \text{if } x < A \\ \frac{B-x}{B-A} & \text{if } A < x < B \\ 0 & \text{if } x > B \end{cases} \tag{2.17}$$

The term C_{IJ} is a geometry factor that represents a measure of the degree of contact between side chains I and J, and it was calculated using a predetermined sets of N_I atoms, A_I, for each side chain I. The matrix M_{IJ} sets the energy that a pair in full contact gets [1].

2.2.2 Pulling protocols

The energy function E of Eq. 2.9 describes an unstretched protein. In our simulation we have used two kinds of pulling protocol: at constant force and at constant velocity. In the first case, constant forces $-\vec{F}$ and $\vec{F}$ act on the N and C termini, respectively. The full energy function is then given by:

$$E_{tot} = E - \vec{F}\cdot\vec{\zeta} \tag{2.18}$$

where $\vec{\zeta}$ is the vector from the N to the C terminus.
For the constant-velocity simulations, the pulling of the protein is modeled using a harmonic potential in the chain extension (end-to-end distance) $\zeta = |\vec{\zeta}|$ whose minimum $z(t)$ varies linearly with Monte Carlo (MC) time t. With this external potential, the full, time-dependent energy function becomes

$$E_{tot}(t) = E + \frac{k}{2}\left[z(t) - \zeta\right]^2 = E + \frac{k}{2}\left[z(0) + v_p t - \zeta\right]^2 \tag{2.19}$$

[1] In the model the hydrophobic amino acids are divided into three categories [10]. The matrix M_{IJ} represents the size of hydrophobicity ineraction when an amino acid of type I is in contact with an amino acid of type J.

where k is a spring constant, v_p is the pulling velocity, and $z(0)$ is the initial equilibrium position of the spring. The spring constant, corresponding to the cantilever stiffness in AFM experiments, is set to $k = 37\,\mathrm{pN/nm}$ (of the same order of magnitude of the typical spring constant of $k \sim 50\,\mathrm{pN/nm}$ used in an experimental study [59] with AFM about FnIII_{10}).
We have considered six constant force magnitude F (50 pN, 80 pN, 100 pN, 120 pN, 150 pN and 192 pN) and four constant pulling velocities v_p (0.03 fm/MC step, 0.05 fm/ MC step, 0.10 fm/MC step and 1.0 fm/MC step).

As a starting point for simulations, we use a model approximation of the experimental FnIII_{10} structure with backbone root-mean-square deviation[2], δ, $\approx 0.2\,\mathrm{nm}$, obtained by simulated annealing[3]. All simulations are started from this initial structure, with different random number seeds. However, in the constant-velocity runs, the system is first thermalized in the potential $E + k(z(0) - \zeta)^2/2$ for 10^7 MC steps ($z(0) = 3.8\,\mathrm{nm}$), before the actual simulation is started at $t = 0$. The thermalization of the initial structure is a prerequisite to apply the Jarzynski reconstruction.

The constant-force simulations are run for a fixed time, which depends on the force magnitude. There are runs in which the protein remains folded over the whole time interval studied. The constant-velocity simulations are run until the spring has been pulled a distance of $v_p \times t = 35\,\mathrm{nm}$. At this point, the protein is always unfolded.

2.2.3 Dynamics

The simulations were performed using Monte Carlo dynamics at a temperature T of 288 K. Moreover, in the constant-velocity simulations, the time-dependent parameter $z(t)$ is changed after every attempted MC step.
Three different types of MC updates have been used: (i) single-variable Metropolis [4] updates of side-chain angles; (ii) Biased Gaussian Steps [54], BGS, which are semi-local updates of backbone angles; and (iii) small rigid-body rotations of the whole

[2] The root-mean-square deviation (RMSD), δ, is a common way to measure the similarity of a configuration to the native state and more generally to compare the structures of two biomolecules. It is defined as $\delta^2_{x,y} = \min \frac{1}{N} \sum_{k=1}^{N} |\vec{r}_{k,x} - \vec{r}_{k,y}|^2$, where the sum is over N pairs of equivalent atoms of the conformations x and y with cartesian coordinate $\vec{r}_{k,x}$, $\vec{r}_{k,y}$ and the minimum is taken with respect to all rigid body rotations and translations of the structures compared.

[3] Simulated annealing [55] is a general global minimization method. It incorporates the Metropolis algorithm (see Sec. 2.2.3) for updates and lowers temperature gradually. Trapping in local minima can be avoided by allowing increases in energy every now and then. Lowering the temperature makes increases in energy less and less probable, allowing the system to stabilize at some point. In our case a model approximation of the experimenatl FnIII_{10} structure was found by simulated annealing-based optimization of the auxiliary function $E' = E + a\delta^2$, where E is the energy function 2.9, a is a parameter and δ the RMSD calculated over the atoms of the backbone C_α, C, O, N.

[4] Let us suppose to have a system in a state X and that a move towards a state X' is suggested. Then, if $p(X)$ and $p(X')$ are the probabilities to be in the respective states, the Metropolis [56]

chain. The BGS move simultaneously updates up to eight consecutive backbone angles, in a manner that keeps the chain ends approximately fixed [5].

accept/reject rule sets the probability for accepting the change to:

$$P(X \to X') = \min(1, p(X')/p(X)) \tag{2.20}$$

where it is assumed that the probabilities for suggesting the changes $X \to X'$ and $X' \to X$ are equal. In thermodynamics the canonical probability to be in the state X is $\propto \mathrm{e}^{-\beta E(X)}$ (where $E(X)$ is the energy of the state, $\beta = (k_B T)^{-1}$ and k_B is Boltzmann's constant), so the acceptance criterion of the move reads as:

$$P(X \to X') = \min(1, \mathrm{e}^{-\beta(E(X')-E(X))}) \tag{2.21}$$

therefore, if the energy decreases, the movement towards X' is always accepted and performed; otherwise the movement is accepted only with a probability < 1.

[5] This kind of move is less drastic than the single backbone angle update (*pivot move*) and it is intended to avoid highly nonlocal, and therefore unphysical, deformation of the chain, which is likely to be rejected if the chain is compact.

Chapter 3

Free energy reconstruction techniques

In this Chapter we describe the techniques used for reconstructing the free energy profile of a protein model as a function of an internal coordinate, namely the end-to-end distance.
First we introduce the umbrella sampling and the weighted histogram analysis method that are equilibrium techniques; from this last technique it's possible to get the equilibrium free energy profile using a series of *biased* molecular dynamics simulations of the protein constrained by an external potential.
Then we illustrate Jarzynski equality, both in the original and in the extended form we use, that we apply in the context of mechanical unfolding process; the amazing feature of this relation relies on the fact that it states a link between the work done on the protein in an *out-of-equilibrium* process with the *equilibrium* free energy.
Finally we describe how it is possible to use the inherent structures formalism; in order to apply this method two data banks of inherent structures (namely local minima of the potential energy) are built up: a thermal data bank obtained by performing equilibrium canonical simulations and a pulling data bank by mechanically unfolding the protein. We compare all these three techniques for reconstucting the free energy landscape of the minimalistic model, while we use only the extended Jarzynski reconstruction for the protein FnIII$_{10}$ simulated with the all-atom model.

3.1 Umbrella sampling and weighted histogram analysis method

A combination of the umbrella sampling technique [49] with the weighted histogram analysis method (WHAM) [50, 51, 52] allow to obtain the equilibrium free energy profile as a function of the end-to-end distance.

The umbrella sampling technique [49] amounts to perform of a series of biased molecular dynamics simulations of the system constrained by an external potential, namely

$$w_i(\zeta) = \frac{1}{2} k_W (\zeta - \bar{\zeta}_i)^2 \qquad . \tag{3.1}$$

The potential w_i forces the heteropolymer to stay in configurations characterized by a certain average end-to-end distance $\bar{\zeta}_i$, even if at the considered temperature such ζ-value is highly unfavored. These simulations allow to obtain a series of M biased end-to-end probability density distributions $\rho_i^B(\zeta)\{i = 1, \dots, M\}$ and M *unbiasing* relations of the form:

$$\rho(\zeta) = f_i(\zeta) \rho_i^B(\zeta) \tag{3.2}$$

where $\rho(\zeta)$ is the unbiased distribution to be find and $f_i(\zeta)$ is the *unbiasing factor* for the *i*th distribution. The WHAM strategy looks for a linear combination of M independent estimates of $\rho(\zeta)$ obtained from the measured biased distributions $\rho_i^B(\zeta)$, such that the variance $\sigma^2[\rho(\zeta)]$ is miminized. In particular, in the case of identical statistics for each biased run, the WHAM formalism prescribes that the optimal estimator for the distribution of interest is the following combination [50, 51]:

$$\rho(\zeta) = \frac{\sum_{i=1}^{M} \rho_i^B(\zeta)}{\sum_{i=1}^{M} f_i^{-1}} = \frac{\sum_{i=1}^{M} \rho_i^B(\zeta)}{\sum_{i=1}^{M} \mathrm{e}^{-\beta[w_i(\zeta) - F_i]}} = \mathrm{e}^{-\beta f_W(\zeta, T)} \tag{3.3}$$

where $\beta^{-1} = k_B T$ and the free energy constants $\{F_i\}$ can be obtained by the normalization condition

$$\mathrm{e}^{-\beta F_i} = \int d\zeta \ \ \mathrm{e}^{-\beta w_i(\zeta)} \rho(\zeta) \qquad . \tag{3.4}$$

Eqs. 3.3 and 3.4 should be solved self-consistently via an iterative procedure, finally this allows to obtain an estimate of the equilibrium free energy $f_W(\zeta, T)$, apart from an additive constant.

We have considered equally spaced $\{\bar{\zeta}_i\}$-values, with a separation $\Delta\bar{\zeta}_i = 0.2$ among them, ranging from the native configuration ζ_0 to the all *trans*-configuration ζ_{trans} [1] . For each of the M runs, after a quite long equilibration time $t \sim 120,000 - 200,000$, we have estimated $\rho_i^B(\zeta)$ over 100,000 configurations taken at regular time intervals $\Delta t = 0.2$. The biased simulations have been performed with a hard and weak spring, corresponding to $k_W = 10$ and 0.5 in (3.1), respectively. The results obtained essentially agree for the two k_W-values, apart when the free energy landscape exhibits steep increases as a function of ζ. In these cases the hard spring is more appropriate, since the weak one allows the protein to refold, thus rendering the ζ-intervals, where $f_W(\zeta)$ is steeper, not accessible to the WHAM reconstruction.

[1] This is an elongated (planar) equilibrium conformation of the protein with all the dihedral angles at their *trans* values, corresponding to $\zeta_{trans} = 35.70$.

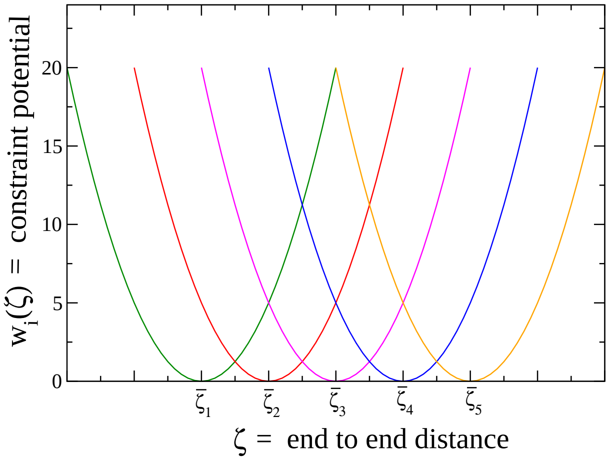

Figure 3.1: Sketch showing the umbrella sampling technique: a series of biased molecular dynamics simulations of the system are performed changing the minimum of the constraint potential.

3.2 Jarzynski equality

Jarzynski equality [4, 5] (JE) relates the difference of the free energy (ΔF) between two *equilibrium* configurations of a system to an ensemble of finite-time measurements of the work performed on the system during an *out-of-equilibrium* process. Therefore the amazing feature of JE relies on the fact that it's possible to get information about an equilibrium quantity from non-equilibrium measurements.
Given a system in a heat reservoir at temperature T; if the system is carried out (see Fig. 3.2a) from the initial *equilibrium* state i at time $t = 0$ to a final state f at time t, with a process during which a *control parameter* λ is switched with an arbitrary rate from λ_i to λ_f, then JE states that:

$$\langle e^{-\beta W} \rangle_t = e^{-\beta \Delta F} \quad , \quad \Delta F = F_{\lambda(t)} - F_{\lambda(0)} \tag{3.5}$$

where:

- the function of the time $\lambda = \lambda(t')$ with $t' \in [0, t]$ defines the so called *manipulation protocol*, lasting for a *finite time* t;
- W is the work done on the system during the process;
- the symbol $\langle ... \rangle_t$ represents an average on repetitions of the *same* experiment (manipulation protocol).

It's necessary to highlight that the control parameter (that can be for example the strenght of an external field, the volume of space where the system is confined, etc.) is a deterministic quantity, *externally* controlled during the process (see Fig. 3.3 for

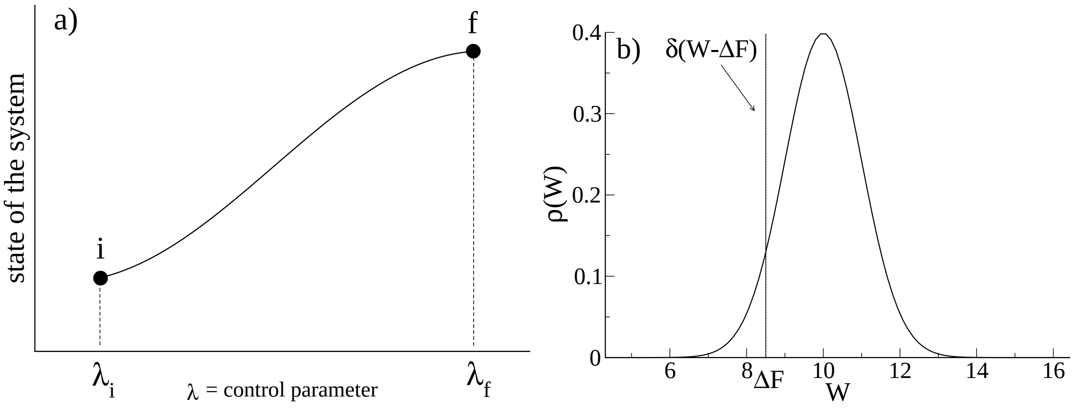

Figure 3.2: (a) Sketch of an out-of-equilibrium process during which a control parameter is switched in a finite time from λ_i to λ_f. (b) Distribution of the work $\rho(W)$ for an out-of-equilibrium process (solid line) and limit ($\delta(W - \Delta F)$ dashed line) for a quasi-static transformation.

a sketch of a pulling experiment with a polypeptide chain). On the other hand the work done on the system at *finite rate* is fluctuating for every different realization, as it depends on the microscopic initial conditions of system and reservoir; starting from different initial conditions (prepared allowing the system to equilibrate with the reservoir) we get a distribution of values for the work, $\rho(W)$.

JE can be seen as a generalization [2] to exponential average of the inequality [62]:

$$\langle W \rangle = \int dW \rho(W) W \geq \Delta F \tag{3.7}$$

where the difference $W_{diss} = \langle W \rangle - \Delta F$ is the *dissipated work* associated with the increase of entropy during the irreversible process. The equal sign in 3.7 holds only in the limit of *infinitely slow* switching (or *quasi-static* transformation):

$$W_\infty = \Delta F \tag{3.8}$$

where the symbol ∞ in the work means a process lasting a time $t \to \infty$; in this case the distribution of work $\rho \to \delta(W - \Delta F)$ (see Fig. 3.2b).

It is worth, finally, to notice that Jarzynski equality is independent of the manipulation protocol used and of the rate of switching of the control parameter but

[2] The inequality 3.7 follows immediately from JE and Jensen's inequality:

$$e^{-\beta\Delta F} = \langle e^{-\beta W} \rangle \geq e^{-\beta \langle W \rangle} \quad \rightarrow \quad \Delta F \leq \langle W \rangle \tag{3.6}$$

only in the limit of infinitely many realizations of the process. In fact, from non-equilibrium experiments or numerical simulations, if we repeat the process a *finite* number of times, N, then:

$$\Delta F \simeq -\frac{1}{\beta}\ln\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{e}^{-\beta W_n}\right] \tag{3.9}$$

where $W_1, W_2, ..., W_N$ represent the work values and the approximation becomes an equality in the limit of $N \to \infty$.

3.3 Extended Jarzynski equality

In Fig. 3.3 it is shown a sketch of a constant velocity pulling experiment with a polypeptide chain; the first bead of the chain is kept fixed and the last bead is attached to a cantilever moving with a constant velocity protocol. In this case the end-to-end distance ζ is an *internal coordinate* (which fluctuates during the pulling experiment); while the distance z between the first bead and the pulling apparatus is the externally controlled parameter (it corresponds to λ parameter of the previous Section 3.2). The spring between the last bead of the chain and the cantilever mimicks the coupling between the system and the device. Using an extended version of the Jarzynski equality (EJE) [12, 13, 14, 15] it's possible to obtain the free energy profile of the system (in this case the polypeptide chain) as a function of an internal collective coordinate. To understand the main differences between the usual JE and the extended version it is worth to highlight that the difference of the free energy ΔF in the JE refers to the *whole* system (in this case the polypeptide chain *and the spring*) and the control parameter λ is not internal but externally controlled.

In this Section we will describe how the extended Jarzynski equality can be obtained in a general way for an arbitrary out-of-equilibrium process; it is straightforward the application to the special case of the pulling experiment showed in Fig. 3.3. Let x be the variable that identifies the system microscopic state, e.g. the collection of the positions and momenta of all the particles in the system $x = \{\boldsymbol{r}_i, \boldsymbol{p}_i\}$. The system Hamiltonian is a function of x, and will be indicated as $H_0(x)$ in the following. Let $X(x)$ be a macroscopic observable of the system whose value can be varied by applying an external force. In the following we will consider only conservative forces, and thus in order to manipulate the system, we can couple it to an external potential $U_\lambda(X)$. This is an explicit function of the observable X and it also depends on a parameter λ, whose value will be modified accordingly to a given time protocol $\lambda(t)$. Thus the system can be characterized by the time-dependent Hamiltonian $H(x,t) = H_0(x) + U_{\lambda(t)}(X(x))$. The work *done* on the system by the external potential (external force) up to time t depends on the trajectory $x(t)$ followed by the system in the phase space:

$$W_t = \int_0^t \mathrm{d}t' \; \dot{\lambda}(t') \; \partial_\lambda U_\lambda(M(x(t'))|_{\lambda=\lambda(t')} \,. \tag{3.10}$$

Due to thermal fluctuations, W_t varies between a realization and another one of the manipulation process.

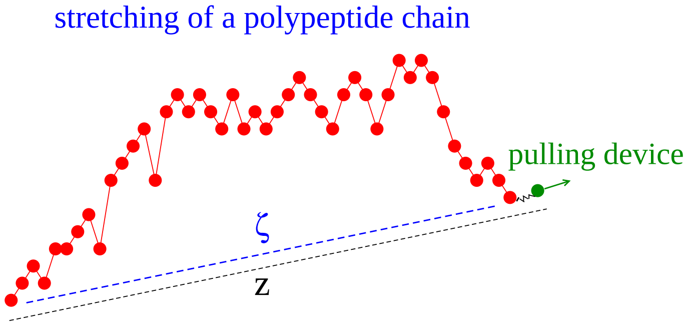

Figure 3.3: Sketch of an out-of-equilibrium process consisting in a constant velocity pulling experiment (for example with AFM) with a polypeptide chain. In this case ζ, the end-to-end distance, is the internal coordinate and z, the distance between the first bead of the chain and the position of the cantilever, is the externally controlled parameter (referred as λ in Section 3.2).

In the following we assume that the time evolution of the system is described by the following Liouville-like equation

$$\frac{\partial p(x,t)}{\partial t} = L_{\lambda(t)}(x) p(x,t), \tag{3.11}$$

where $L_\lambda(x)$ represents the evolution operator ruling the dynamics. The only requirement on L_λ is its compatibility with the canonical distribution associated to the Hamiltonian $H_\lambda(x) = H_0(x) + U_\lambda(M(x))$:

$$L_\lambda \frac{\mathrm{e}^{-\beta H_\lambda(x)}}{Z_\lambda} = 0, \tag{3.12}$$

where $Z_\lambda = \int \mathrm{d}x \; \mathrm{e}^{-\beta H_\lambda(x)}$. Let us introduce the joint probability distribution $\phi(x, W, t)$ that the system is found in state x, having subjected to a work W, at time t. The time evolution of this function is governed by the partial differential equation [12, 13, 14, 15]

$$\frac{\partial \phi}{\partial t} = L_{\lambda(t)}\phi - \dot{\lambda}(t) \left[\partial_\lambda U_{\lambda(t)}(M(x))\right] \frac{\partial \phi}{\partial W}. \tag{3.13}$$

Assuming that the system is initially at equilibrium for vanishing external force, the initial condition of Eq. 3.13 reads $\phi(x, W, t = 0) = p_0(x)$, where $p_0(x)$ is the canonical ensemble probability distribution of the unperturbed system

$$p_0(x) = \frac{\mathrm{e}^{-\beta H_0(x)}}{Z_0}, \tag{3.14}$$

and

$$Z_0 = \int \mathrm{d}x \; \mathrm{e}^{-\beta H_0(x)} \tag{3.15}$$

is the corresponding partition function.

