Politecnico di Torino

DIPARTIMENTO DI FISICA

Corso di Dottorato di Ricerca in Fisica - XXIV ciclo

Tesi di Dottorato

Mechanical unfolding and confinement of proteins investigated through an Ising-like model

Candidato
Michele Caraglio

Relatore
Prof. Alessandro Pelizzola

Anno Accademico 2011–2012

Contents

Introduction

Proteins are heteropolymers made up of an ordered sequence of amino acids that forms the so called primary structure. In nature there are twenty different amino acids and their sequence in a protein is specified by the sequence of a gene, which is encoded in the genetic code. The peculiar feature of proteins is that, under physiological conditions, they fold spontaneously in a unique, specific and compact three–dimensional structure, called native configuration. It is its three–dimensional shape that determines the macroscopic properties of a protein and makes it able to perform its biological functions. 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].

The big challenge in protein science is the prediction of the folded configuration starting from the primary structure. This issue, also known as the protein folding problem, has been the subject of intense investigation for decades, involving biologists, chemists, mathematicians, physicists and computer scientists, but unfortunately, a complete understanding of the mechanisms involved in the folding process is still lacking. The importance of finding a solution of the protein folding problem can be easily understood by considering that many diseases, such as Alzheimer's and Parkinson's disease, are associated with the accumulation of misfolded proteins and their aggregation in structures called amyloid fibrils [2, 3]. To solve protein folding problem will thus be a crucial step towards the achievement of a successful treatments of these diseases. Moreover, once the solution of the inverse protein folding problem (i.e. the problem of finding the primary structure of a protein given its final three–dimensional shape) is known, one can easily attempt at designing proteins which will have applications in medicine and bioengineering. Furthermore, a protein is a system which is in a nearly–zero–entropy equilibrium state (the native state) for a wide interval of temperatures ranging, from about 0°to 60°C. Such equilibrium state has essentially no symmetries. Notwithstanding the complicated and heterogeneous interactions which contribute to the formation of the native state, proteins do have neither slow dynamics, nor the large number of competing low–energy states and kinetic traps typical of "frustrated" systems. Thus protein folding problem is also extremely intriguing by itself from the physical point of view.

The ability of proteins to fold spontaneously and in a time scale ranging from some milliseconds to few minutes, immediately raised a fundamental question that nowadays is known as the Levinthal paradox [4, 5]. It reads as follows. A protein

has billions of possible conformations. For a 100–monomers chain, in the simplest case in which for each residue only two states are possible, they are $2^{100} \sim 10^{30}$. If proteins had to randomly sample all these configurations, even at the fast test rate of one sampling each picosecond, the chain would need $\sim 2^{100}$ picoseconds, or $\sim 10^{10}$ years, to find the folded state, which is clearly not the case. To solve this puzzle, Levinthal himself proposed that there exists a specific folding pathway, and the native fold is simply the end of this pathway.

The current perspective is that the folding proceeds via a nucleation and growth mechanism [6] in which only the folding nucleus needs to form its native contacts by "chance" and then the rest of folding process is downhill with the polypeptide chain that attains fast its native structure by accretion. The folding nucleus is generally defined as the folded part of a particular ensemble of configurations, called the transition state ensemble. From the physical point of view this is the ensemble of states that corresponds to the saddle point of the free energy between the folded and the unfolded state [7]. However, for some proteins, the free energy landscape shows, besides the minima corresponding to the native and unfolded states, other local minima, and in this case the folding pathway may also go through metastable states, called intermediates [8,9]. Finally it is worth noting that proteins exist which have multiple folding pathways [8,9].

In the last two decades, the study of protein folding received a boost thanks to new experimental techniques based on atomic force microscopy [10] and on laser optical tweezers [11]. These techniques allowed the microscopic manipulation of a single biomolecule with the advantage that it is thus possible to separate out the fluctuations of a single molecule from the ensemble average behavior observed in traditional biochemical–biophysical experiments. Manipulation experiments on single biological molecules may thus greatly increase the knowledge of the structural properties of such molecules. Exerting a mechanical force, it is possible to induce unfolding of the molecule, to measure the binding forces responsible for the stability of biomolecules and to explore the unfolding pathways and the possible intermediate states. Specific functions of proteins can often be exerted thanks to these structural properties. Unfortunately experiments have shown that mechanical unfolding may be different from the thermal or chemical one [12]. Nevertheless pulling experiments remain a powerful tool to extract information on the internal structure of proteins, as well as on unfolding and refolding pathways.

To this day, many physical model have been proposed with the goal to find the basic features of the protein folding thermodynamics and kinetics. Some of them, called phenomenological models, generally aim to solve the Levinthal's paradox [6, 13,14,15]. Others try to simulate the folding of proteins at a greater level of details. Among this second class of models there are the all–atom models [16, 17, 18] and the less detailed but more practical coarse–grained models [19, 20, 21, 22, 23]. This Thesis deals with some of the possible applications of a particular coarse–grained model known as the Wako–Saitô–Muñoz–Eaton (WSME) model.