We now introduce the generating function $\psi(x, \lambda, t)$ of the distribution of $\phi(x, W, t)$ defined as

$$\psi(x, \lambda, t) = \int \mathrm{d}W \; \mathrm{e}^{\lambda W} \; \phi(x, W, t). \tag{3.16}$$

The function $\psi(x, \lambda, t)$ satisfies the differential equation

$$\frac{\partial \psi}{\partial t} = L_{\lambda(t)}\psi + \dot{\lambda}(t) \left[\partial_\lambda U_{\lambda(t)}(M(x))\right] \psi. \tag{3.17}$$

By taking $\lambda = -\beta$, and exploiting Eq. 3.12, it is easy to show that Eq. 3.17, with the initial condition $\psi(x, \lambda, t{=}0) = p_0(x)$, is identically satisfied by

$$\psi(x, -\beta, t) = \frac{\mathrm{e}^{-\beta H(x,t)}}{Z_0}. \tag{3.18}$$

Integrating this relation over x one obtains the usual form of the JE:

$$\begin{aligned} \left\langle \mathrm{e}^{-\beta W} \right\rangle_t &= \int \mathrm{d}x \int \mathrm{d}W \; \mathrm{e}^{-\beta W} \; \phi(x, W, t) \\ &= \frac{Z_{\lambda(t)}}{Z_0} = \exp\left[-\beta\left(F_{\lambda(t)} - F_0\right)\right]. \end{aligned} \tag{3.19}$$

Here $\beta F_\lambda = -\ln Z_\lambda$ is the free energy associated to the Hamiltonian H_λ. A more general relation is obtained by multiplying both sides of Eq. 3.18 by $\delta(X - X(x))$ before integrating over x:

$$\begin{aligned} \left\langle \delta(X - X(x))\mathrm{e}^{-\beta W} \right\rangle_t &= \int \mathrm{d}x \; \delta(X - X(x)) \frac{\mathrm{e}^{-\beta H(x,t)}}{Z_0} \\ &= \mathrm{e}^{-\beta\left(f(X,T) + U_{\lambda(t)}(X) - F_0\right)}. \end{aligned} \tag{3.20}$$

Where $f(X, T)$ is the free energy of a constrained ensemble, in which the value $X(x)$ is fixed at X:

$$\beta f(X, T) = -\ln \int \mathrm{d}x \; \delta(X - X(x)) \, \mathrm{e}^{-\beta H_0(x)} \qquad . \tag{3.21}$$

By multiplying both sides of Eq. 3.20 by $\mathrm{e}^{\beta U_{\lambda(t)}(X)}$, we obtain the *extended Jarzynski equality*:

$$\mathrm{e}^{\beta U_{\lambda(t)}(X)} \left\langle \delta(X - X(x))\mathrm{e}^{-\beta W} \right\rangle_t = \mathrm{e}^{-\beta(f(X,T)-F_0)}. \tag{3.22}$$

Equation 3.22 provides thus a method to evaluate the unperturbed free energy $f(X,T)$ as long as one has a reliable estimate of the lhs of this equation. The optimal estimate of $f(X,T)$ can be obtained by combining Eq. 3.22 with the method of weighted histograms [12, 78, 77] (that we have introduced in Section 3.1) as we are going to show below.

For a constant velocity pulling experiment with AFM showed in Fig. 3.3 the variable X corresponds to the end-to-end distance ζ and the externally controlled parameter λ corresponds to z, the distance between the first bead of the chain and the position of the cantilever. In this case the external coupling potential between the protein and the tip of the cantilever (with a spring constant k) is of the form:

$$U_t(\zeta) = \frac{1}{2}k(\zeta - z(t))^2 = \frac{1}{2}k(\zeta - vt)^2 \qquad . \tag{3.23}$$

In this framework we can rewrite JE using the parameters of AFM geometry as:

$$e^{-\beta F(z(t))} = \langle e^{-\beta W_t} \rangle \tag{3.24}$$

where $F(z(t))$ is the free energy difference between the initial state at $t = 0$ and the final state at time t. Please note that $F(z(t))$ is the free energy obtained in presence of the *constraint* potential $U_t(\zeta)$, and so it is a *biased* free energy. In this geometry the extended Jarzynski equality reads as:

$$\mathrm{e}^{-\beta f(\zeta)} = \langle \delta(\zeta - \zeta_t)\mathrm{e}^{-\beta W_t} \rangle \mathrm{e}^{\beta U_t(\zeta)} \tag{3.25}$$

where $f(\zeta)$ is (except for an additive constant) the unperturbed (or unbiased) free energy. The relation 3.25 is of the same form of Eq. 3.2 ($\rho(\zeta) = f_i(\zeta)\rho_i^B(\zeta)$) seen in Section 3.1.

Given the positions of the cantilever $z(t)$ obtained from repeated pulling experiments we can reconstruct $f(\zeta)$ using Eq. 3.25. In fact, at each time slice one can *in principle* get an estimation of the whole (meaning for all the values of the coordinate ζ) free energy. But in practise, as Hummer and Szabo note in [12], at any given time (or equivalently at any given $z(t)$), only a small window around the equilibrium position $z = v \times t$ will be adequately sampled. Therefore an average over several time slices and repeated trajectories is required to obtain an optimal estimate of the whole free energy.

From 3.24 and integrating over ζ the equation 3.25 we can derive the following relation:

$$\mathrm{e}^{-\beta F(z(t))} = \langle \mathrm{e}^{-\beta W_t} \rangle = \langle \mathrm{e}^{-\beta U_t(\zeta)} \rangle \tag{3.26}$$

where the first average is over all the possible realizations of the process and the second average is defined as

$$\langle \mathrm{e}^{-\beta U_t(\zeta)} \rangle \equiv \int d\zeta \mathrm{e}^{-\beta f(\zeta)} \mathrm{e}^{-\beta U_t(\zeta)} \tag{3.27}$$

Therefore using the relation 3.26 the Equation 3.25 can be rewritten as:

$$\mathrm{e}^{-\beta f(\zeta)} = \frac{\langle \delta(\zeta - \zeta_t) \mathrm{e}^{-\beta W_t} \rangle}{\langle \mathrm{e}^{-\beta W_t} \rangle} \mathrm{e}^{\beta U_t(\zeta)} \langle \mathrm{e}^{-\beta U_t(\zeta)} \rangle = \rho_t^B(\zeta) f_t \tag{3.28}$$

where the second equality in 3.28 is written using the WHAM formalism with the unbiasing factor f_t given by:

$$f_t = \mathrm{e}^{\beta U_t(\zeta)} \langle \mathrm{e}^{-\beta U_t(\zeta)} \rangle \tag{3.29}$$

Applying the WHAM procedure (see Section 3.1) we finally get [12]:

$$f_J(\zeta, T) = -\frac{1}{\beta} \ln \left[\frac{\sum_t \frac{\langle \delta(\zeta - \zeta_t) \exp(-\beta W_t) \rangle_t}{\langle \exp(-\beta W_t) \rangle_t}}{\sum_t \frac{\exp(-\beta U(\zeta, t))}{\langle \exp(-\beta W_t) \rangle_t}} \right]. \tag{3.30}$$

where the letter J in $f_J(\zeta, T)$ is meant to distinguish Jarzynski reconstruction from reconstruction of the free energy with different methods.

3.4 Inherent structures formalism

Inherent structures (IS) correspond to local minima of the potential energy, in particular the phase space visited by the protein during its dynamical evolution can be decomposed into a set of disjoint attraction basins, each corresponding to a specific IS. Therefore, the canonical partition function can be expressed within the IS formalism as a sum over the non overlapping basins of attraction, each corresponding to a specific minimum (IS) a [22, 31]:

$$Z_{IS}(T) = \frac{1}{\lambda^{3N'}} \sum_a \mathrm{e}^{-\beta V_a} \int_{\Gamma_a} \mathrm{e}^{-\beta \Delta V_a(\Gamma)} d\Gamma = \sum_a \mathrm{e}^{-\beta [V_a + R_a(T)]} \tag{3.31}$$

where N' is the number of degrees of freedom of the system, λ is the thermal wavelength and Γ represents one of the possible conformations of the protein within the basin of attraction of a, V_a is the potential energy associated to the minimum a, $\Delta V_a(\Gamma) = V(\Gamma) - V_a$ and $R_a(T)$ the vibrational free energy due to the fluctuations around the minimum.

The vibrational term $R_a(T)$ can be estimated by assuming a harmonic basin of attraction:

$$\mathrm{e}^{-\beta R_a(T)} = \frac{1}{\lambda^{3N-6}} \int_{\Gamma_a} \mathrm{e}^{-\beta \Delta V_a(\Gamma)} d\Gamma = \prod_{j=1}^{3N-6} \frac{T}{\omega_a^j} \tag{3.32}$$

where ω_a^j are the frequencies of the vibrational modes around the IS a and unitary reduced Planck and Boltzmann constants have been considered.

Therefore the probability to be in the basin of attraction of the IS a is

$$p_a(T) = \frac{1}{Z_{IS}(T)} \mathrm{e}^{-\beta(V_a + R_a(T))} \qquad . \tag{3.33}$$

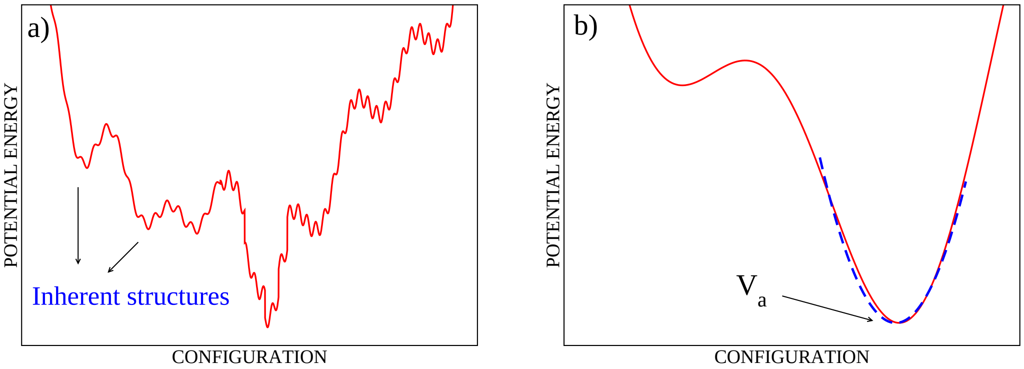

Figure 3.4: (a)Inherent structures of the potential energy. (b)Harmonic approximation (dotted line) of an inherent structure.

The free energy of the whole system at equilibrium is simply given by $f_{IS}(T) = -T \ln[Z_{IS}(T)]$. However if one is interested to construct a free energy landscape as a function of a parameter characterizing the different IS, like e.g. a similarity measure to the native state as the Kabsch distance δ_K[3] or the end-to-end distance ζ, this is possible by defining a partition function restricted to IS with an end-to-end distance within the narrow interval $[\zeta; \zeta + d\zeta]$

$$Z_{IS}(\zeta, T) = \sum_a{}' \mathrm{e}^{-\beta[V_a + R_a(T)]} \tag{3.34}$$

where the $\sum'$ indicates that the sum is not over the whole ensemble of ISs $\{a\}$ but restricted. The free energy profile as a function of ζ can be simply obtained by the relationship:

$$f_{IS}(\zeta, T) = -T \ln[Z_{IS}(\zeta, T)] \qquad ; \tag{3.35}$$

while the average potential and free vibrational energy, corresponding to ISs characterized by a certain ζ, can be estimated as follows:

$$V_{IS}(\zeta, T) = \frac{\sum_a{}' V_a \ \mathrm{e}^{-\beta[V_a + R_a(T)]}}{Z_{IS}(\zeta, T)} \qquad ; \qquad R_{IS}(\zeta, T) = \frac{\sum_a{}' R_a(T) \ \mathrm{e}^{-\beta[V_a + R_a(T)]}}{Z_{IS}(\zeta, T)} \qquad . \tag{3.36}$$

[3] The Kabsch algorithm [75] allows to find analytically the optimal rotation matrix U to minimize the distance $\frac{1}{N} \sum_{k=1}^{N} |U\vec{r}_{k,x} - \vec{r}_{k,y}|^2$ between two conformations of protein and more generally between two point sets $\vec{r}_{k,x}$, $\vec{r}_{k,y}$.

T	goodfolder	badfolder
0.1	2,843	456
0.2	5,875	1,763
0.3	12,359	6,477
0.4	35,409	21,060
0.5	52,546	45,950
0.6	51,971	—
0.7	54,736	—

Table 3.1: Number of distinct ISs contained in the PBD at different temperatures. These have been obtained by sampling, during out-of-equilibrium mechanical unfoldings, several Langevin trajectories at constant elongation increments $\delta\zeta = 0.1$. The total number of relaxations performed for each temperature amounts to $\sim 60,000$ corresponding to ~ 200 repetitions of the same pulling experiment. For the bad folder not all temperatures have been examined.

In order to find the different ISs one can perform Monte Carlo samplings or molecular dynamics simulations. We have chosen to examine molecular dynamic trajectories at constant temperature via a Langevin integration scheme. In particular, we have built up two data banks of ISs: the thermal data bank (TDB) obtained by performing equilibrium canonical simulations and the pulling data bank (PDB) by mechanically unfolding the protein. In order to find the different ISs the equilibrium (resp. out-of-equilibrium) Langevin trajectory is sampled at constant time intervals $\delta t = 5$ (resp. at constant elongation increments $\delta\zeta = 0.1$) to pinpoint a series of configurations, which afterward are relaxed via a steepest descent dynamics [4] and finally refined by means of a standard Newton's method. In the case of the TDB, in order to speed up the search of ISs we have employed a so-called "quasi-Newton" method [76] [5]. For mechanical unfolding, the protein is unblocked and the pulling apparatus removed before the relaxation stage. Two local minima are identified as distinct whenever their energies differ more than 1×10^{-5}. The TDB for the good (resp. bad) folder contains $579,749$ (resp. $210,782$) distinct ISs collected via equilibrium simulations at various temperatures in the range $[0.3; 2.0]$. The PDB contains $3,000 - 50,000$ ISs depending on the examined temperature as detailed in the Table 3.1.

[4] In the steepest descent dynamics the configuration of the protein $\mathbf{r}_i$, $i = 1, L$ is evolved according to the *gradient dynamics*: $\dot{\mathbf{r}}_i = -\frac{1}{\gamma}\nabla_i V$, where V is the intermolecular potential and γ the friction coefficient.

[5] The comparison between the steepest descent and the quasi-Newton methods has revealed that this second minimization scheme is somehow faster (1.8 times faster at $T = 0.5$ for the good folder), but while the steepest descent algorithm is able to identify the metastable stationary states in the 99.8 % of examined cases the quasi-Newton scheme was successful in the 98.7 % of situations. However the distributions of the identified minima (by considering the same trajectory) obtained with the two schemes were essentially coincident.

Chapter 4

Free energy landscape of mechanically unfolded model proteins

In this Chapter we firstly describe the main thermodynamical properties of the two studied sequences of the minimalistic model. Then we compare and discuss the free energy landscape reconstruction as a function of the end-to-end distance obtained in terms of the extended Jarzynski equality, weighted histogram analysis method and inherent structures approach. Futhermore we employ the free energy landscape to characterize the unfolding stages.
From the investigation of the ISs it's possible to get an estimate of the (free) energetic and entropic barriers separating the native from the completely stretched configuration. These barriers are associated to the structural transitions induced by the protein manipulation and for the good folder they can be put in direct relationship with the transition temperatures usually identified during thermal folding/unfolding process (the glassy, the folding and the hydrophobic collapse temperature).

4.1 Thermodynamical properties

The main thermodynamical features of the examined model can be summarized by reporting three different transition temperatures [22, 30, 79, 33, 32]: namely, the hydrophobic collapse temperature T_θ, the folding temperature T_f, and the glassy temperature T_g.
The collapse temperature discriminates between phases dominated by random-coil configurations rather than collapsed ones [80], T_θ has been usually identified as the temperature where the heat capacity $C(T)$ reaches its maximal value, namely (within the canonical formalism):

$$C(T_\theta) \equiv C^{max} \quad , \quad \text{where} \quad C(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{T^2}, \tag{4.1}$$

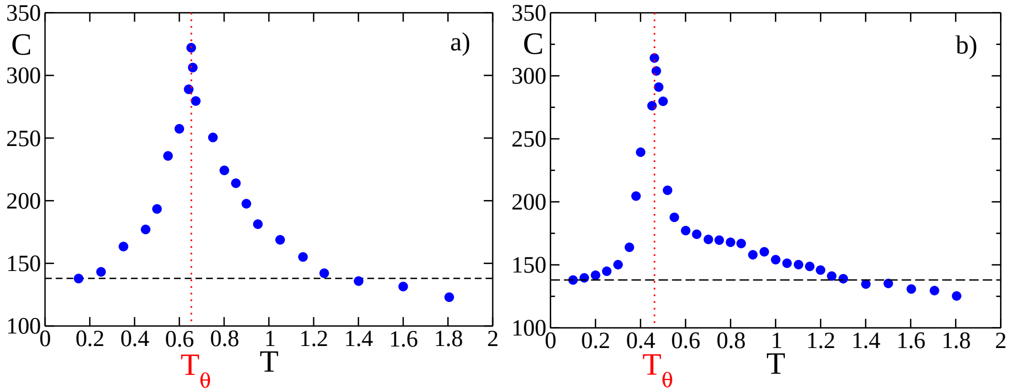

Figure 4.1: Heat capacity C as a function of the temperature T for good (a) and bad (b) folder; the vertical (red) dotted line indicates the hydrophobic collapse temperature T_θ and the horizontal (black) dashed line the value C_{sol}.

and $< \cdot >$ represents a time average performed over an interval $t \simeq 10^5$ by following an US trajectory. From Fig. 4.1, it is evident that for both sequences $C(T) \sim 138$ up to temperatures $T \sim 0.25$. This result can be understood by noticing that at low temperatures the thermal features of heteropolymers resemble that of a disordered 3D solid, with an associated heat capacity $C_{sol} \equiv 3L$. Moreover, the high temperature values are smaller than C_{sol}, since in this limit we expect that a one dimensional chain in a three dimensional space would have a specific heat $C = 2L$ [79]. However, as shown in Fig. 4.1, these extreme temperatures have not yet been reached. The comparison of the heat capacity curves for the GF and BF reveals that $C(T)$ obtained for the GF has a much broader peak with respect to the BF. This indicates that the transition from the NC to the random coil state is definitely sharper for the bad folder.

The folding temperature has been defined in many different ways [65, 69, 79], however we have chosen to define the folding temperature by employing the IS reconstruction of the phase space. In practice, quite long USs have been performed at various temperatures , up to duration $t = 5,000,000$. During each of this US the visited ISs have been identified at regular intervals $\delta t = 5$, and from these data we have estimated the probability $P_{nc}(T)$ to visit the NC at such temperature. The folding temperature T_f (see Fig. 4.2) is then defined as

$$P_{nc}(T_f) \equiv 0.5 \qquad . \tag{4.2}$$

Indeed, it should be noticed that for the GF P_{nc} is the probability to stay in the two lowest lying energy minima (ISs) and not in the NC only. These two minima can be associated to an unique attraction basin, since their energy separation is extremely small with respect to $|V_{NC}|$ (namely, 0.04) and also the corresponding configurations are almost identical, being separated by a Kabsch distance $\delta_K = 0.128$. Moreover, at any examined temperature we have always observed a rapid switching between the

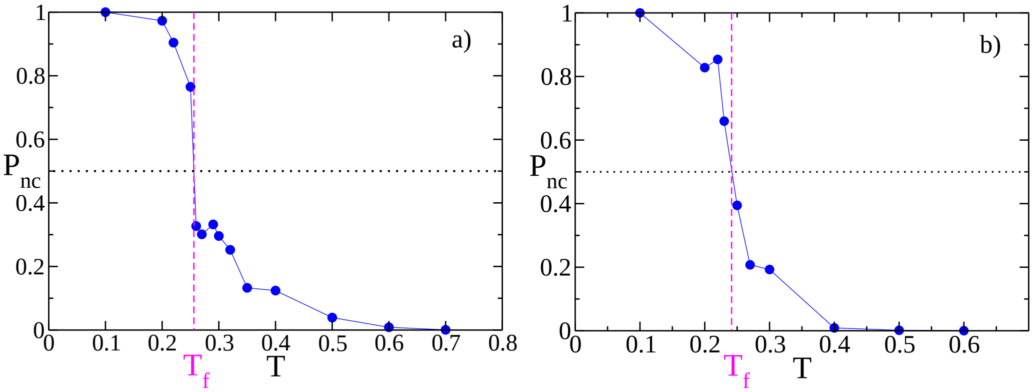

Figure 4.2: Probability P_{nc} as a function of the temperature T for good (a) and bad (b) folder; the vertical (magenta) dashed line indicates the folding temperature T_f, while the horizontal (black) dotted line refers to the value 0.5.

two configurations, indicating that there is an extremely low energy barrier among these two states.

The glassy temperature T_g indicates the temperature below which freezing of large conformational rearrangements occurs: below such a temperature the system can be trapped in local minima of the potential. By following [79], in order to locate T_g we have made a comparison among results obtained from FS and US. In particular, we have examined, at the same temperatures, the average total energy $\langle E \rangle$ of the system evaluated over finite time intervals. As shown in Fig. 4.3, these quantities, when obtained from USs and FSs, coincide at temperatures larger than T_g, below which the structural arrest takes place. In particular, unfolding averages have been performed over intervals of duration $t = 10^5$ by following a single trajectory. On the other hand, folding simulations have been followed up to times $t \simeq 1.1 \cdot 10^7$ and the averages taken over 5-7 different initial conditions by considering for each trajectory only the last time span of duration $t \simeq 5 \cdot 10^4$. The error bars (standard deviation) shown in Fig. 4.3 should be interpreted, at sufficiently low temperatures, as a sign of the dependence of the results on the initial conditions.

The three transition temperatures estimated for the good and bad folder are reported in table 4.1 [1]. One can notice that T_θ is larger for the good folder, thus indicating that the collapsed state has a greater stability with respect to the bad folder. Moreover, while for the good folder $T_f > T_g$, for the bad one this order is reversed. Therefore the BF will most likely remain trapped in some misfolded configurations before reaching the NC even at temperatures $T \sim T_f$.

[1] In [65] for the sequence GF it has been found $T_\theta = 0.65$ and $T_f \sim 0.34$; however in the same paper the authors suggested that the folding transition was associated to a shoulder in the C, but this result has been recently criticized [70]. Moreover, more recent estimates, obtained by employing different protocols, suggest that $T_f \sim 0.24 - 0.25$ [29, 32] and $T_g \sim 0.15$ [32], values that are essentially in agreement with our results

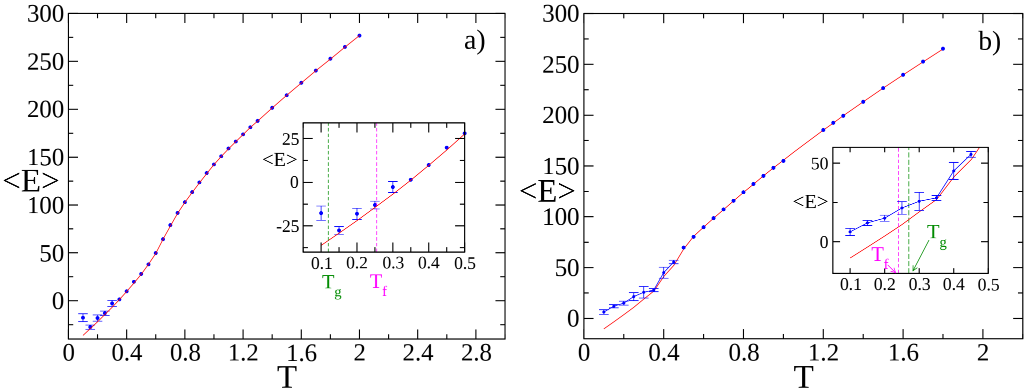

Figure 4.3: Total energy $\langle E \rangle$ as a function of the temperature T for good (a) and bad (b) folder; the solid (red) line corresponds to US's and the (blue) symbols to FS's. In the inset an enlargement for low temperatures: the dashed lines indicate the glassy (T_g) (magenta) and folding (T_f) (green) temperatures.

	GF	BF
T_θ	0.65(1)	0.46(2)
T_f	0.255(5)	0.24(1)
T_g	0.12(2)	0.27(2)

Table 4.1: Transition temperatures estimated for good and bad folder with the corresponding error.

4.2 Extended Jarzynski equality reconstruction

In this section we present for both the sequences GF and BF the reconstruction of the FEL, at various temperatures, as a function of the end-to-end distance ζ starting from out-of-equilibrium measurements. The free energy profiles have been obtained via the EJE by averaging over $28 - 250$ repetitions of the same pulling protocol depending on the pulling velocity as described in Section 2.1.3. We have generally used the pulling configuration where the first bead is kept fixed and the 46th bead is pulled (tail-pulled case). However, by considering the head-pulled case, where the roles of the first and last bead are reversed. we obtain, for sufficiently low velocities (namely, $v_p \leq 5 \times 10^{-4}$ for the GF and $v_p \leq 5 \times 10^{-5}$ for the BF), exactly the same free energy profile (see Section 4.4). These results are essentially in agreement with those reported in [72] for the GF.

4.2.1 Good folder

Fig. 4.4a) shows the EJE reconstructions $f_J(\zeta)$ (symbols) for T=0.3 obtained at various pulling velocities for the good folder together with the corresponding equilibrium WHAM estimate $f_W(\zeta)$ obtained with the procedure described in Section 3.1 (dashed lines). It is worth to stress that *in principle* the Jarzynski reconstruction is independent of the protocol and of the pulling velocity used but only in the limit of infinitely many realizations of the same manipulation experiment. Fig. 4.4a) shows, instead, what happens when the manipulation protocol is repeated a finite number of times and with an almost fixed number of experiments for all the velocities: in this case to get the equilibrium profile, represented by the WHAM reconstruction, it's necessary to use a pulling protocol with a velocity sufficiently low. In particular, for the good folder the equilibrium profile is reached for small ζ-values at a somehow larger velocity (namely, for $\zeta < 10$ already for $v_p = 5 \times 10^{-4}$) than at larger ζ. In particular, to reproduce $f_W(\zeta)$ up to ζ_{trans} the pulling should be performed at $v_p = 5 \times 10^{-6}$. Moreover, referring to Fig. 4.4, it is possible to identify the structural transitions (STs) induced by the pulling experiment. As shown in Fig. 4.4b), the equilibrium $f_J(\zeta)$ profile exhibits a clear minimum in correspondence of the end-to-end distance of the NC (namely, $\zeta_0 \sim 1.9$). In more detail, up to $\zeta \sim 5.6$, the protein remains in native-like configurations characterized by a β-barrel made up of 4 strands, while the escape from the native valley is signaled by the small dip at $\zeta \sim 5.6$ and it is indicated as ST1 in Fig. 4.4b). This ST has been firstly identified in [71] by analyzing the the potential energy of ISs measured during a mechanical unfolding (numerical) experiment. In particular, Lacks [71] identifies this transition as an irreversible transition, in the sense that above this transition it is no more sufficient to reverse the stretching to recover the previously visited configurations [2].

For $\zeta > 6$ the configurations are characterized by an almost intact core (made of 3 strands) plus a stretched tail corresponding to the pulled fourth strand (see configuration (b) in Fig. 4.5a)). The second ST amounts to pull the strand $(PB)_5P$ out of the barrel. In the range $13 < \zeta < 18.5$ the curve $f_J(\zeta)$ appears as essentially flat, thus indicating that almost no work is needed to completely stretch the tail once detached from the barrel (see configuration (c) in Fig. 4.5a)). The pulling of the third strand (that is part of the core of the NC) leads to a definitive destabilization of the β-barrel. This transition is denoted as ST3 in Fig. 4.4a). The second plateau in $f_J(\zeta)$ corresponds to protein structures made up of a single strand (similar to configuration (d) in Fig. 4.5a)).

To distinguish between entropic and energetic costs associated to each ST we have also evaluated separately the potential energy contributions V_i $(i = 1, \ldots, 4)$ during the pulling experiment, these data are reported in Fig. 4.5b). From the

[2] Please notice that we observe this transition at $\zeta \sim 5.6$ and not at $\zeta = 4.782$ as Lacks has reported, since we are considering the free energy profile at $T = 0.3$, while Lacks' analysis concerns potential energies of the ISs. Our inspection of the average potential energies estimated during the pulling experiments and reported in Fig. 4.13a) confirms this small mismatch.

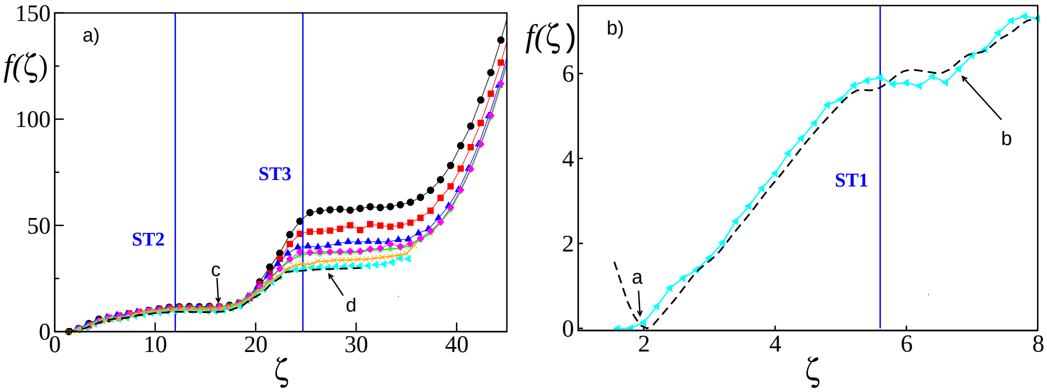

Figure 4.4: (a) Free energy profiles f_J for the good folder as a function of the end-to-end distance ζ at $T = 0.3$, obtained with the EJE for various pulling velocities: from top to bottom $v_p = 5{\times}10^{-2}$, $1{\times}10^{-2}$, $5{\times}10^{-3}$, $5{\times}10^{-4}$, $2{\times}10^{-4}$, $2{\times}10^{-5}$ and $5{\times}10^{-6}$. In (b) an enlargement of the curve for $v_p = 5\times10^{-6}$ at low ζ is reported. The (black) dashed curve in (a) and (b) refers to the WHAM reconstruction $f_W(\zeta)$ with $k_W = 10$. The number of different pulling experiments performed to estimate the profiles ranges between 150 and 250 at the higher velocities and 28 at the lowest velocity $v_p = 5 \times 10^{-6}$. The letters indicate the value of $f(\zeta)$ for the pulled configurations reported in Fig. 4.5a) and the (blue) vertical solid lines the location of the STs.

figure it is clear that the variation of the potential energy during the stretching is essentially due to the Lennard-Jones term V_4, while the other terms contribute to a much smaller extent, at least up to $\zeta \sim 35$. The transition ST1 has essentially only energetic costs, since $\Delta f = 7(1)$ and the potential energy varies almost of the same amount, in particular $\Delta V \sim \Delta V_4 = 8(1)$. The other transitions instead have not negligible entropic costs, since the free energy barrier heights associate to ST2 and ST3 are 10(1) and 29(2), respectively; while the corresponding potential energy barriers are higher, namely $\Delta V = 16(1)$ for ST2 and $\Delta V = 43(1)$ for ST3. The complete stretching of the protein up to $\zeta = 35$ has a free (resp. potential) energy cost corresponding to $\Delta f = 30(2)$ (resp. $\Delta V = 49(1)$). Above $\zeta \sim 35$, while the Lennard-Jones and dihedral contributions vanish, the final (almost quadratic) rise of the free energy is due to the harmonic and angular contributions, since we are now stretching bond distances and angles beyond their equilibrium values. Due to computational constraints and to the fact that this part of the FEL is not particularly relevant, the reconstructions at the lowest velocities and the WHAM estimations have been not performed for these large ζ-values.

In Fig. 4.6 the reconstruction of the FEL obtained at various temperatures is shown. For temperatures around T_f one still observes a FEL resembling the one found for $T = 0.3$, while by increasing the temperature the dip around $\zeta \sim 6 - 7$ (associated to ST1) disappears and the heights of the other two barriers reduce. By approaching T_θ the first plateau, characterizing the transition from the NC to con-

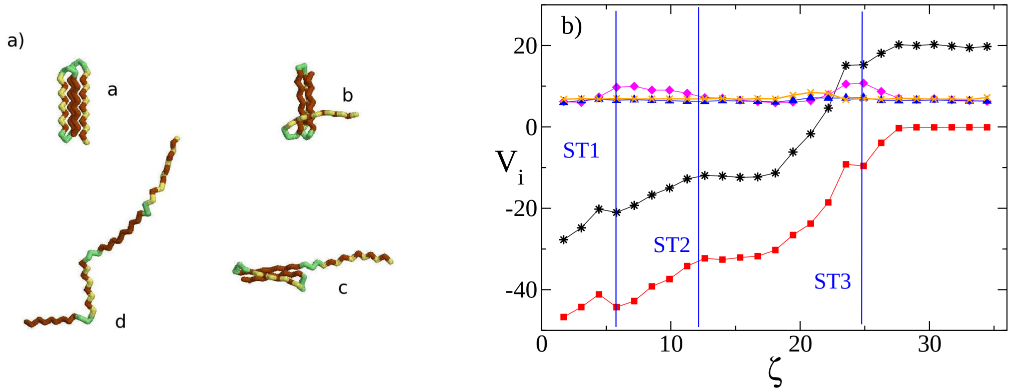

Figure 4.5: (a) Pulled configurations of the good folder at $T = 0.3$: the NC (a) has $\zeta_0 \sim 1.9$; the others are characterized by $\zeta = 6.8$ (b), $\zeta = 16.8$ (c), and $\zeta = 27.1$ (d). The beads of type N, B, and P are colored in green, red and yellow, respectively. (b) Potential energies contributions as a function of the end-to-end distance ζ estimated during a pulling experiment with speed $v_p = 5 \times 10^{-6}$ and obtained by averaging over 28 different realizations at $T = 0.3$. (Black) Stars indicate the entire potential energy V, (orange) crosses V_1, (blue) triangles V_2, (magenta) diamonds V_3, and (red) squares V_4. The (blue) vertical solid lines indicate the transitions previously discussed in the text.

figurations of type (c), essentially disappears, and it is substituted by a monotonous increase of $f_J(\zeta)$. This suggests that 4 stranded β-barrel configurations coexist with partially unfolded ones. Above T_θ only one barrier remains indicating that at these temperatures the protein unfolds completely in one step process.

The connection between dynamical properties of the system and the free energy profile is still an open problem. In particular, the relationship between the unfolding times and the free energy barriers has been previously discussed in Ref. [81] for proteins and more recently the same problem has been addresses for Ising-like lattice protein model in Ref. [82]. We have estimated average first passage times τ via USs by recording the time needed to the protein to reach a certain end-to-end threshold ζ_{th} once it starts from the NC at different temperatures. Our data, reported in Fig. 4.7, clearly indicate that at low temperatures the simple result of the transition state theory [83, 84, 85], namely

$$\tau = \frac{\mathrm{e}^{\Delta f/T}}{T} \qquad , \tag{4.3}$$

where $\Delta f = f(\zeta_{th}) - f(\zeta_0)$, is in very good agreement with the numerics. However, at high temperatures the agreement worsens. Therefore, in order to take in account all the details of the free energy profile and not only the barrier height, we have generalized a result of the Smoluchowski theory for the escape of a particle from a

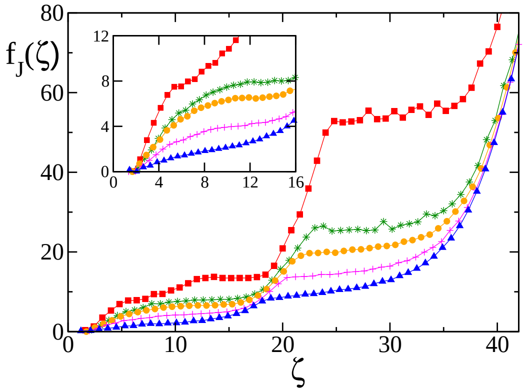

Figure 4.6: Free energy profiles $f_J(\zeta)$ obtained with the EJE for good folder at various temperatures: $T = 0.2$ (red squares), 0.4 (green stars), 0.5 (orange circles), 0.6 (magenta plus) and 0.7 (blue triangles). In the inset, an enlargement is reported at small ζ. Data refer to $v_p = 5 \times 10^{-4}$. The number of different realizations performed to estimate the averages at the different temperatures ranges between 160 and 250.

potential well [85] as follows [82]:

$$\tau \propto \frac{1}{T} \int_{\zeta_0}^{\zeta_{th}} dy \;\; \mathrm{e}^{f(y)/T} \int_{\zeta_0}^{y} dz \mathrm{e}^{-f(z)/T} \tag{4.4}$$

where the potential energy has been substituted by the free energy profile. The estimation obtained via Eq. 4.4 compare well with the numerical results at all the considered temperatures, unfortunately apart an arbitrary scaling factor common to all the temperatures that we are unable to estimate (see Fig. 4.7).

4.2.2 Bad folder

Fig. 4.8a) shows the free energy profiles $f_J(\zeta)$ reconstructed via the EJE at $T = 0.3$ for different pulling speeds (symbols) together with the estimated $f_W(\zeta)$ (dashed line); as in the case of the GF one observes a collapse to the equilibrium FEL (represented by $f_W(\zeta)$) for a sufficiently small speed. In particular, at $v_p = 5 \times 10^{-6}$ a reasonably good agreement between f_J and f_W is already achieved.

For the BF the mechanically induced unfolding transition are less clearly identifiable from the inspection of the free energy profile for two reasons. Firstly, for the BF not only the LJ interactions play a role in the STs but also the dihedral terms: these two terms contribute with opposite signs to the whole potential energy thus partially canceling each other. Moreover, as we will show in the following the main contribution to the free energy is due to entropic terms. Therefore, in order to identify the STs it is better to consider the distinct average profile of the single potential contributions V_i ($i = 1, \ldots, 4$) reported in Fig. 4.8b). In particular,

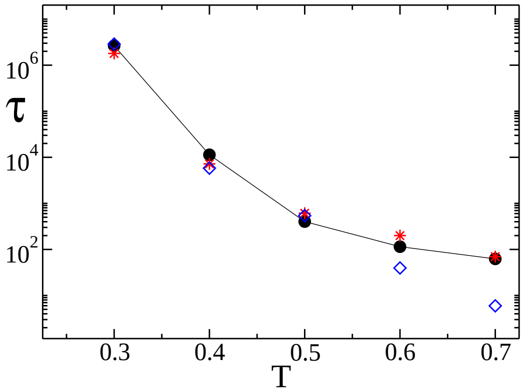

Figure 4.7: Average unfolding times τ for the GF at various temperatures corresponding to $\zeta_{th} = 4$. Filled (black) circles denote the numerical data, the estimations obtained via Eq. 4.3 and Eq. 4.4 are represented by empty (blue) diamonds and (red) stars, respectively. The arbitrary scaling factor entering in Eq. 4.4 (see text) has be set equal to 8. The average times have been estimated over $100,000 - 200,000$ unfolding events for $T = 0.7$ and 0.6, $12,000$ events at $T = 0.5$ and as few as 200 and 60 events at the lowest temperatures, namely $T = 0.4$ and 0.3.

the most relevant is the Lennard-Jones term V_4, due to the stabilizing effect of the hydrophobic interactions on the protein structure. From the inspection of V_4, at least four different STs can be single out, occurring at $\zeta \sim 7.3$, 14.5, 19.3, and 26.3, respectively.

The first transition amounts to pull the last part of the tail out of the NC, namely the 6th and 5th strand that we have previously identified. To this ST is associated a free energy increase of 3.1(5) and a potential energy variation of 8.0(5), once the ST1 is completed the protein assumes the configuration (b) shown in Fig. 4.9. ST2 consists in pulling out from the compact configuration the whole tail (therefore to detach also the 4th strand) and leaving the protein in a configuration composed by the core (represented by the first three strands) plus a long tail (see configuration (c) in Fig. 4.9). The entropic contributions to ST2 is quite relevant since to pass from the NC to (c) the free energy increases of 3.8(5), while the associated potential energy variation is almost the triple, i.e. 11.5(5). The third transition amounts to detach the first β-strand ($BNBPB_3NP$) from the core and this operation has much greater costs with respect to the previous STs, namely, $\Delta f = 7.0(5)$ and $\Delta V = 15(1)$. The complete opening of the core structure (now made only of the second and third strand) occurs at $\zeta \sim 27$ amounting to a total free (resp. potential) energy barrier to overcome of height 11(1) (resp. 23(1)). At variance with the GF case, for the BF the entropic costs are never negligible and instead they always amount at least at the half of the potential energy contributions in all the four examined transitions. Finally, analogously to the GF for $\zeta > 35$ the LJ and dihedral contributions essentially vanish and the free energy increase is due to the harmonic

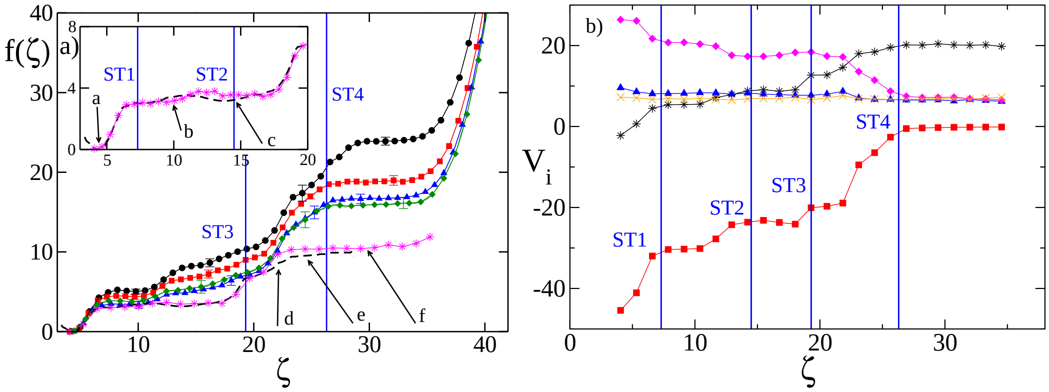

Figure 4.8: (a) Free energy profiles f_J for the bad folder as a function of the end-to-end distance ζ, obtained with the EJE for various pulling velocities: from top to bottom $v_p = 5 \times 10^{-4}$ and 160 realizations (black circles), 2×10^{-4} and 200 realizations (red squares), 1×10^{-4} and 200 realizations (blue triangles), 5×10^{-5} and 100 realizations (green diamonds), 5×10^{-6} and 28 realizations (magenta stars). The WHAM estimate $f_W(\zeta)$ is also shown (black dashed line). In the inset an enlargement of the curve at low ζ for $v_p = 5 \times 10^{-6}$ is reported together with $f_W(\zeta)$. Data have been obtained at $T = 0.3$. (b) Potential energies contributions as a function of the end-to-end distance ζ estimated during a pulling experiment with velocity $v_p = 5 \times 10^{-6}$ and obtained by averaging over 28 different realizations at $T = 0.3$. Black stars indicate the entire potential energy V, (orange) crosses V_1, (blue) triangles V_2, (magenta) diamonds V_3, and (red) squares V_4. The (blue) solid lines indicate the transitions discussed in the text.

and angular terms, only.

In Fig. 4.10 the reconstruction f_J of the FEL for the bad folder is reported at three temperatures below T_θ. As one can notice the bad folder exhibits at comparable temperatures much lower free energy barriers, indicating that the NC and the partially folded structures are less stable, with respect to the GF. This is reflected also in the value of T_θ that has a smaller value with respect to the GF: namely, 0.46 for BF and 0.65 for GF. By increasing T the heights of the free energy barriers rapidly decrease and the various STs become less clearly defined. Moreover, the FEL of the BF at the lower examined temperature ($T = 0.2$) reveals, besides the absolute minimum (corresponding to the NC), other two local minima at $\zeta \sim 7$ and $\zeta \sim 11$. This indicates that, at variance with the GF, the BF can remain trapped even at $T \sim T_f$, for some finite time, in intermediate (misfolded) states far from the NC.

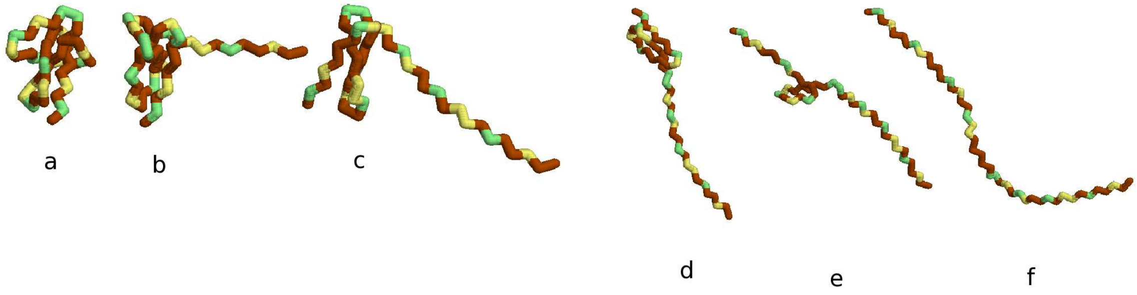

Figure 4.9: Pulled configurations of the bad folder at $T = 0.3$: the reported configurations refer to $\zeta_0 = 4.7$ (NC) (a), $\zeta = 9.9$ (b), 14.5 (c), 22.1 (d), 24.6 (e), and 29.7 (f).

4.3 Inherent structures landscape

In this section we compare the reconstructions of the FEL for the good and bad folder obtained via the EJE and the IS approach with the WHAM equilibrium estimation. As already explained in Section 3.4, we have created two IS data banks: the thermal data bank (TDB) obtained by performing equilibrium canonical simulations and the pulling data bank (PDB) by mechanically unfolding the protein. Fig. 4.11 shows for the GF the comparison, at three temperatures, between the estimate $f_W(\zeta)$ with $f_{IS}(\zeta)$ and the $f_J(\zeta)$, obtained via the EJE reconstruction. The results reveal an astonishingly good coincidence between $f_W(\zeta)$ and $f_{IS}(\zeta)$, obtained by employing the PDB, at all the examined temperatures. For what concerns the EJE reconstructions: at $T = 0.3$ $f_J(\zeta)$ is essentially in good agreement with the other two estimations, while at higher temperatures the f_J curves slightly overestimate the equilibrium free energy f_W for $\zeta > 10$. This discrepancy is probably due to a non complete convergence of the EJE approach at the considered pulling velocities, smaller velocities are required to recover the equilibrium profile at all th end-to-end distances.

The further comparison reported in Fig. 4.11 between the IS reconstructions obtained via the TDB and the PDB indicates a perfect coincidence up to $\zeta \sim 17$. On the contrary, during the last stage of the unfolding process the two f_{IS} differ: the TDB FEL is steeper than the PDB one. This suggests that during the mechanical unfolding the protein can easier reach states with low energies, even at large ζ. These states have a very low probability to be visited during thermal equilibrium dynamics. However, at $T = 0.3$ the value of the barrier to overcome and that of the final plateau are quite similar to those of the PDB FEL, while at higher temperatures the final energy plateaus of the TDB FEL are slightly larger than the f_W-plateaus. The reason of these discrepancies is related to the fact that, despite the high number of IS forming the TDB, this data bank is far from containing all the

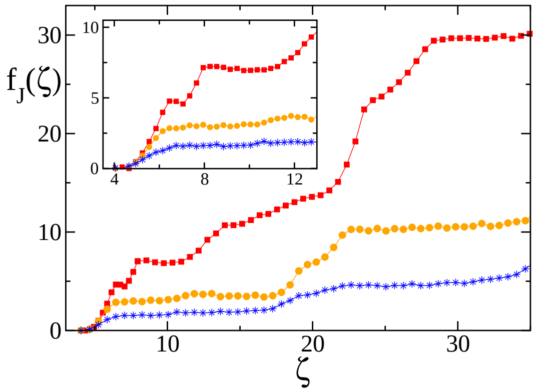

Figure 4.10: Free energy profiles $f_J(\zeta)$ obtained via the EJE for bad folder at three temperatures: namely, $T = 0.2$ (red squares), $T = 0.3$ (orange circles) , $T = 0.4$ (blue stars). In the inset an enlargement is reported at small ζ. Data refer to pulling velocity $v_p = 5 \times 10^{-6}$ and the averages are performed over 28 samples of the same protocol.

relevant ISs, in particular those associated to high ζ-values are lacking. It should be remarked that the IS conformation with the maximal end-to-end distance is the all *trans*-configuration, corresponding to $\zeta_{trans} = 35.70$, therefore the IS approach does not allow to evaluate the FEL for $\zeta > \zeta_{trans}$. For the GF, we can safely affirm that the out-of-equilibrium process consisting in stretching the protein is more efficient to investigate the FEL, since a much smaller number of ISs are needed to reliably reconstruct it, as reported in Table 3.1.

The comparison for the BF case is reported in Fig. 4.12 at $T = 0.3$ and 0.4. Also in this case the $f_W(\zeta)$ and $f_{IS}(\zeta)$ essentially coincide, apart at $T = 0.3$ and $\zeta > 20$ where f_W is slighty higher than f_{IS}. In this case the agreement between the two IS reconstructions is quite good at both the considered temperatures and for all ζ-values. As far as the EJE reconstructions are concerned, at the employed pulling velocity (namely, $v_P = 5 \times 10^{-6}$) f_J can be considered as asymptotic at $T = 0.3$, while probably at $T = 0.4$ is still slightly overestimating f_W, but please notice the really small range of the free energy scale reported in Fig. 4.12b) with respect to the GF.

Furthermore, from the IS analysis by employing Eq. 3.36 we can obtain an estimate of the profiles of the potential and vibrational free energies $V_{IS}(\zeta)$ and $R_{IS}(\zeta)$, respectively. From the latter quantity, the entropic costs associated to the various unfolding stages can be estimated. As shown in Fig. 4.13a), for the GF at $T = 0.3$, the structural transitions ST2 and ST3 previously described correspond to clear "entropic" barriers, while the ST1 transition has only energetic costs since $\Delta R_{IS} \sim 0$. This last result is in good agreement with the previously reported EJE analysis. For what concerns the other two transitions, ST2 (resp. ST3) is associated to a decrease $\sim 6(1)$ (resp. 15(2)) of $R_{IS}(\zeta)$ once more in agreement with the

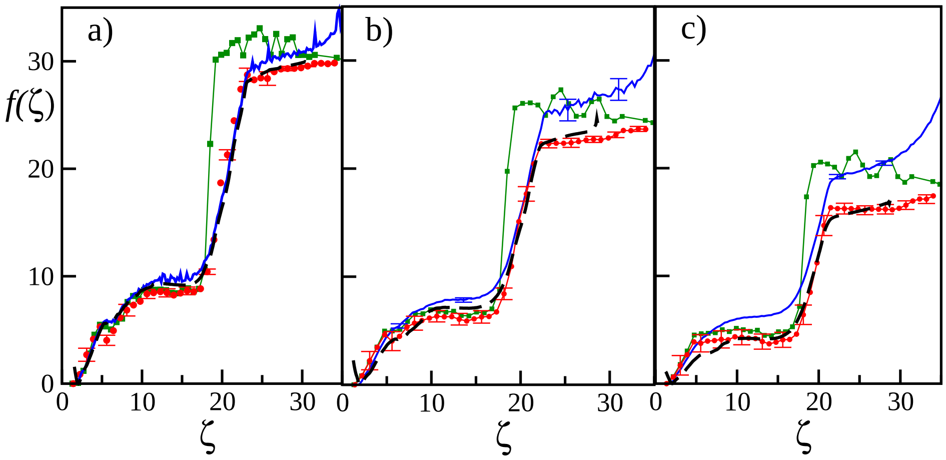

Figure 4.11: Free energy profiles f_J (blue solid lines) as a function of ζ for various temperatures for the good folder: a) $T = 0.3$ for $v_p = 5 \times 10^{-6}$ and 28 repetitions; b) $T = 0.4$ for $v_p = 5 \times 10^{-4}$ and 240 experiments; c) $T = 0.5$ for $v_p = 5 \times 10^{-4}$ and 240 repetitions. The (black) dashed lines refer to the WHAM estimation $f_W(\zeta)$, (green) squares to $f_{IS}(\zeta)$ obtained by employing the TDB and (red) circles to $f_{IS}(\zeta)$ obtained by employing the ISs in the PDB for each considered T.

EJE reconstruction. The complete opening of the protein is associated to a barrier $\Delta R_{IS}(\zeta) = 20(2)$, while the analysis reported in Sect. 4.2 indicates an entropic barrier to overcome corresponding to $\sim 19(2)$. These results suggest that for the good folder the entropic contributions to the free energy are essentially of the vibrational type. Moreover, the reconstructed potential energies $V_{IS}(\zeta)$ are in very good agreement with the average potential energy evaluated during the corresponding pulling experiments as shown in Fig. 4.13a).

Finally, one can try to put in correspondence the three unfolding stages previously discussed for the GF with thermodynamical aspects of the protein folding. In particular, by considering the energy profile $V_{IS}(\zeta)$, an energy barrier ΔV_{IS} and a typical transition temperature $T_t = (2\Delta V_{IS})/(3N)$ can be associated to each of the STs. The first transition ST1 corresponds to a barrier to overcome $\Delta V_{IS} = 8(1)$ and therefore to $T_t = 0.11(1)$, that, within error bars, coincide with T_g. For the ST2 transition the barrier to overcome is $\Delta V_{IS} = 16(1)$ and this is associated to a temperature $T_t \simeq 0.23(2)$ (slightly smaller than T_f). At the ST3 transition $\Delta V_{IS} = 43(2)$ corresponding to $T_t = 0.62(2)$, while the energetic cost to completely stretch the protein is $50(2)$ with an associated transition temperature $T_t = 0.72(2)$: the θ-temperature ($T_\theta = 0.65(1)$) is well bracketed within these two transition temperatures. At least for the GF , our results indicate that the observed STs induced by pulling can be put in direct relationship with the thermal transitions usually identified for the folding/unfolding process.

Also for the BF the IS approach is able to well reproduce not only the average

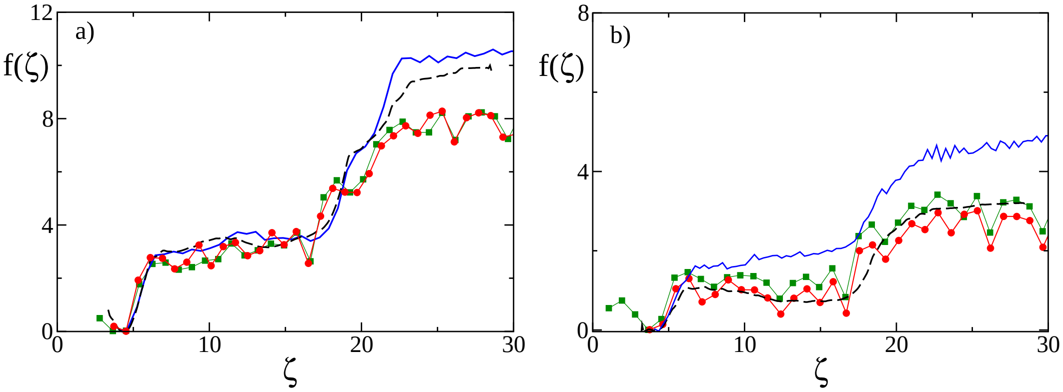

Figure 4.12: Free energy profiles f_J as a function of ζ for various temperatures for the bad folder: a) $T = 0.3$ and b) $T = 0.4$. The data refer to a pulling velocity $v_P = 5 \times 10^{-6}$ and 28 repetitions of the same pulling protocol. The symbols are the same as in Fig. 4.11.

potential energy during the pulling experiment, as clearly shown in Fig. 4.13b), but also to provide a good estimate of the "entropic" barriers associated to the structural transitions. In particular, at $T = 0.3$ the vibrational free energy barriers to overcome are $\Delta R_{IS} = 5.3(5)$ at ST1, 8(1) at ST2, 10(1) at ST3 and 16(1) at ST4. These values are in reasonably good agreement with those previously obtained from the EJE reconstruction, apart at ST3 and ST4, where the analysis performed in Sect. 4.2.2 indicates entropic barriers to overcome corresponding to $\sim 8(1)$ and $\sim 12(2)$, respectively. These underestimations at large ζ-values are probably due to the fact that at this temperature the estimated f_J has not reached its equilibrium profile at the employed velocity.

As already previously pointed out, the entropic contributions for the BF are more relevant than for the GF: e.g while the ST2 transition is clearly visible by the potential energy inspection it is almost absent by looking to the free energy profile (compare the data reported in Fig. 4.8). Therefore we cannot expect to infer information on the thermal transitions from the knowledge of the potential energy barriers at the STs, as done for the GF. Indeed the estimated transition temperatures T_t for the four examined structural transitions give values not corresponding to any of the relevant temperatures reported in Table 4.1 for the BF.

To better understand this difference we have performed USs for the GF and BF for $T_g \lesssim T \lesssim T_\theta$ and we have estimated the average, the minimal and the maximal ζ associated to the visited ISs. The corresponding data are reported in Fig. 4.14. While for the GF the minimal value remains essentially ζ_0 for all the temperatures and the maximum ζ increases smoothly up to ~ 18 at $T = T_\theta$, the dependence of the minimal and maximal ζ-values on T are more dramatic for the BF. Up to the temperatures $T \sim 0.5 \times T_\theta$, average , minimal and maximal ζ-values almost coincide indicating that the protein is still confined around the NC, please remember that for

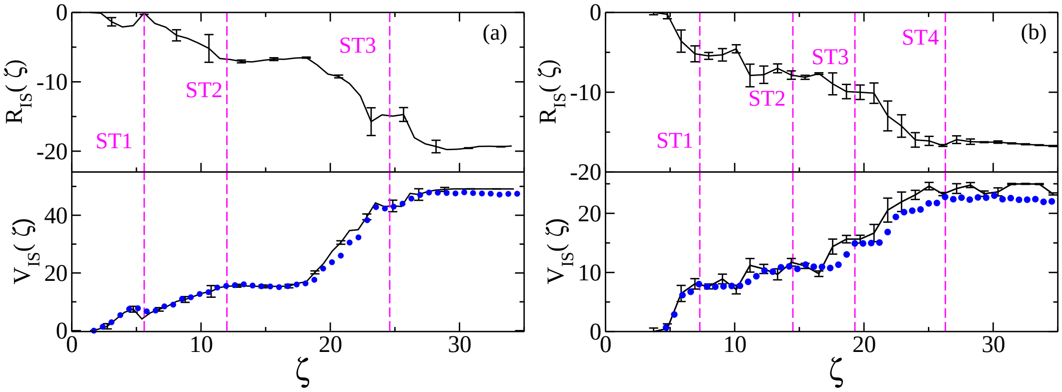

Figure 4.13: Reconstructed $V_{IS}(\zeta)$ (lower panel) and $R_{IS}(\zeta)$ (upper panel) for good folder (a) and bad folder (b) by employing ISs in the PDB at $T = 0.3$. In the lower panel the blue dotted line refers to the average potential energy evaluated during the corresponding pulling experiments (this has been already reported in Fig. 4.5b) for the GF and in Fig. 4.8b) for the BF). Please notice that the data have been vertically translated in order to have zero energy at the NC.

the BF $T_g = 0.58 \times T_\theta$. As soon as $T > 0.6 \times T_\theta$ the maximum grows abruptly and reach the upper bound corresponding to ζ_{trans} already at $T \sim T_\theta$, on the other hand the minimum value decreases indicating that at higher temperatures the protein can access basins of ISs with end-to-end distance lower than ζ_0. This last result indicates that there is not a clear monotonic correspondence between the temperature increase and the achievable protein extensions. Moreover, the fact that the protein can easily attain also extremely stretched configurations at not too high temperatures suggests that in the case of the BF the protein can easily escape form the native valley and reach any part of the phase space, while for the GF the accessible IS configurations are much more limited at comparable temperatures. All this amounts to say that the end-to-end distance cannot be considered as a good reaction coordinate for the BF.

4.4 Tail-pulling versus head-pulling

Since recent experimental results have shown that the pulling geometry influences the mechanical resistance of a protein [60], besides the usual pulling configuration (tail-pulled case) we have also considered a situation where the role of the first and last bead are reversed (head-pulled case). In Fig. 4.15, the results obtained for the tail and head-pulled configuration are compared: we find that the free energy profiles for GF and BF, obtained at $T = 0.3$ by employing the largest speeds ensuring asymptotic results, are identical. This implies that for the considered pulling speed the mechanical unfolding pathways are the same, namely the various strands of the

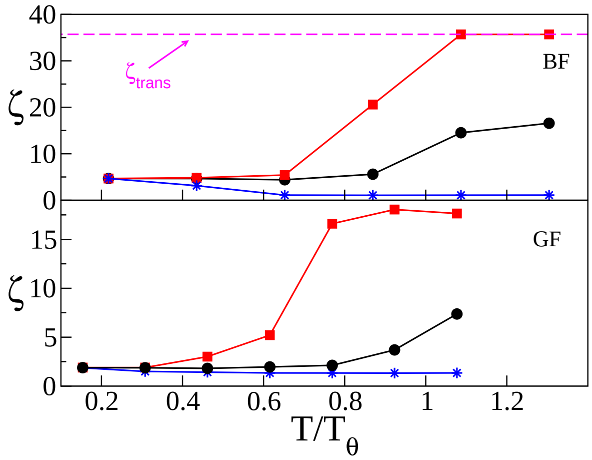

Figure 4.14: End-to-end distance of the ISs estimated during USs at various temperatures: (black) circles represent the average value; (blue) stars the minimal value; and (red) squares the maximal one. The upper panel refer to the BF and the lower one to the GF. The horizontal magenta dashed line indicates the ζ_{trans}-value. For the GF (resp. BF) trajectories of duration $t \sim 100,000 - 500,000$ (resp. $t \sim 50,000 - 250,000$) have been examined to obtain the ISs at constant time intervals $\Delta t = 5$.

protein open up following the same rupture order in both cases. These findings essentially agree with those reported in [72] where the authors have found that the sequence of unfolding events depends on pulling velocities and temperature, moreover they observe that the strands open in the same order for sufficiently low pulling velocities and temperatures[3].

4.5 Concluding remarks

In conclusion, we can safely affirm that the reconstructions of the free energy landscape as a function of the end-to-end distance in terms of the ISs, obtained via out-of-equilibrium mechanical unfolding of the heteropolymers, are in very good agreement with the equilibrium weighted histogram estimate for the good and bad folder sequences at all the examined temperatures. In particular, this result indicates that the harmonic approximation employed to estimate the vibrational term is quite good for temperatures in the range $[T_f; T_\theta]$, as already pointed out in [32] by considering the average potential energy. Moreover, the EJE reconstructions of the free energy profile compare quite well with the other two approaches for sufficiently

[3] The lowest pulling velocity analyzed by the authors in Ref. [72] for the GF was $v_p = 0.36$m/s while the lowest temperature was $T = 20K$. In our adimensional units these values correspond to $v_p \simeq 2.3 \times 10^{-3}$ and $T \simeq 0.17$. Therefore we are considering a much smaller speed, but a somehow larger temperature. For a comparison with physical units see [69].

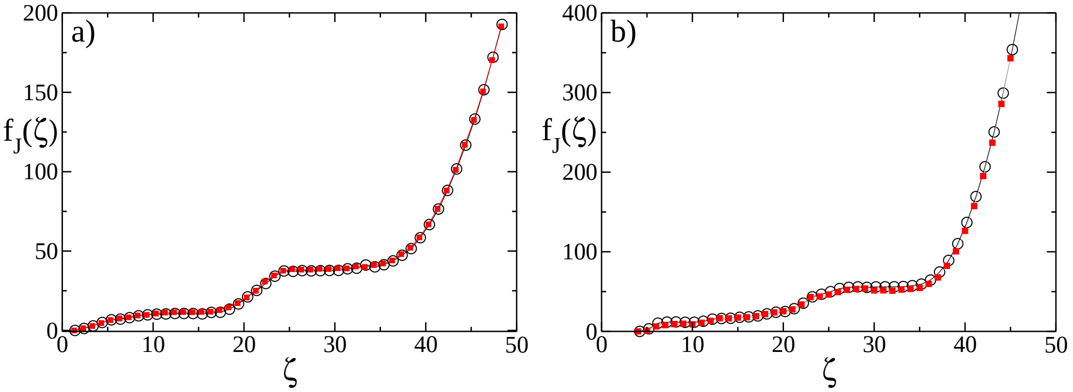

Figure 4.15: Comparison of the EJE reconstruction of the free energy profile for good folder in the tail-pulled (open black circles) and head-pulled (red squares) case. (a) Good folder sequence at T=0.3 and $v_p = 5 \times 10^{-4}$; (b) bad folder sequence at $T = 0.3$ and $v_p = 5 \times 10^{-5}$.

low pulling velocities. For the good folder, the quality of the free energy landscape reconstruction via the extended Jarzinsky equality can be well appreciated by stressing that from pure structural information about the landscape a good estimate of dynamical quantities, like the unfolding times from the native configuration, can be obtained.

Furthermore, for the good folder the information obtained by the equilibrium FEL both with the EJE and the IS methodologies can be usefully combined to give substantiated hints about the thermal unfolding. In particular the investigation of the ISs allows us to give an estimate of the (free) energetic and entropic barriers separating the native state from the completely stretched configuration. These barriers are associated to the structural transitions induced by the protein manipulation and for the good folder they can put in direct relationship with the thermal transitions usually identified during folding/unfolding process.

On the other hand for the bad folder the end-to-end distance appears not to represent a good reaction coordinate, since mechanical and thermal unfolding seem to follow different paths. In other terms the unfolding process for the good folder consists of many small successive rearrangements of the NC, which are well captured by the distribution of the corresponding ISs on the landscape. While for the bad folder the thermal unfolding can involve also large conformational rearrangements, thus implying jumps from one valley to another of the landscape associated to large variations in the end-to-end distance, that cannot be well reproduces by the mechanical stretching of the heteropolymer. Future work on more realistic heteropolymer models is needed to clarify if the observed features, distinguishing good folders from bad folders, can be really considered as a specific trademark of proteins.

A drawback of the EJE reconstruction is that extremely small velocities or an extremely large number of repetitions of the protocol are needed to achieve the collapse towards the equilibrium profile, thus rendering the implementation of the method quite time consuming. However, new optimized methods to obtain the asymptotic FEL, by combining the Jarzinsky equality with the Crooks' path ensemble average theorem, have been recently published [117, 118] and it will be definitely worth to test their performances in the next future with respect to complex landscapes, like those of heteropolymers.

Chapter 5

Mechanical unfolding of FNIII$_{10}$

This Chapter is devoted to the study of mechanical unfolding of a real protein, the tenth type III domain from fibronectin, FnIII$_{10}$, both at constant force and at constant pulling velocity, by all-atom Monte Carlo simulations. We observe both apparent two-state unfolding and several unfolding pathways involving one of three major, mutually exclusive *intermediate states*. All the three major intermediates lack two of seven native β-strands, and share a quite similar extension. The unfolding behavior is found to depend strongly on the pulling conditions. At low constant force or low constant velocity, all the three major intermediates occur with a significant frequency. At high constant force or high constant velocity, one of them dominates over the other two. Using the extended Jarzynski equality, we also estimate the equilibrium free-energy landscape, calculated as a function of chain extension. The application of a constant pulling force leads to a free-energy profile with three major local minima. Two of these correspond to the native and fully unfolded states, respectively, whereas the third one can be associated with the major unfolding intermediates.

5.1 Introduction

Fibronectin is a giant multimodular protein that exists in both soluble (dimeric) and fibrillar forms. In its fibrillar form, it plays a central role in cell adhesion to the extracellular matrix. Increasing evidence indicates that mechanical forces exerted by cells are a key player in initiation of fibronectin fibrillogenesis as well as in modulation of cell-fibronectin adhesion, and thus may regulate the form and function of fibronectin [35, 36].

Each fibronectin monomer contains more than 20 modules of three types, called FnI-II-III. The most common type is FnIII, with ∼90 amino acids and a β-sandwich fold. Two critical sites for the interaction between cells and fibronectin are the RGD motif Arg78-Gly79-Asp80 [86] on the tenth FnIII module, FnIII$_{10}$, and a synergistic site [87] on the ninth FnIII module, which bind to cell-surface integrins. In the native structure of FnIII$_{10}$, shown in Fig. 5.1, the RGD motif is found on the loop connecting the C-terminal β-strands F and G. It has been suggested that a stretching force can change the distance between these two binding sites sufficiently to affect the cell-adhesion properties, without deforming the sites themselves [36]. Force could also influence the adhesion properties by causing full or partial unfolding of the FnIII$_{10}$ module, and thereby deformation of the RGD motif [88]. Whether or not mechanical unfolding of fibronectin modules occurs in vivo is controversial. It is known that cell-generated force can extend fibronectin fibrils to several times their unstretched length [89]. There are experiments indicating that this extensibility is due to changes in quaternary structure rather than unfolding [90], while other experiments indicate that the extensibility originates from force-induced unfolding of FnIII modules [91, 92]. Also worth noting is that the FnIII$_{10}$ module is capable of fast refolding [93].

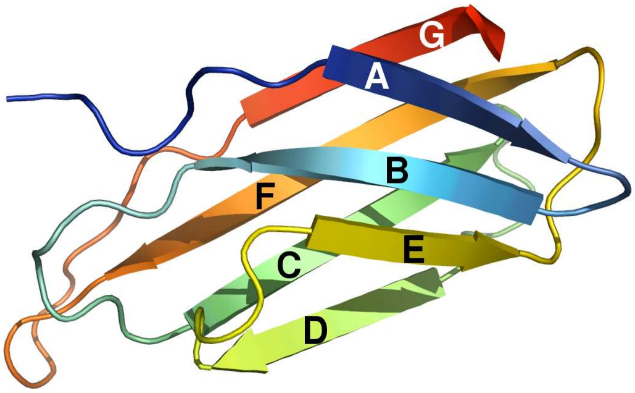

Figure 5.1: Illustration of the native structure of domain 10 of type III of fibronectin, FnIII$_{10}$ (Protein Data Bank ID 1ttf). The letters A-G label its seven β-strands.

Atomic force microscopy experiments have provided important insights into the mechanical properties of FnIII modules [94, 11, 95]. Interestingly, it was found that, although thermodynamically very stable [96], the cell-binding module FnIII$_{10}$ is mechanically one of the least stable FnIII modules [94]. Anyway for the aim of this thesis the most relevant result was found by Li *et al.* [11] where it was shown that the force-induced unfolding of FnIII$_{10}$ using AFM with constant velocity protocol often occurs through *intermediate states*, meaning that partially unfolded but stable structures, capable to opposite to the unfolding forces, are visited in the way from the native to the fully unfolded state. The presence of intermediates states is deduced from the typical sawtooth pattern in the AFM force extension profile, where, also if apparent one-step events (direct path from folded to unfolded configuration) were seen as well, the majority of the unfolding events had a clear two-step character [11].

Several groups have used computer simulations to investigate the force-induced unfolding of FnIII_{10} [88, 97, 98, 99, 100, 101, 102]. An early study predicted the occurrence of intermediate states [97]. In these simulations, two unfolding pathways were seen, both proceeding through partially unfolded intermediate states. Both intermediates lacked two of the seven native β-strands. The missing strands were A and B in one case, and A and G in the other (for strand labels, see Fig. 5.1). A more recent study reached somewhat different conclusions [99]. This study found three different pathways, only one of which involved a partially unfolded intermediate state, with strands A and B detached. The experiments [11] are consistent with the existence of the two different intermediates seen in the early simulations [97], but do not permit an unambiguous identification of the states. When comparing the experiments with these simulations, it should be kept in mind that the forces studied in the simulations were larger than those studied experimentally.

In this work we use the implicit-water all-atom model introduced by Irbäck *et al.* [9, 10], described in Chapter 2, to investigate how the response of FnIII_{10} to a stretching force depends on the pulling strength. We study the unfolding behavior both at constant force and at constant pulling velocity. Some previous studies were carried out using explicit-solvent models [88, 99, 100]. These models might capture important details that the implicit-solvent model we used ignores, like weakening of specific hydrogen bonds through interactions between water molecules and the protein backbone [103]. The advantage of model that we use is computational convenience. The relative simplicity of the model makes it possible for us to generate a large set of unfolding events, which is important when studying a system with multiple unfolding pathways.

Our analysis of the generated unfolding trajectories consists of two parts. The first part aims at characterizing the major unfolding pathways and unfolding intermediates. In the second part, we use the extended Jarzynski equality (EJE) to estimate the equilibrium free-energy landscape, calculated as a function of end-to-end distance. This analysis extends previous work on simplified protein models [16, 17, 18] to an atomic-level model. This level of detail may be needed to facilitate comparisons with future EJE reconstructions based on experimental data. Indeed quite recently this approach has been successfully applied for the first time to data obtained from nanomanipulation of titin I27 domain with atomic force microscopy [20, 21].

5.2 Analysis of pathways and intermediates

To characterize pathways and intermediates, we study the evolution of the native secondary-structure elements along the unfolding trajectories. For this purpose, during the course of the simulations, all native hydrogen bonds connecting two β-strands (see Fig. 5.1) are monitored. A bond is defined as present if the energy of that bond is lower than a cutoff ($-2.4k_{\mathrm{B}}T$). Using this data, we can describe a configuration by which pairs of β-strands are formed (see Section 5.4 for the description of the standard used to label intermediates). A β-strand pair is said to be formed if more than a fraction 0.3 of its native hydrogen bonds are present. Whether individual β-strands are present or absent is determined based on which β-strand pairs the conformation contains.

The characterization of intermediate states requires slightly different procedures in the respective cases of constant force and constant velocity. For constant force simulations (see Fig. 5.2), a histogram of the end-to-end distance ζ, covering the interval $3\,\mathrm{nm} < \zeta < 27\,\mathrm{nm}$, is made for each unfolding trajectory. Each peak in the histogram corresponds to a metastable state along the unfolding pathway. To reduce noise the histogram is smoothed with a sliding ζ window of 0.3 nm. Peaks higher than a given cutoff are identified. Two peaks that are close to each other are only considered separate states if the values between them drop below half the height of the smallest peak. The position of an intermediate, ζ_{I}, is calculated as a weighted mean over the corresponding peak.

In the constant-velocity runs, the unraveling of the native state or an intermediate state is associated with a rupture event, at which a large drop in force occurs (see Fig. 5.3). To ascertain that we register actual rupture events and not fluctuations due to thermal noise, the force versus time curves are smoothed with a sliding time window of $T_{\mathrm{w}} = 0.3\,\mathrm{nm}/v_p$, where v_p is the pulling velocity. Rupture events are identified as drops in force that are larger than 25 pN within a time less than T_{w}. The point of highest force just before the drop defines the rupture force, F_{I}, and the end-to-end distance, ζ_{I}, of the corresponding state. Only rupture events with a time separation of at least $2T_{\mathrm{w}}$ are considered separate events. The rupture force F_{I} is a stability measure statistically easier to estimate than the life time τ_{I} at constant force.

For a peak with a given ζ_{I}, to decide which β-strands the corresponding state contains, we consider all stored configurations with $|\zeta - \zeta_{\mathrm{I}}| < 0.1\,\mathrm{nm}$. All β-strand pairs occurring at least once in these configurations are considered formed in the state. With this prescription, it happens that separate peaks from a single run exhibit the same set of β-strand pairs. Distinguishing between different substates with the same secondary-structure elements is beyond the scope of the present work. Such peaks are counted as a single state, with ζ_{I} set to the weighted average position of the merged peaks.

5.3 Description of the calculated unfolding traces

We study the mechanical unfolding of FnIII$_{10}$ for six constant forces and four constant velocities. Table 5.1 shows the number of runs and the length of each run in these ten cases. At low force or low velocity, it takes longer for the protein to unfold, which makes it necessary to use longer and computationally more expensive trajectories.

Fig. 5.2 shows the time evolution of the end-to-end distance ζ in a representative set of runs at constant force (100 pN). Typically each trajectory starts with a long waiting phase with $\zeta \sim 5$ nm, where the molecule stays close to the native conformation. In this phase, the relative orientation of the two β-sheets (see Fig. 5.1) might change, but all native β-strands remain unbroken. The waiting phase is followed by a sudden increase in the end-to-end distance. This step typically leads either directly to the completely unfolded state with $\zeta \sim 30$ nm or, more commonly, to an intermediate state at $\zeta \sim 12$–16 nm. The intermediate is in turn unfolded in another abrupt step that leads to the completely stretched state. In a small fraction of the trajectories, depending on force, the protein is still in the native state or an intermediate state when the simulation stops. Intermediates outside the range 12–16 nm are unusual but occur in some runs. For example, a relatively long-lived intermediate at 21 nm can be seen in one of the runs in Fig. 5.2.

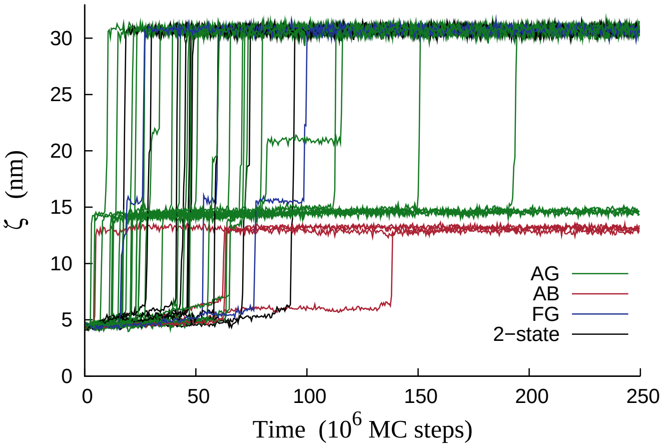

Figure 5.2: MC time evolution of the end-to-end distance (ζ) in 42 independent simulations with a constant pulling force of 100 pN. The three most frequent intermediates lack different pairs of native β-strands: AG, FG, or AB. Trajectories in which these states occur are labeled green, blue and red, respectively. Apparent two-state events are colored black.

Fig. 5.3 shows samples of unfolding traces at constant velocity (0.05 fm/MC step). Here force is plotted against end-to-end distance. As in the constant-force runs, there are two main events in most trajectories. First, the native state is pulled until it ruptures at $\zeta \sim 5$ nm. The chain is then elongated without much resistance until it, in most cases, reaches an intermediate at $\zeta \sim 12$–16 nm. Here the force increases until there is a second rupture event. After that, the molecule is free to elongate towards the fully unfolded state with $\zeta \sim 30$ nm. Some trajectories have force peaks at other ζ. An unusually large peak of this kind can be seen at 22 nm in Fig. 5.3. Inspection of the corresponding structure reveals that it contains a

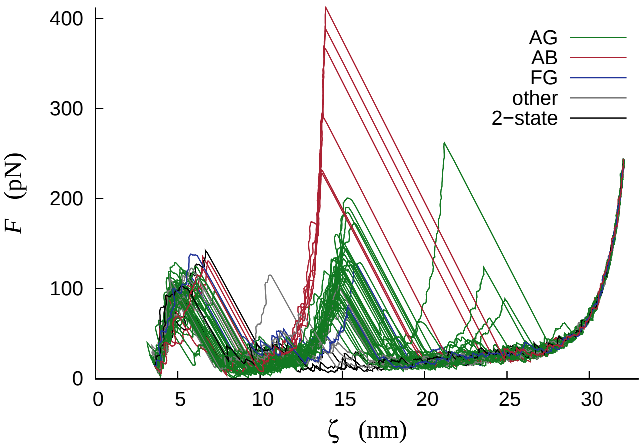

Figure 5.3: Force (F) versus end-to-end distance (ζ) in 55 independent simulations with a constant pulling velocity of 0.05 fm/MC step. Noise has been filtered out using a sliding time window of $6 \cdot 10^6$ MC steps. The color coding is the same as in Fig. 5.2, with the addition of a new category for a few trajectories not belonging to any of the four categories in that figure. These trajectories are colored grey.

three-stranded β-sheet composed of the native CD hairpin and a non-native strand. This sheet is pulled longitudinally, which explains why the stability is high. Another feature worth noting in Fig. 5.3 is that the pulling velocity is sufficiently small to permit the force to drop to small values between the peaks.

There are several similarities between the unfolding events seen at constant force and at constant velocity. In most trajectories, there are stable intermediates, and the unfolding from both the native and intermediate states is abrupt. Also, the vast majority of the observed intermediates have a similar end-to-end distance, in the range 12–16 nm. It should be noticed that experiments typically measure contour-length differences rather than end-to-end distances. Below we analyze contour-length differences between the native state and our calculated intermediates, which turn out to be in good agreement with experimental data.

The trajectories can be divided into three categories: apparent two-state unfolding, unfolding through intermediate states, and trajectories in which no unfolding takes place. Table 5.2 shows the relative frequencies of these groups at the different pulling conditions. The number of trajectories in which the protein remains folded throughout the run obviously depends on the trajectory length. More interesting to analyze is the ratio between the two kinds of unfolding, with or without intermediate states. In the constant-force runs, this ratio depends strongly on the magnitude of the applied force; unfolding through intermediates dominates at the lowest force, but is less common than apparent two-state unfolding at the highest force. In the constant-velocity runs, unfolding through intermediates is much more probable than apparent two-state unfolding at all the velocities studied.

5.4 Identifying pathways and intermediates

The fact that most observed intermediates fall in the relatively narrow ζ interval of 12–16 nm does not mean that they are structurally similar. Actually, the data in Figs. 5.2 and 5.3 clearly indicate that these intermediates can be divided into three groups with similar but not identical end-to-end distances. The β-strand analysis (see Section 5.2) reveals that these three groups correspond to the detachment of different pairs of β-strands, namely A and G, A and B, or F and G. The prevalence of these particular intermediate states is not surprising, given the native topology. When pulling the native structure of FnIII$_{10}$, the interior of the molecule is shielded from force by the N- and C-terminal β-strands, A and G. Consequently, in 95 % or more of our runs, either strand A or G is the first to detach, for all the pulling conditions studied. Most commonly, this detachment is followed by a release of the other strand of the two. But, when A (G) is detached, B (F) is also exposed to force. We thus have three main options for detaching two strands, AG, AB or FG, which actually correspond to the three major intermediates we observe.

Intermediates outside the interval 12–16 nm also occur in our simulations. When applied to the intermediates with $\zeta < 12$ nm, the β-strand analysis identifies two states with one strand detached, A or G. The intermediates with $\zeta > 16$ nm are scattered in ζ and correspond to rare states with more than two strands detached. The intermediate at 21 nm seen in one of the runs in Fig. 5.2 lacks, for example, four strands (A, B, F and G). However, in these relatively unstructured states with more than two strands detached, the remaining strands are often disrupted, which makes the binary classification of strands as either present or absent somewhat ambiguous. Moreover, it is not uncommon that these large-ζ intermediates contain some non-native secondary structure. In what follows, we therefore focus on the five states seen with only one or two strands detached.

For convenience, the intermediates will be referred to by which strands are detached. The intermediate with strands A and B unfolded will thus be labeled AB, etc. Tables 5.3 and 5.4 show basic properties of the A, G, AB, AG and FG intermediates, as observed at constant force and constant velocity, respectively.

From Tables 5.3 and 5.4, several observations can be made. A first one is that the average end-to-end distance, $\bar{\zeta}_{\mathrm{I}}$, of a given state increases slightly with increasing force. More importantly, it can be seen that the relative frequencies with which the different intermediates occur depend strongly on the pulling conditions. At high force or high velocity, the AG intermediate stands out as the by far most common one. By contrast, at low force or low velocity, there is no single dominant state. In fact, at $F = 50$ pN as well as at $v_p = 0.03$ fm/MC step, all the five states occur with a significant frequency.

Table 5.4 also shows the average rupture force, $\bar{F}_{\mathrm{I}}$, of the different states, at the different pulling velocities. Although the data are somewhat noisy, there is a clear tendency that $\bar{F}_{\mathrm{I}}$, for a given state, slowly increases with increasing pulling velocity, which is in line with the expected logarithmic v_p dependence [108]. Comparing the different states, we find that those with only one strand detached (A and G) are markedly weaker than those with two strands detached (AG, AB and FG), as will be further discussed below. Most force-resistant is the AB intermediate. This state occurs much less frequently than the AG intermediate, especially at high velocity, but is harder to break once formed. Compared to experimental data, our $\bar{F}_{\mathrm{I}}$ values for the intermediates are somewhat large. For example the experiments by Li *et al* found a relatively wide distribution of unfolding forces centered at 40–50 pN [11], which is a factor two or more lower than what we find for the AG, AB and FG intermediates. Our results for the unfolding force of the native state are consistent with experimental data. For the native state, the experiments found unfolding forces of 75 ± 20 pN [94] and 90 ± 20 pN [11]. Our corresponding results are 88 ± 2 pN, 99 ± 2 pN and 114 ± 3 pN at $v_p = 0.03$ fm/MC step, $v_p = 0.05$ fm/MC step and $v_p = 0.10$ fm/MC step, respectively.

The AG, AB and FG intermediates do not only require a significant rupture force in our constant-velocity runs, but are also long-lived in our constant-force simulations. In fact, in many runs, the system is still in one of these states when the simulation ends, which means that their average life times, unfortunately, are too long to be determined from the present set of simulations. Nevertheless, there is a clear trend that the AB intermediate is more long-lived than the other two, which in turn have similar life times. The relative life times of these states in the constant-force runs are thus fully consistent with their force-resistance in the constant-velocity runs.

At high constant force, we see a single dominant intermediate, the AG state, but also a large fraction of events without any detectable intermediate. Interestingly, it turns out that the same two strands, A and G, are almost always the first to break in the apparent two-state events as well. Table 5.5 shows the fraction of all trajectories, with or without intermediates, in which A and G are the first two strands to break, at the different forces studied. At 192 pN, this fraction is as large as 98 %. Although the time spent in the state with strands A and G detached varies from run to run, there is thus an essentially deterministic component in the simulated events at high force.

The unfolding behavior at low force or velocity is, by contrast, complex, with several possible pathways. Fig. 5.9 illustrates the relations between observed pathways at the lowest pulling velocity, 0.03 fm/MC step. The main unfolding path begins with the detachment of strand G, followed by the formation of the AG intermediate, through the detachment of A. There are also runs in which the same intermediate occurs but A and G detach in the opposite order. Note that for the majority of the trajectories the boxes A and G in Fig. 5.9 only indicate passage through these states, not the formation of an intermediate state. In a few events, it is impossible to say which strand breaks first. In these events, the initial step is either that the hairpin AB detaches as one unit, or that strands A and G are unzipped simultaneously. Detachment of the FG hairpin in one chunk does not occur in the set of trajectories analyzed for Fig. 5.9. Finally, we note that in the few trajectories where G occurs as an intermediate, the FG intermediate is always visited as well, but never AG. Similarly, the few trajectories where the A intermediate occurs also contain the AG intermediate, but not AB. We find no example where the AB intermediate is preceded by another intermediate.

The unfolding pattern illustrated in Fig. 5.9 can be partly understood by counting native hydrogen bonds. The numbers of hydrogen bonds connecting the strand pairs AB, BE, CF and FG are $n_{\mathrm{AB}} = 7$, $n_{\mathrm{BE}} = 5$, $n_{\mathrm{CF}} = 8$ and $n_{\mathrm{FG}} = 6$, respectively. In our as well as in a previous study [99], two hydrogen bonds near the C terminus break early in some cases, which reduces the number of FG bonds to $n_{\mathrm{FG}} = 4$. The transition frequencies seen in Fig. 5.9 match well with the ordering $n_{\mathrm{BE}} \sim n_{\mathrm{FG}} < n_{\mathrm{AB}} < n_{\mathrm{CF}}$. The first branch point in Fig. 5.9 is the native state. Transitions from this state to the G state, N→G, are more common than N→A transitions, in line with the relation $n_{\mathrm{FG}} < n_{\mathrm{AB}}$. The second layer of branch points is the A and G states. That transitions G→AG are more common than G→GF and that A→AG and A→AB have similar frequencies, match well with the relations $n_{\mathrm{AB}} < n_{\mathrm{CF}}$ and $n_{\mathrm{FG}} \sim n_{\mathrm{BE}}$, respectively. Finally, there are fewer hydrogen bonds connecting the AB hairpin to the rest of the native structure than what is the case for the FG hairpin, $n_{\mathrm{BE}} < n_{\mathrm{CF}}$, which may explain why the AB hairpin, unlike the FG hairpin, detaches as one unit in some runs.

Another feature seen from Fig. 5.9 is that the remaining native-like core rotates during the course of the unfolding process. The orientation of the core is crucial, because a strand is much more easily released if it can be unzipped one hydrogen bond at a time, rather than by longitudinal pulling. The detachment of the first strand leads, irrespective of whether it is A or G, to an arrangement such that two strands are favorably positioned for unzipping, which explains why the intermediates with only A or G detached have a low force-resistance (see Tables 5.3 and 5.4). The AG, AB and FG intermediates, on the other hand, have cores that are pulled longitudinally, which makes them more resistant. Also worth noting is that the core of the AG intermediate is flipped 180°, which is not the case for the AB and FG intermediates.

5.5 Worm-like chain model analysis

The end-to-end distance of the intermediates cannot be directly compared with experimental data. The experiments with AFM by Li *et al.* [11] measured differences of the contour-lengths, L_{c}, rather than ζ, by fitting the constant-velocity data with worm-like chain (WLC) [109] model; this is a polymers elasticity model commonly used in the AFM experiments to retrieve the increment in the protein's lenght between successive rupture events. Using data at our lowest pulling velocity

(0.03 fm/MC step), we now mimic this procedure. For each force peak, we determine a contour length L_c by fitting the WLC expression

$$F = \frac{k_BT}{p}\left[\frac{1}{4(1-d/L_c)^2} - \frac{1}{4} + \frac{d}{L_c}\right] \qquad (5.1)$$

to data. Here p denotes the persistence length [1] and d is the elongation, defined as $d = \zeta - \zeta_N$, where ζ_N is the end-to-end distance of the native state. Following Li *et al.* [11], we use a fixed persistence length of $p = 0.4$ nm.

After each rupture peak follows a region where the force is relatively low. Here it sometimes happens that the newly released chain segment forms α-helical structures, indicating that our system is not perfectly described by the simple WLC model. Nevertheless, the WLC model provides a quite good description of our unfolding traces, as illustrated by Fig. 5.4. The figure shows a typical unfolding trajectory with three force peaks, corresponding to the native (N), intermediate (I) and unfolded (U) states, respectively. From the fitted L_c values, the contour-length differences $\Delta L_c(N \to I)$, $\Delta L_c(I \to U)$ and $\Delta L_c(N \to U)$ can be calculated.

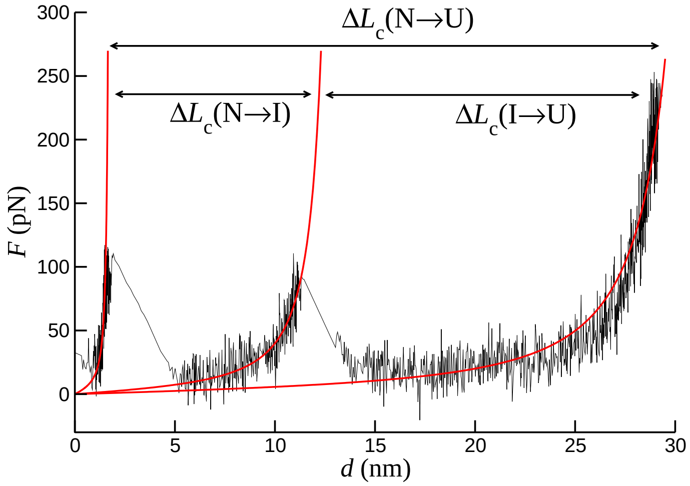

Figure 5.4: WLC fits (Eq. 5.1) to a typical force-extension curve at v_p = 0.03 fm/MC step. The arrows indicate contour-length differences extracted from the fits: $\Delta L_c(N \to I)$, $\Delta L_c(I \to U)$ and $\Delta L_c(N \to U)$.

Fig. 5.5 shows a histogram of $\Delta L_c(N \to I)$, based on our 100 trajectories for v_p = 0.03 fm/MC step. For a small fraction of the force peaks, a WLC fit is not

[1] The persistence length is a mechanical property linked to the stiffness of a polymer and for WLC model it can be shown that it is the characteristic distance along the chain over which the tangent vector correlations die off.

possible; e.g., the A state cannot be analyzed due to its closeness to the native state. All intermediates analyzed have a $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ in the range 6–27 nm. They are divided into five groups: AB, AG, FG, G and "other". Most of those in the category "other" have five strands detached (CDEFG or ABEFG) and a $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ larger than 21 nm. These intermediates were not identified in the experimental study by Li *et al.* [11], which did not report any $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ values larger than 18 nm. These high-ζ intermediates mainly occur as a second intermediate, following one of the main intermediates, which perhaps explains why they were not observed in the experiments. The few remaining intermediates in the category "other" are all of the same kind, ABG, but show a large variation in $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$, from 10 to 19 nm. The small values correspond to states where strand B actually is attached to the structured core, but through non-native hydrogen bonds.

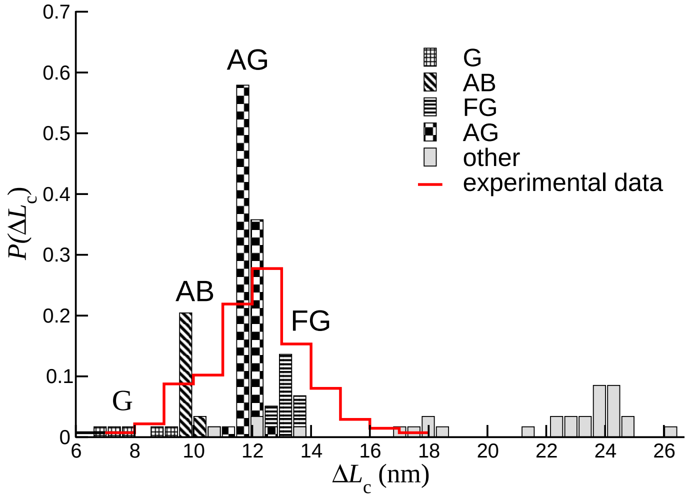

Figure 5.5: Histogram of the contour-length difference $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$, obtained by WLC fits (Eq. 5.1) to our data for $v_p = 0.03\,\mathrm{fm/MC}$ step. A total of 121 force peaks corresponding to intermediate states are analyzed. The intermediates are divided into five groups: AB, AG, FG, G and "other". The experimental $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ distribution, obtained by Li *et al.* [11], is also indicated.

The three major peaks in the $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ histogram (Fig. 5.5) correspond to the AG, AB and FG intermediates. Although similar in size, these states give rise to well separated peaks, the means of which differ in a statistically significant way (see Table 5.6). For comparison, Fig. 5.5 also shows the experimental $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ distribution [11]. The statistical uncertainties appear to be larger in the experiments, because the distribution has a single broad peak extending from 6 to 18 nm. All our $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ data for the AB, AG, FG and G intermediates fall within this region. The occurrence of these four intermediates is thus consistent with the experimental

$\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ distribution. The highest peak, corresponding to the AG intermediate, is located near the center of the experimental distribution.

Transitions from the native state directly to the unfolded state do not occur in the trajectories analyzed for Fig. 5.5. For the contour-length difference between these two states, we find a value of $\Delta L_c(\mathrm{N} \rightarrow \mathrm{U}) = 30.9 \pm 0.1\,\mathrm{nm}$, in perfect agreement with experimental data [11].

5.6 EJE reconstruction

In this section we present the reconstruction of the free-energy profile as a function of the end-to-end distance, $f_J^0(\zeta)$ (where the superscript means *at zero force*), obtained by applying the extended Jarzynski equality, EJE (see Eqs. 3.22 and 3.30), to the constant-velocity trajectories. The number of trajectories analyzed can be seen in Table 5.1. Fig. 5.6 shows the free-energy landscape $f_J^0(\zeta)$ as obtained using different velocities v_p. We observe a collapse of the curves in the region of small-to-moderate ζ. Furthermore, the range of ζ where the curves superimpose, expands as v_p decreases.

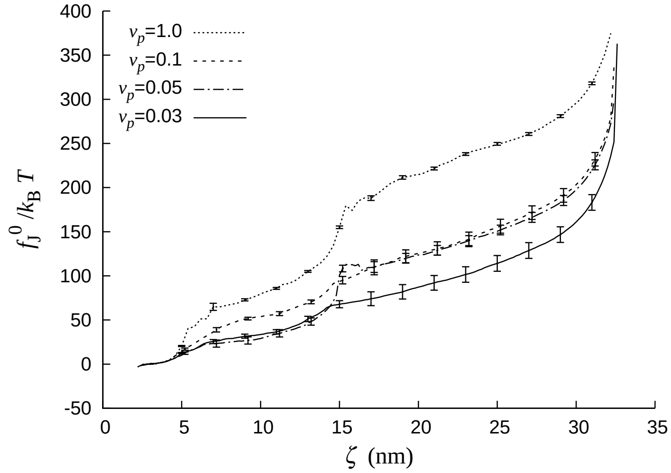

Figure 5.6: EJE reconstruction of the free-energy landscape at zero force $f_J^0(\zeta)$ as a function of the end-to-end distance ζ, using data at different pulling velocities v_p (given in fm/MC step).

As discussed in [17, 18, 19, 107] and in Chapter 4, the collapse of the reconstructed free-energy curves, as the manipulation rate is decreased, is a clear signature of the reliability of the evaluated free-energy landscape. Given our computational resources, we are not able to further decrease the velocity v_p, and for $\zeta > 15$ nm there is still a difference of $\sim 40\,k_B T$ between the two curves corresponding to the lowest velocities. The best estimate we currently have for $f_J^0(\zeta)$ is the curve obtained with $v_p = 0.03\,\mathrm{fm/MC}$ step. This curve will be used in the following analysis.

Let us consider the case where a constant force F is applied to the chain ends. The free energy then becomes $f_J(\zeta) = f_J^0(\zeta) - F\cdot\zeta$. The tilted free-energy landscape $f_J(\zeta)$ is especially interesting for small forces for which the unfolding process is too slow to be studied through direct simulation.

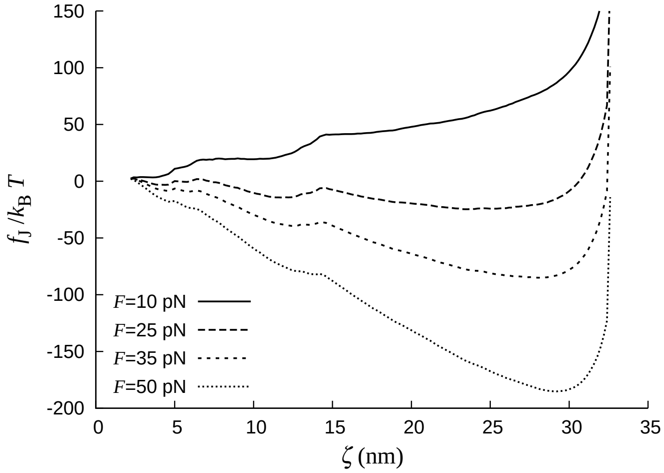

Figure 5.7: Tilted free-energy landscape $f_J(\zeta) = f_J^0(\zeta) - F\cdot\zeta$ for four different forces F. The unperturbed landscape $f_J^0(\zeta)$ corresponds to the curve shown in Fig. 5.6 for $v_p = 0.03\,\mathrm{fm/MC}$ step.

Fig. 5.7 shows our calculated $f_J(\zeta)$ for four external forces in the range 10–50 pN. At $F = 10\,\mathrm{pN}$, the state with minimum free energy is still the native one, and no additional local minima have appeared. At $F = 25\,\mathrm{pN}$, the situation has changed. For $20 \lesssim F \lesssim 60\,\mathrm{pN}$, we find that $f_J(\zeta)$ exhibits three major minima: the native minimum and two other minima, one of which corresponds to the fully unfolded state. The fully unfolded state takes over as the global minimum beyond $F = F_\mathrm{c} \approx 22\,\mathrm{pN}$. The statistical uncertainty on the force at which this happens, F_c, is large, due to uncertainties on $f_J(\zeta)$ for large ζ. For $F = 25\,\mathrm{pN}$, the positions of the three major minima are 4.3 nm, 12 nm and 25 nm. As F increases, the minima move slightly toward larger ζ; for $F = 50\,\mathrm{pN}$, their positions are 4.6 nm, 14 nm and 29 nm. The first two minima become increasingly shallow with increasing F. For $F \gtrsim 60\,\mathrm{pN}$, the only surviving minimum is the third one, corresponding to the completely unfolded state.

These results have to be compared with the analysis above, which showed that the system, on its way from the native to the fully unfolded state, often spends a significant amount of time in some partially unfolded intermediate state with

ζ around 12–16 nm. These intermediates should correspond to local free-energy minima along different unfolding pathways, but in principle it's not obvious that they correspond to local minima of the *one-dimensional* global free energy $f_J(\zeta)$, which is, roughly speaking, based on an average over the full conformational space. As we just saw, we found that it turns out that $f_J(\zeta)$ actually exhibits a minimum around 12–16 nm, where the most common intermediates are sited. It is worth noting that above $\sim$25 pN this minimum gets weaker with increasing force. This trend is in agreement with the results shown in Table 5.2: the fraction of apparent two-state events, without any detectable intermediate, increases with increasing force.

For $F = 25$ pN and $F = 35$ pN, a fourth local minimum can also be seen in Fig. 5.7, close to the native state. Its position is $\approx$6 nm. This minimum is weak and has already disappeared for $F = 50$ pN. It corresponds to a state in which the two native β-sheets are slightly shifted relative to each other and aligned along the direction of the force, with all strands essentially intact. The appearance of this minimum is in good agreement with the results of Gao *et al.* [99]. In their constant force unfolding trajectories, Gao *et al.* saw two early plateaus with small ζ, which in terms of our $f_J(\zeta)$ should correspond to the native minimum and to this $\zeta \approx 6$ nm minimum. In our model, the $\zeta \approx 6$ nm minimum represents a non-obligatory intermediate state; in many unfolding events, especially at high force, the molecule does not pass through this state.

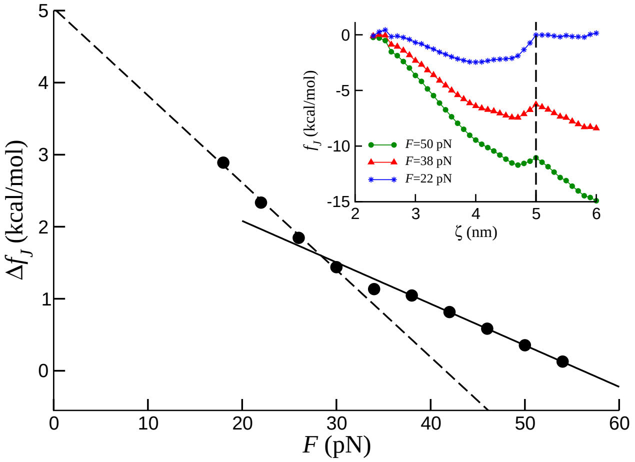

Figure 5.8: Free-energy barrier Δf_J, separating the native state from extended conformations, as a function of the pulling force F. The solid line is a linear fit to the data for forces $F > 25$ pN, while the dashed line refers to a linear fit to the data in the interval $15\,\text{pN} \leq F \leq 30\,\text{pN}$. The inset shows the free energy $f_J(\zeta)$ in the vicinity of the native state for three values of the force (the vertical dashed line indicates the position of the barrier).

Finally, Fig. 5.8 illustrates a more detailed analysis of the native minimum of

$f_J(\zeta)$, for $20\,\text{pN} < F < 60\,\text{pN}$. In this force range, we find that the first barrier is always located at $\zeta = 5.0\,\text{nm}$, whereas the position of the native minimum varies with force (see inset of Fig. 5.8). Hence, the distance between the native minimum and the barrier, x_u, depends on the applied force, as expected [110, 111, 81, 112]. Fig. 5.8 shows the force-dependence of the barrier height, $\Delta f_J(F)$. The solid line is a linear fit with slope $x_\text{u} = 0.4\,\text{nm}$, which describes the data quite well in the force range 25–56 pN. At lower force, the force-dependence is steeper; a linear fit to the data at low force gives a slope of $x_\text{u} = 0.8\,\text{nm}$ (dashed line). Using this latter fit to extrapolate to zero force, we obtain a barrier estimate of $\Delta f_J(0) \approx 5\,\text{kcal/mol}$. Due to the existence of the non-obligatory $\zeta \approx 6\,\text{nm}$ intermediate, it is unclear how to relate this one-dimensional free-energy barrier to unfolding rates. Experimentally, barriers are indirectly probed, using unfolding kinetics. For FnIII_{10}, experiments found a zero-force barrier of 22.2 kcal/mol [94], using kinetics. For the unfolding length, an experimental value of $x_\text{u} = 0.38\,\text{nm}$ was reported [94], based on data in the force range 50–115 pN. Our result $x_\text{u} = 0.4\,\text{nm}$ obtained using the overlapping force range 25–56 pN, is in good agreement with this value.

5.7 Role of pulling strenght in mechanical unfolding

By AFM experiments, Li *et al.* [11] showed that FnIII_{10} unfolds through intermediates when stretched by an external force. AFM data for the wild-type sequence and some engineered mutants were consistent with the existence of two distinct unfolding pathways with different intermediates, one being the AB state with strands A and B detached and the other being either the AG or the FG state [11]. This conclusion is in broad agreement with simulation results obtained by Paci and Karplus [97] and by Gao *et al.* [99].

Comparing our results with these previous simulations, one finds both differences and similarities. In our simulations, three major intermediates are observed: AB, which was seen by Paci and Karplus as well as by Gao et al.; AG, also seen by Paci and Karplus; and FG, which was not observed in previous studies. The most force-resistant intermediate is AB in our as well as in previous studies. Frequencies of occurrence of the intermediates are difficult to compare because the previous studies were based on fewer trajectories. Nevertheless, one may note that the most common intermediate in our simulations, AG, is one of two intermediates seen by Paci and Karplus, and corresponds to one of three pathways observed by Gao et al. A and G often being the first two strands to break is also in agreement with the simulation results of Klimov and Thirumalai [98], who studied several different proteins using a simplified model. Unlike us, these authors found a definite unfolding order for the β-strands. The first strand to break was G, followed by A.

A key issue in our study is how the unfolding pathway depends on the pulling strength. This question was addressed by Gao *et al.* [99]. Based on a simple analytical model rather than simulations, it was argued that there is a single unfolding pathway at low force and multiple unfolding pathways at high force. Our results show the opposite trend. At our lowest force, 50 pN, we observe several different unfolding pathways, and all the three major intermediates occur with a significant frequency. At our highest force, 192 pN, unfolding occurs either in one step or through one particular intermediate, the AG state. Moreover, at 192 pN, the same two strands, A and G, are almost always the first to break in the apparent one-step events as well. Hence, at our highest force, we find that the unfolding behavior has an essentially deterministic component. The trend that the unfolding pathway becomes more deterministic with increasing force can probably be attributed to a reduced relative importance of random thermal fluctuations.

There is a point of disagreement between our results and experimental data, which is that the rupture forces of the three major intermediates are higher in our constant-velocity simulations than they were in the experiments [11]. Although the statistical uncertainties are non-negligible and the pulling conditions are not identical (e.g., we consider a single FnIII_{10} module, while the experiments studied multimodular constructs), we do not see any plausible explanation of this discrepancy. It thus seems that our model overestimates the rupture force of these intermediates. Our calculated rupture force for the native state is consistent with experimental data (see above). To make sure that this agreement is not accidental, we also measured the rupture force of the native state for three other domains, namely FnIII_{12}, FnIII_{13} and the titin I27 domain. AFM experiments (at 0.6 μm/s) found that these domains differ in force-resistance, following the order $\mathrm{FnIII}_{13}\,(\sim 90\,\mathrm{pN}) < \mathrm{FnIII}_{12}\,(\sim 120\,\mathrm{pN}) < \mathrm{I27}\,(\sim 200\,\mathrm{pN})$ [94]. For each of these domains, we carried out a set of 60 unfolding simulations, at a constant velocity of 0.10 fm/MC step. The average rupture forces were 108 ± 4 pN for FnIII_{13}, 135 ± 4 pN for FnIII_{12}, and 159 ± 6 pN for I27, which is in reasonable agreement with experimental data. In particular, our model correctly predicts that the force-resistance of the native state decreases as follows: $\mathrm{I27} > \mathrm{FnIII}_{12} > \mathrm{FnIII}_{13} \sim \mathrm{FnIII}_{10}$. Similar findings have been reported for another model [100].

In this work times have been given in MC steps. In order to roughly estimate what one MC step corresponds to in physical units, we use the average unfolding time of the native state, which is $\sim 4 \cdot 10^8$ MC steps at our lowest force, 50 pN. Assuming that the force-dependence of the unfolding rate is given by $k(F) = k_0 \exp(Fx_\mathrm{u}/k_\mathrm{B}T)$ [113] with $x_\mathrm{u} = 0.38\,\mathrm{nm}$ [94], this unfolding time corresponds to a zero-force unfolding rate of $k_0 \sim 1/(4 \cdot 10^{10}\,\mathrm{MC\ steps})$. Setting this quantity equal to its experimental value, $k_0 = 0.02\,\mathrm{s}^{-1}$ [94], gives the relation that one MC step corresponds to $1 \cdot 10^{-9}$ s. Using this relation to translate our pulling velocities into physical units, one finds, for example, that 0.05 fm/MC step corresponds to 0.05μm/s. This estimate suggests that the effective pulling velocities in our simulations are comparable to or lower than the typical pulling velocity in the experiments by Li *et al.* [11], which was 0.4μm/s. That the effective pulling velocity is low in our simulations is supported by the observation made earlier that the force drops to very small values between the rupture peaks.

The force range studied in our simulations is comparable to that studied in AFM experiments [94, 11, 95]. The exact forces acting on fibronectin under physiological conditions are not known, but might be considerably smaller. For comparison, it was estimated that physiologically relevant forces for the muscle protein titin are $\sim 4\,\mathrm{pN}$ per I-band molecule [115]. For so small forces, the unfolding of FnIII$_{10}$ occurs too slowly in the model to permit direct simulation. Therefore, we cannot characterize unfolding pathways and possible intermediates for these forces. On the other hand, we have an estimate of the free-energy profile $f_J(\zeta)$ for arbitrary force, which can be used, in particular, to estimate the force F_c, beyond which the fully extended state has minimum free energy. Using our best estimate of $f_J(\zeta)$, one finds an F_c of 22 pN (see above), but F_c depends on the behavior of $f_J(\zeta)$ for large ζ, where the uncertainties are large and not easy to accurately estimate. Our estimate of F_c indicates that unfolding of FnIII$_{10}$ to its fully stretched state is a rare event for small stretching forces. The major intermediates are also suppressed compared to the native state for $F \lesssim 10\,\mathrm{pN}$ (see Fig. 5.7). However, our results indicate that the major intermediates are more likely to be observed than the fully stretched state for these forces.

The reconstructed free energies $f_J^0(\zeta)$ and $f_J(\zeta)$ are thermodynamical potentials describing the equilibrium behavior of the system in the absence and presence of an external force F, respectively. For $20 \lesssim F \lesssim 60\,\mathrm{pN}$, this function exhibits three major minima corresponding to the folded state, the most common intermediates, and the fully unfolded state, respectively. However, since $f_J(\zeta)$ describes the system in terms of a single coordinate ζ and "hides" the microscopic configuration, one cannot extract the full details of individual unfolding pathways from this function. For example, one cannot, based on $f_J(\zeta)$, distinguish the AG, AB and FG intermediates, which have quite similar ζ.

The height of the first free-energy barrier, Δf_J, can be related to the unfolding length x_u, a parameter typically extracted from unfolding kinetics, assuming the linear relationship $\Delta f_J(F) = \Delta f_J^0 - F \cdot x_\mathrm{u}$. The parameter x_u measures the distance between the native state and the free-energy barrier, which generally depends on force. Our data for x_u indeed show a clear force-dependence (see inset of Fig. 5.8). However, over a quite large force interval, our x_u is almost constant ($x_\mathrm{u} = 0.4\,\mathrm{nm}$ in the force range 25–56 pN) and similar to its experimental value of $x_\mathrm{u} = 0.38\,\mathrm{nm}$ obtained by [94] in a force range of 50–115 pN.

5.8 Concluding remarks

We have used all-atom MC simulations to study the force-induced unfolding of the fibronectin module FnIII$_{10}$, and in particular how the unfolding pathway depends on the pulling conditions. Both at constant force and at constant pulling velocity, the same three major intermediates were seen, all with two native β-strands missing: AG, AB or FG. Contour-length differences $\Delta L_c(\mathrm{N} \to \mathrm{I})$ for these states were analyzed, through WLC fits to constant-velocity data. We found not only a perfect agreement with the experimenatl data but we also showed that the states, in principle, can be distinguished based on their $\Delta L_c(\mathrm{N} \to \mathrm{I})$ distributions; unfortunately the differences between the distributions are small compared to the resolution of existing experimental data.

The unfolding behavior at constant force was examined in the range 50–192 pN. The following picture emerges from this analysis:

1. At the lowest forces studied, several different unfolding pathways can be seen, and all the three major intermediates occur with a significant frequency.
2. At the highest forces studied, the AB and FG intermediates are very rare. Unfolding occurs either in an apparent single step or through the AG intermediate.
3. The unfolding behavior becomes more deterministic with increasing force. At 192 pN, the first strand pair to break is almost always A and G, also in apparent two-state events.

The dependence on pulling velocity in the constant-velocity simulations was found to be somewhat less pronounced, compared to the force-dependence in the constant-force simulations. Nevertheless, some clear trends could be seen in this case as well. In particular, with increasing velocity, we found that the AG state becomes increasingly dominant among the intermediates. Our results thus suggest that the AG state is the most important intermediate both at high constant force and at high constant velocity.

The response to weak pulling forces is expensive to simulate; our calculations, based on a relatively simple and computationally efficient model, extended down to 50 pN. The reconstruction, based on extended Jarzynski equality, of the free energy $f_J(\zeta)$ opens up a possibility to partially circumvent this problem. Our estimated $f_J(\zeta)$, which matches well with several direct observations from the simulations, indicates, in particular, that stretching forces below 10 pN only rarely unfold FnIII_{10} to its fully extended state, but this conclusion should be verified by further studies, because accurately determining $f_J(\zeta)$ for large ζ is a challenge.

Table 5.1: Number of runs and the length of each run, in number of elementary MC steps, at the different pulling conditions studied.

pulling force or velocity	runs	MC steps/10^6
50 pN	98	1 000
80 pN	100	1 000
100 pN	100	250
120 pN	200	100
150 pN	340	50
192 pN	600	30
0.03 fm/MC step	100	1 167
0.05 fm/MC step	99	700
0.10 fm/MC step	99	350
1.0 fm/MC step	200	35

Table 5.2: The fractions of trajectories in which unfolding occurs either in an apparent two-state manner (labeled $n = 2$) or through intermediate states (labeled $n \geq 3$). "No unfolding" refers to the fraction of trajectories in which the protein remains folded throughout the run (with $\zeta < 8\,\text{nm}$).

pulling force or velocity	$n = 2$	$n \geq 3$	no unfolding
50 pN	0.01	0.79	0.20
80 pN	0.21	0.79	0
100 pN	0.23	0.77	0
120 pN	0.24	0.76	0
150 pN	0.29	0.72	<0.01
192 pN	0.54	0.46	0
0.03 fm/MC step	0.04	0.96	0
0.05 fm/MC step	0.07	0.93	0
0.10 fm/MC step	0.03	0.97	0
1.0 fm/MC step	0	1.0	0

Table 5.3: Frequency f and average extension $\bar{\zeta}_{\rm I}$ (in nm) of intermediate states in the constant-force simulations. The label of a state indicates which β-strands are detached, that is the state AG lacks strands A and G, etc. The frequency f is the number of runs in which a given state was seen, divided by the total number of runs in which unfolding occurred. The statistical uncertainties on $\bar{\zeta}_{\rm I}$ are about 0.1 nm or smaller. "—" indicates not applicable.

	50 pN		80 pN		100 pN		120 pN		150 pN		192 pN	
state	f	$\bar{\zeta}_{\rm I}$	f	$\bar{\zeta}_{\rm I}$	f	$\bar{\zeta}_{\rm I}$	f	$\bar{\zeta}_{\rm I}$	f	$\bar{\zeta}_{\rm I}$	f	$\bar{\zeta}_{\rm I}$
AG	0.46	13.9	0.49	14.3	0.65	14.3	0.69	14.5	0.69	14.6	0.45	14.7
AB	0.35	12.4	0.14	12.9	0.09	13.1	0.03	13.2	<0.01	—	<0.01	—
FG	0.15	14.8	0.13	15.2	0.03	15.5	0.03	15.7	<0.01	—	<0.01	—
G	0.19	11.1	0.04	11.8	0	—	0	—	0	—	0	—
A	0.13	6.7	0	—	0	—	0	—	0	—	0	—

Table 5.4: Frequency f, average rupture force $\bar{F}_{\rm I}$ (in pN) and average extension $\bar{\zeta}_{\rm I}$ (in nm) of intermediate states in the constant-velocity simulations. The statistical uncertainties are 10–20 % on $\bar{F}_{\rm I}$, about 0.1 nm or smaller on $\bar{\zeta}_{\rm I}$ for AG and AB, and about 0.5 nm on $\bar{\zeta}_{\rm I}$ for FG, G and A.

	0.03 fm/MC step			0.05 fm/MC step			0.10 fm/MC step			1.0 fm/MC step		
state	f	$\bar{F}_{\rm I}$	$\bar{\zeta}_{\rm I}$	f	$\bar{F}_{\rm I}$	$\bar{\zeta}_{\rm I}$	f	$\bar{F}_{\rm I}$	$\bar{\zeta}_{\rm I}$	f	$\bar{F}_{\rm I}$	$\bar{\zeta}_{\rm I}$
AG	0.60	115	14.9	0.69	121	14.9	0.78	131	14.8	0.81	198	15.0
AB	0.14	283	13.7	0.09	289	13.8	0.08	333	13.9	0.04	318	13.9
FG	0.15	119	15.6	0.08	107	15.3	0.08	162	16.0	0.04	216	15.7
G	0.05	54	10.5	0.08	73	10.8	0.20	46	9.9	0.06	67	10.3
A	0.06	43	6.2	0.07	53	7.2	0.09	57	6.9	0.03	81	7.2

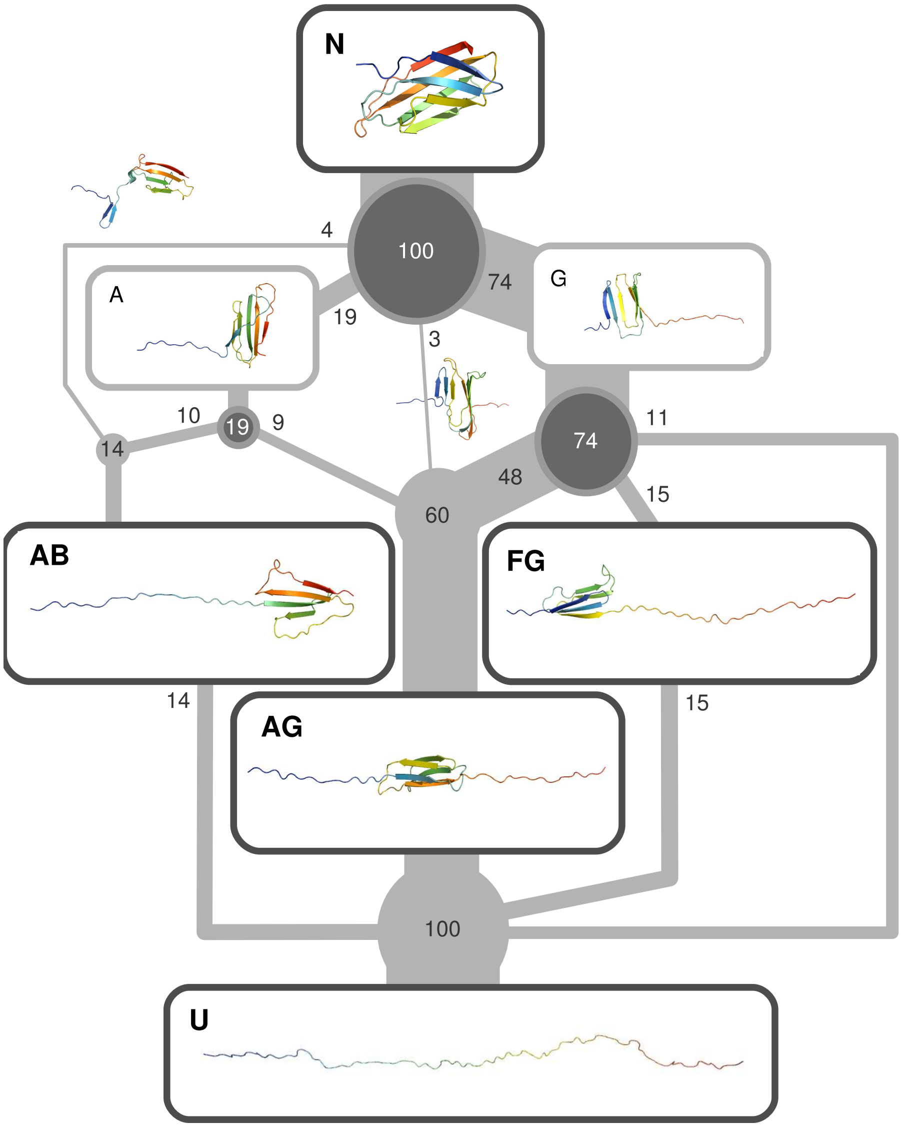

Figure 5.9: Illustration of the diversity of unfolding pathways in the 100 constant-velocity unfolding simulations at $v_p = 0.03\,\mathrm{fm/MC}$ step. The numbers indicate how many of the trajectories follow a certain path. The boxes illustrate important structures along the pathways and boxes with dark rims correspond to the most long-lived states. Dark circles mark branch points. Most trajectories pass through G or A, but only a fraction spend a significant amount of time there (see Table 5.4). The line directly from G to U corresponds to events that either have no intermediate at all or only have intermediates other than the main three. The direct lines N→AB and N→AG describe events that do not clearly pass through A or G and examples of structures seen in those events are illustrated by the unboxed cartoons next to the lines.

Table 5.5: The fractions of all unfolding events in which the first two strands to break are A & G, F & G, and A & B, respectively, at different constant forces. The first pair to break was always one of these three.

first pair	50 pN	80 pN	100 pN	120 pN	150 pN	192 pN
A & G	0.50	0.69	0.87	0.935	0.973	0.980
A & B	0.35	0.15	0.09	0.025	0.006	0.007
F & G	0.15	0.16	0.04	0.040	0.021	0.013

Table 5.6: The average contour-length difference $\overline{\Delta L_{\mathrm{c}}(\mathrm{N} \rightarrow \mathrm{I})}$ for different intermediates, as obtained by WLC fits (Eq. 5.1), to our data for $v_p = 0.03\,\mathrm{fm/MC}$ step.

state	$\overline{\Delta L_{\mathrm{c}}(\mathrm{N} \rightarrow \mathrm{I})}$ (nm)
AG	12.1 ± 0.3
AB	10.1 ± 0.1
FG	13.4 ± 0.3
G	8.2 ± 0.9

Conclusions and perspectives

The original research project of this thesis concerned the numerical study of the mechanical unfolding of two different protein models: the first one is a minimalistic model but, despite its semplicity, it shows the main thermodynamic features of a protein-like behavior; the second one is a realistic model that allowed us the comparison with experimental data.

The first part of this thesis was devoted to the results about the simplified model, that is a variant of the model originally introduced by Honeycutt-Thirumalai [7]. It was simulated via a Langevin molecular dynamics and with a constant pulling velocity protocol. We have studied both a sequence of monomers previously identified as a reasonably fast folder (good folder), and a sequence randomly chosen (bad folder) that it was not expected to have the folding properties of a protein.

In the second part we reported the results concerning the investigation of the mechanical unfolding, both at constant force and at constant pulling velocity, of a real protein, the tenth type III domain from fibronectin, FnIII_{10}, by using an implicit water all-atom model developed by Irbäck and coworkers [9, 10] and simulated via Monte Carlo dynamics. In this case the results of the simulations led to a good agreement with findings coming from available AFM experiments on FnIII_{10}.

For what concerns the simplified model we reconstructed the free energy landscape as a function of an internal reaction coordinate, namely the extension of the chain. At variance with previous studies here we exploited two independent methods: the first one relies on an extended version of the Jarzynski equality (EJE) [12, 13, 14, 15], that links the work done in an out-of-equilibrium process, as the mechanical unfolding actually is, to the difference of the free energy between two equilibrium states; the second method is based on thermodynamic averages over the local minima, or inherent structures (IS), of the potential energy of the protein and on the approximation that such basins of attraction are harmonic. Both methods are compared with the reconstruction performed using a standard equilibrium technique (i.e. umbrella sampling used in conjunction with the weighted histogram analysis method).
In order to apply the EJE reconstruction many realizations of the same constant velocity pulling protocol are performed.

For the ISs approach it was necessary to sample the different ISs and so we built up two data banks of local mimina: the thermal data bank (TDB) obtained by performing equilibrium canonical simulations without any constraints and the pulling data bank (PDB) obtained starting from the configurations visited during the pulling process of the protein. It is known in fact that thermal and mechanical unfolding can follow different pathways in between the native and the completely open configuration and so these two methods can have a different efficiency in sampling the ISs of the potential energy landscape.

We showed that the reconstructions of the free energy landscape in terms of inherent structures, obtained via out-of-equilibrium mechanical unfolding of the heteropolymers, are in very good agreement with the equilibrium umbrella sampling technique, used in conjunction with the weighted histogram method, for the good and bad folder sequences at all the examined temperatures. In particular, this result indicates that the harmonic approximation (employed to estimate the vibrational contribution to the free energy that describes the fluctuations of the protein's configuration in the attraction basin) is quite good for temperatures in the range between the folding and the hydrophobic collapse temperature, as already pointed out in [32] by considering the average potential energy. Moreover, the EJE reconstructions of the free energy profile compare quite well with the other two approaches for sufficiently low pulling velocities.

Furthermore, for the good folder the information obtained by the equilibrium FEL both with the EJE and the IS methodologies can be usefully combined to give substantiated hints about the thermal unfolding. In particular the investigation of the ISs allows us to give an estimate of the (free) energetic and entropic barriers separating the native state from the completely stretched configuration. These barriers are associated to the structural transitions induced by the protein manipulation and for the good folder they can be put in direct relationship with the transition temperatures usually identified during thermal folding/unfolding processes, namely: the glassy temperature, below which the freezing of large conformational rearrangements occurs and so the system can be trapped in local minima of the potential energy without reaching in the finite time the native configuration; the folding temperature, below which the protein stays predominantly in the native valley; and finally the hydrophobic collapse temperature, that discriminates between phases dominated by random-coil configurations and phases with collapsed structures.

It is worth to mention that for the good folder, from the free energy landscape reconstructed as a function of only a single reaction coordinate it was possible to get a good estimate of dynamical quantities, like the unfolding times from the native configuration.

On the other hand for the bad folder from the ISs approach is not possible to get the thermal transition temperatures: in this case the chain extension appears not to represent a good reaction coordinate and this kind of unidimensional reconstruction is not sufficient to completely describe the system's dynamics. Therefore in this case mechanical and thermal unfolding seem to follow different paths. In other terms the unfolding process for the good folder consists of many small successive rearrangements of the native configuration, which are well captured by the distribution of the corresponding ISs on the landscape. While for the bad folder the thermal unfolding can involve also large conformational rearrangements, thus implying jumps from one valley to another of the landscape associated to large variations in the chain extension, that cannot be well reproduced by the mechanical stretching of the heteropolymer.

Anyway further work on more realistic heteropolymer models is needed to clarify if the observed features, distinguishing good folders from bad folders, can be really considered as a specific trademark of proteins.

The main aim of the project concerning the mechanical unfolding of FnIII_{10} was devoted to study how the pulling conditions affect the unfolding pathway followed by this protein from the folded to the completely unfolded conformation. FnIII_{10} has a β-sandwich structure with seven β strands (labelled with letters A-G) and so the unfolding pathway describes the order of rupture of these sub-structures.

Atomic force microscopy experiments have provided important insights into the mechanical properties of FnIII modules [94, 95]. Anyway for the aim of this thesis the most relevant result was found by Li *et al.* [11] where it was shown that the force-induced unfolding of FnIII_{10} using AFM with constant velocity protocol often occurs through *intermediate states*; this means that, during the manipulation process, partially unfolded but *stable* structures are visited. The presence of intermediates states is deduced from the typical sawtooth pattern in the AFM force extension profile, where, also if apparent one-step events (direct path from folded to unfolded configuration) were seen as well, the majority of the unfolding events had a clear two-step character [11].

Several research groups have used computer simulations to investigate the force-induced unfolding of FnIII_{10} [88, 97, 98, 99, 100, 101, 102]. An early study by Paci and Karplus [97] predicted the occurrence of intermediate states. In these simulations two unfolding pathways were seen, both proceeding through partially unfolded intermediate states. Both intermediates lacked two of the seven β-strands of the native structure; but the missing strands were A and B in one case, and A and G in the other. A more recent study by Gao *et al.* [99] reached somewhat different conclusions, because three different pathways were found and only one of which involved a partially unfolded intermediate state, with strands A and B detached. The experiments by Li *et al.*, above cited, are consistent with the existence of the two different intermediates seen in the early simulations by Paci and Karplus, but do not permit an unambiguous identification of the states.

Anyway when comparing the experiments with these simulations it's absolutely crucial to keep in mind that the forces studied in the simulations were larger than those studied experimentally, because the pulling strenght can play a role in changing the mechanical unfolding pathway.

The results obtained in this thesis revealed first of all that both apparent two-state

unfolding and *several* unfolding pathways are present, both at constant force and at constant pulling velocity. The unfolding pathways involve one of three major, mutually exclusive intermediate states, that lack two of the seven native β-strands and share a quite similar extension. The unfolding behavior is found to depend strongly on the pulling conditions. In particular, we observe large variations in the relative frequencies of occurrence for the intermediates. At low constant force or low constant velocity, the behavior of the system is characterized by a wide variability, meaning that several different unfolding pathways can be seen and all the three major intermediates occur with a significant frequency. On the other hand at high constant force or high constant velocity, one of them dominates over the other two, and so in this regime the unfolding behavior becomes more deterministic. To compare the numerical results obtained via the constant velocity pulling protocol with the results of Li *et al.* we used the worm-like chain (WLC) model analysis that is a standard technique used in the AFM experiments to retrieve the increment in the protein's lenght between successive unfolding events represented by the rupture peaks in the sawtooth pattern. One of the main result of this thesis is the good agreement obtained with the experimental data of Li *et al.*; moreover we found that these intermediates states, in principle, can be distinguished using the analysis of the increment in the protein's lenght, but the differences are small compared to the resolution of existing experimental data.
As a further test of the obtained results, from the constant velocity unfolding trajectories and using the extended Jarzynski equality, we also estimated the equilibrium free-energy landscape as a function of chain extension for FnIII_{10} as we did for the simplified model. Once we reconstructed the *zero force* landscape, the application of a constant pulling force leads to *tilted* free-energy profile wich exhibits three major local minima: two of these correspond to the native and fully unfolded states, respectively, whereas the third one can be associated with the major unfolding intermediates found with the direct observation of the unfolding trajectories from the simulations.

We would like to remember that, in the context of glassy systems, the concept of ISs has been critically compared to that of pure states [119], the latter being local minima of the free energy landscape, while the ISs are minima of the potential energy, as discussed above. In FnIII_{10} mechanical unfolding the relevance of pure states for protein dynamics has been shown by putting them in direct correspondence with unfolding intermediates observable in pulling trajectories.
One of the main technique used in this thesis for reconstructing the free energy profile was the extended Jarzynski equality. Anyway a drawback of the EJE reconstruction is that extremely small velocities or an extremely large number of repetitions of the protocol are needed to achieve the collapse towards the equilibrium profile, thus rendering the implementation of the method quite time consuming. However, new optimized methods to obtain the equilibrium FEL, by combining the Jarzynski equality with the Crooks' path ensemble average theorem, have been recently pub-

lished [117, 118] and it will be definitely worth to test their performances in the next future with respect to complex landscapes, like those of heteropolymers. The power of these new *bidirectional methods*, that can also be applied to pulling experiments on real proteins, relies on the optimal combination of pulling trajectories got in the *forward* process, from the native to the stretched structure, in conjunction with the pulling trajectories obtained in the *backward* process, where the protein is driven from the unfolded towards the compact conformation.

Bibliography

[1] T.E. Creighton, *Proteins*, W. H. Freeman and Company, sixth printing (2002).

[2] J. Kurchan, J. Stat. Mech. P07005 (2007).

[3] G.E. Crooks, Phys. Rev. E **60** , 2721 (1999).

[4] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[5] C. Jarzynski, Phys. Rev. Lett. **56**, 5018 (1997).

[6] J. Liphardt, S. Dumont, S.B. Smith, I. Tinoco Jr. and C. Bustamante, Science **296**, 1832 (2002).

[7] J.D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990).

[8] R.S. Berry, N. Elmaci, J.P. Rose, and B. Vekhter, Proc. Natl. Acad. Sci. USA **94**, 9520 (1997).

[9] A. Irbäck, B. Samuelsson, F. Sjunnesson, and S. Wallin, Biophys. J. **85**, 1466-1473 (2003).

[10] A. Irbäck and S. Mohanty, Biophys. J. **88**, 1560-1569 (2005).

[11] L. Li, H.H.-L. Huang, C.L. Badilla, and J.M. Fernandez, J. Mol. Biol. **345**, 817–826 (2005).

[12] G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. USA **98**, 3658 (2001).

[13] T. Speck and U. Seifert, Phys. Rev. E **70**, 066112 (2004).

[14] A. Imparato and L. Peliti, Europhys. Lett. **70**, 740-746 (2005).

[15] A. Imparato and L. Peliti, Phys. Rev. E **72**, 046114 (2005).

[16] D.K. West, P.D. Olmsted, and E. Paci, J. Chem. Phys. **125**, 204910 (2006).

[17] A. Imparato, A. Pelizzola, and M. Zamparo, Phys. Rev. Lett. **98**, 148102 (2007).

[18] A. Imparato, A. Pelizzola, and M. Zamparo, J. Chem. Phys. **127**, 145105 (2007).

[19]

[20] N.C. Harris, Y. Song, and C.-H. Kiang, Phys. Rev. Lett. **99** 068101 (2007).

[21] A. Imparato, F. Sbrana, and M. Vassalli, *EPL* **82**, 58006 (2008).

[22] D.J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.

[23] F.H. Stillinger and T.A. Weber, Science **225**, 983 (1984).

[24] S. Sastry, P.G. Debenedetti, and F.H. Stillinger, Nature **393**, 554 (1998).

[25] L. Angelani, R. Di Leonardo, G. Ruocco, A. Scala, and F. Sciortino, Phys. Rev. Lett. **85**, 5356 (2000).

[26] Z. Guo and D. Thirumalai, J. Mol. Biol., **263**, 323 (1996).

[27] M.A. Miller and D.J. Wales, J. Chem. Phys. **111**, 6610 (1999); P.N. Mortenson and D.J. Wales, J. Chem. Phys. **114**, 6443 (2001); P.N. Mortenson, D.A. Evans and D.J. Wales, J. Chem. Phys. **117**, 1363 (2002).

[28] S.V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).

[29] D.A. Evans and D.J. Wales, J. Chem. Phys. **118**, 3891 (2003).

[30] A. Baumketner, J.-E. Shea, and Y. Hiwatari, Phys. Rev. E **67**, 011912 (2003).

[31] N. Nakagawa and M. Peyrard, Proc. Natl. Acad. Sci. USA **103**, 5279 (2006); Phys. Rev. E **74**, 041916 (2006).

[32] J. Kim and T. Keyes, J. Phys. Chem. B **111**, 2647 (2007).

[33] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **68**, 061111 (2003); *ibidem* **72**, 051929 (2005).

[34] P.E. Leopold, M. Montal, and J.N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).

[35] B. Geiger, A. Bershadsky, R. Pankov, and K.M. Yamada. Nat. Rev. Mol. Cell Biol. **2**, 793–805 (2001).

[36] V. Vogel, Annu. Rev. Bioph. Biom. **35**, 459–488 (2006).

[37] A. Fersht, *Structure and mechanism in protein science*, W. H. Freeman and Company, fifth printing (2003).

[38] J.T. Kellis Jr, K. Nyberg, Dasa Sali and A.R. Fersht, Nature **333**, 784 (1988).

[39] P.L. Privalov and S.J. Gill, Adv. Prot. Chem. **39**, 191 (1988).

[40] C. Levinthal, J. Chem. Phys. **65**, 44 (1968).

[41] C.B. Anfinsen, Science **181**, 223 (1973).

[42] P.E. Leopold, M. Montal and J.N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).

[43] J.N. Onuchic, P.G. Wolynes, Z. Luthey-Schulten and N.D. Socci, Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).

[44] P.G. Wolynes, J.N. Onuchic and D. Thirumalai, Science **267**, 1619 (1995).

[45] R.B. Best, D.J. Brockwell, J.L. Toca-Herrera, A.W. Blake, D. Alastair Smith, S.E. Radford and J. Clarke, Anal. Chim. Acta **479**, 87 (2003).

[46] J. Brujić, R. I. Hermans Z., K. A. Walther and J. M. Fernandez, Nature Physics **2**, 282 (2006)

[47] T.S. Grigera, A. Cavagna, I. Giardina, and G. Parisi, Phys. Rev. Lett. **88**, 055502 (2002).

[48] D.A. Evans and D.J. Wales, J. Chem. Phys. **121**, 1080 (2004).

[49] G.M. Torrie and J.P. Valleau, Chem. Phys. Lett. **28** (1974) 578.

[50] A.M. Ferrenberg and R.H. Swendsen, Phys. Rev. Lett. **63** (1989) 1195.

[51] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman and J.M. Rosenberg, J. Comput. Chem. **13** (1992) 1011.

[52] B. Roux, Computer Physics Communications **91** (1995) 275.

[53] M. Carrion-Vazquez *et al.* Proc. Natl. Acad. Sci. USA **96**, 3694 (1999); B. Onoa *et al.*, Science **299**, 1892 (2003).

[54] G. Favrin, A. Irbäck, and F. Sjunnesson, J. Chem. Phys. **114**, 8154-8158 (2001).

[55] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Science **220**, 671-680 (1983).

[56] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, J. Chem. Phys. **21**, 1087-1092 (1953).

[57] M. Carrion-Vazquez *et al.*, Nature Structural Biology **10**, 738 (2003).

[58] A. Irbäck, S. Mitternacht, and S. Mohanty, Proc. Natl. Acad. Sci. USA **102**, 13427 (2005).

[59] L. Li, H.H.-L. Huang, C.L. Badilla, and J.M. Fernandez, J. Mol. Biol. **345**, 817-826 (2005).

[60] D.J. Brockwell *et al.*, Nature Structural Biology **10**, 731 (2003).

[61] A. Imparato and L. Peliti, Eur. Phys. J. B **39**, 357 (2004).

[62] L.D. Landau and E.M. Lifshitz, *Statistical Physics*, Pergamon Press, Oxford, 1990, 3rd ed., Part. 1, Sect. 15;

[63] D. Collin *et al.*, Nature **437**,231 (2005).

[64] Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).

[65] Z. Guo and C.L. Brooks III, Biopolymers **42**, 745-757 (1997).

[66] S. Fowler *et al.*, J. Mol. Biol. **322**, 841 (2002).

[67] E. Paci and M. Karplus, Proc. Natl. Acad. Sci. USA **97**, 6521 (2000).

[68] D.K. West, P.D. Olmsted, and E. Paci, J. Chem. Phys. **124**, 154909 (2006).

[69] T. Veitshans, D. Klimov, and D. Thirumalai, Folding & Design **2**, 1 (1997).

[70] J. Kim, J.E. Straub, and T. Keyes, Phys. Rev. Lett. **97**, 050601 (2006).

[71] D.J. Lacks, Biophys. J. **88**, 3494 (2005).

[72] F.-Y. Li, J.-M. Yuan, and C.-Y. Mou, Phys. Rev. E **63**, 021905 (2001).

[73] A. Blondel and M. Karplus, J. Computational Chemistry **17**, 1132-1141 (1996).

[74] A. Rampioni, *Caratterizzazione del panorama energetico di piccoli peptidi al variare della loro lunghezza*, PhD Thesis (Firenze, 2005).

[75] W. Kabsch, Acta Cryst. **A32**, 922-923 (1976).

[76] W.H. Press *et al.*, Numerical Recipes (Cambridge University Press, 1994, New York).

[77] A. Imparato and L. Peliti, J. Stat. Mech., P03005 (2006).

[78] Braun, A. Hanke, and U. Seifert, Phys. Rev. Lett **93**, 158105 (2004).

[79] A. Torcini, R. Livi, and A. Politi, J. Biol. Phys. **27**, 181 (2001).

[80] P-G De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, 1979).

[81] D.K. West, P.D. Olmsted, and E. Paci, Phys. Rev. E **74**, 061912 (2006).

[82] A. Imparato, A. Pelizzola, and M. Zamparo, Phys. Rev. Lett. **98**, 148102 (2007).

[83] J.S. Langer, Ann. Phys. **54**, 258 (1969).

[84] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).

[85] R. Zwanzig, *Nonequilibrium statistical mechanics* (Oxford University Press, 2001).

[86] E. Ruoslahti and M.D. Pierschbacher, Science **238**, 491–497 (1987).

[87] S.-I. Aota, M. Nomizu, and K.M. Yamada, J. Biol. Chem. **269**, 24756–24761 (2004).

[88] A. Krammer, H. Lu, B. Isralewitz, K. Schulten, and V. Vogel, Proc. Natl. Acad. Sci. USA **96**, 1351–1356 (1999).

[89] T. Ohashi, D.P. Kiehart, and H.P. Erickson, Proc. Natl. Acad. Sci. USA **96**, 2153–2158 (1999).

[90] N.I. Abu-Lail, T. Ohashi, R.L. Clark, H.P. Erickson, S. Zauscher, Matrix Biol. **25**, 175–184 (2006).

[91] G. Baneyx, L. Baugh, and V. Vogel, Proc. Natl. Acad. Sci. USA **99**, 5139–5143 (2002).

[92] M.L. Smith, D. Gourdon, W.C. Little, K.E. Kubow, R. Andresen Eguiluz, S. Luna-Morris, and V. Vogel, PLoS Biol. **5**, e268 (2007).

[93] K.W. Plaxco, C. Spitzfaden, I.D. Campbell, and C.M. Dobson, J. Mol. Biol. **270**, 763–770 (1997).

[94] A.F. Oberhauser, C. Badilla-Fernandez, M. Carrion-Vazquez, and J.M. Fernandez, J. Mol. Biol. **319**, 433–447 (2002).

[95] Ng, S.P., and J. Clarke, J. Mol. Biol. **371**, 851–854 (2007).

[96] E. Cota and J. Clarke, Protein Sci. **9**, 112–120 (2000).

[97] E. Paci and M. Karplus, J. Mol. Biol. **288**, 441–459 (1999).

[98] D.K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. USA **97**, 7254–7259 (2000).

[99] M. Gao, D. Craig, V. Vogel, and K. Schulten, J. Mol. Biol. **323**, 939–950 (2002).

[100] D. Craig, M. Gao, K. Schulten, and V. Vogel, Structure **12**, 21–30 (2004).

[101] J.I. Sułkowska and M. Cieplak, J. Phys.: Condens. Matter **19**, 283201(2007).

[102] M.S. Li, Biophys. J. **93**, 2644–2654 (2007).

[103] H. Lu and K. Schulten, Biophys. J. **79**, 51–65 (2000).

[104] C. Jarzynski, Phys. Rev. Lett. **78**, 2690–2693 (1997).

[105] G.E. Crooks, Phys. Rev. E **60**, 2721–2726 (1999).

[106] G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. USA **98**,3658–3661 (2001).

[107] A. Imparato and L. Peliti, J. Stat. Mech. P03005 (2006).

[108] E. Evans and K. Ritchie, Biophys. J. **72**, 1541–1555 (1997).

[109] J.F. Marko and E.D. Siggia, Macromolecules **28**, 8759–8770 (1995).

[110] P.-C. Li and D.E. Makarov, J. Chem. Phys. **119**, 9260–9268 (2003).

[111] C. Hyeon and D. Thirumalai, Biophys. J. **90**, 3410–3427 (2006).

[112] O.K. Dudko, J. Mathé, A. Szabo, A. Meller, and G. Hummer, Biophys. J. **92**, 4186–4195 (2007).

[113] G.I. Bell, Science **200**, 618–627 (1978).

[114] A. Imparato and A. Pelizzola, Phys. Rev. Lett. **100**, 158104 (2008).

[115] H. Li, W.A. Linke, A.F. Oberhauser, M. Carrion-Vazquez, J.G. Kerkvliet, H. Lu, P.E. Marszalek, and J.M. Fernandez, Nature **418**, 998–1002 (2002).

[116] H.P. Erickson, Proc. Natl. Acad. Sci. USA **91**, 10114–10118 (1994).

[117] R. Chelli, S. Marsili, and P. Procacci, Phys. Rev. E **77**, 031104 (2008).

[118] D.D.L. Minh and A.B. Adib, Phys. Rev. Lett. **100**, 180602 (2008).

[119] G. Biroli and R. Monasson, Europhys. lett. **50**, 155 (2000).

Ringraziamenti

Bisogna cominciare subito con un grazie al dott. Alessandro Torcini che mi ha seguito in questi tre anni e che ha anche alternato le funzioni di tutore con quelle di allenatore...e quindi gli dico che ancora i Renai ci mancano!

Ringrazio quindi il prof. Alberto Imparato, il prof. Anders Irbäck ed il dott. Simon Mitternacht con i quali, insieme ad Alessandro Torcini, è stato svolto questo lavoro di ricerca. Ringrazio anche il prof. Roberto Livi, il dott. Antonio Politi ed il dott. Stefano Lepri per gli utili suggerimenti.

Passando ai compagni di stanza passati e presenti, ringrazio Luca Delfini e proseguendo l'antico discorso gli dico che solo in un caso...io la promessa l'avrei mantenuta! Grazie anche agli altri compagni della stanza 126: la Raffaella anche se è scappata dopo pochi mesi..., Thomas che ha lasciato molti ricordi di sè...Valerio l'Ispanico, Massimo e l'astice; la Simona grazie alla quale nella stanza 126 non mancano mai arance e pinguini...

Grazie a Stefano L., al Leprisky e alla URL.

Ringrazio le serate al Terzolle di Gianni che poi divenuto Duca Conte ci ha ospitato insieme alla Contessa al maniero.

Grazie ai miei compagni di dottorato, Andrea Guazzini, Francesca di Patti, Elisa Calistri e Fausto Baroncini e le tante, ma per carità mai abbastanza, serate.

Ringrazio le sorelle Cacciari (almeno due...), Marina, nu' poco rispustera, Carlos il mangiatore di arance... Il Professore, il professore, il professore sovrannumerario, Garlo, il nostro amico Andrea Barducci e gli amici di Parma Elena e Fernet Maria.

Grazie a Simpan e alla ”green room” ed alle ”conversazioni sulla panchina” con Caroline.

Grazie a Gianpiero Puccioni, Francesca Sbrana, Lapo Casetti, Lorenzo Bongini, Francesco Ginelli, Stefano Ruffo, Franco Bagnoli, Pietro Poggi, Francesco e Marzenka e Antonellnka, Ruediger, Alessandro Vindigni e l'amico Giorgione.

Non si può infine non ricordare Giovannone, il Sig. Verdirame Augusto (da Brescia) e almeno i' Nedo.