WSME model was introduced for the first time in 1978 by Wako and Saitô [24,25] and then forgotten until it was independently reconsidered by Muñoz and Eaton [26, 27,28] as a simple and efficient theoretical tool to interpret their experimental data. It is a one–dimensional model, with long–range and many–body interactions, where a binary variable is associated with each amino acid, denoting its ordered or disor-

dered conformation. For this reason we will also refer to the WSME model as an Ising–like model. Two residues can interact only if they are in contact in the native state and all the residues between them are ordered. The non–native interactions are neglected in the spirit of Gō–type models [19]. Moreover, in the original version of the model, an entropic cost is associated with each ordered residue. The WSME model is particularly alluring because it has remarkable mathematical properties which make it possible to obtain an exact solutions in equilibrium conditions [29]. Later on, Imparato *et al.* [30], by introducing a protein–length dependent potential in the Hamiltonian, proposed a modified version of the model able to treat mechanical unfolding of proteins. In the case in which the protein is pulled by a constant force, the equilibrium model turned out to be exactly solvable once more. Non equilibrium behavior can be instead studied by means of Monte Carlo simulations.

In this Thesis we will show that this generalized WSME model is a good one to describe mechanical unfolding of proteins. On one side, its simplicity permits to probe forces and speed ranges close to *in vivo* and experimental conditions, which is often not possible for other more detailed models because they require higher computational efforts. On the other hand, the model generally obtains results in accordance with experiments and other simulations. Indeed in the case of two well studied molecules such as a fibronectin domain and the green fluorescent protein, it is able to predict correctly intermediate states, folding pathways and binding forces.

This version of the model, when no force is applied to pull the molecule, can be further generalized to investigate another interesting phenomenon which has great relevance, particularly when one takes into account that folding process occurs in the cellular environment, namely confinement of proteins. *In vivo*, such a phenomenon occurs, for example, in the exit–tunnel of ribosomes or in the chaperonin cavity. There is experimental evidence [31,32] that confinement can alter both thermodynamics and kinetics of folding by enhancing folding temperature and folding rate. Furthermore, confinement is interesting also because molecular crowding, another phenomenon that occurs in the interior of the cell, upon certain conditions, has effects similar to those of confinement [33]. Understanding better the role of confinement is thus desirable to get an improved knowledge of how protein folding works in, and is modified by, cellular environment, in particular it should shed more light on the functioning of chaperones and ribosomes and, on the other hand, on the effects of crowding. We will show that WSME model can be a useful tool also to reproduce confinement effects.

Outline of the thesis

The plan of this Thesis is the following. The first chapter is devoted to a brief introduction about the structure of proteins, necessary in order to better understand contents of the subsequent chapters. After a brief description of the amino acids, the notions of primary, secondary and tertiary structure will be discussed.

In the second chapter we will introduce some accepted facts concerning the protein folding process. We will rapidly consider the forces that fold a protein and then we will review some important notions useful to characterize protein phases and thermodynamics and kinetics of protein folding. These concepts are those of free energy landscape, folding funnel, transition state ensemble, folding rate, folding

pathways and intermediate states.

The third chapter is devoted to describe mechanical unfolding of proteins and to review the kinetic theory developed to describe it. A brief description of the experimental techniques used in this field will be also given.

The fourth chapter introduces the WSME model and its generalization for mechanical unfolding. It will be shown how to find the exact solution of the model, i.e. how to find an algorithm with a polynomial complexity in the number of amino acids that allows to compute quantities of physical interest related to the protein.

The fifth chapter is divided in two parts, the first and second one dealing respectively with the mechanical unfolding of the tenth type III domain of Fibronectin and of the Green Fluorescent Protein. We will investigate the mechanical unfolding pathways of the two proteins, using both constant force and constant velocity protocols and trying to predict the unfolding pathways and to estimate the kinetic parameters of the energy landscape. In the Green Fluorescent Protein case we will pull the molecule not only from its ends but also along different directions, trying to find the magnitude and ranking of the unfolding forces. As it will be shown, such a procedure could pave the way to the use of the Green Fluorescent Protein as a force sensor.

Finally, the sixth chapter deals with confinement of proteins. We will show how it is possible to modify the algorithm, which constitutes the iterative solution of the equilibrium thermodynamics of WSME model, in order to handle with confinement of the polypeptide chain between two inert hard walls. Also in this case the equilibrium solution of the model is exact. Exploiting this fact, we will study the equilibrium thermodynamics upon confinement of three real and three ideal protein structures. Using Monte Carlo simulations we will consider also the changes induced in the folding kinetics.

Conclusions will be drawn at the end of the Thesis.

Chapter 1

Protein structure

Proteins are biological complex systems whose study is currently one of the most intriguing and challenging topics in physics, chemistry and biology. From physico-chemical point of view proteins are biopolymer made up of monomer called *amino acids* which link to each other giving rise to a chain called the *polypeptide chain* which folds into a compact globular structure with a well defined three dimensional state whose shape is strictly connected to the biological function of the protein [34]. Amino acids that occur in proteins are of 20 different standard types and a protein typically consists of between 100 and 1000 amino acids.

The word protein comes from the Greek word "proteios" that means primary. Proteins were first described by the dutch chemist Gerardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838. However, the central role of proteins in living organisms was not fully appreciated until 1926, when James B. Sumner (nobel prize in chemistry in 1946) showed that the enzyme urease is a protein [35]. Early nutritional scientists believed that proteins were the most important nutrient for maintaining the structure of the body, because it was generally believed that "flesh makes flesh". This old but still quite widespread belief is true only to a certain extent in the sense animals cannot synthesize all the amino acids they need and must obtain them from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism to build other proteins which can also have different roles than forming muscles and other tissues. In fact, because of their ability to bind other molecules specifically and tightly, proteins are among the most versatile macromolecules which contribute to the functioning of living beings.

Some proteins (*structural proteins*) polymerize to form stiff and long fibers which confer stiffness and rigidity to otherwise fluid biological components and thus play an important role in the architecture of cells and extracellular matrix. There are proteins which are capable of generating mechanical forces from chemical energy (*motor proteins*), such as myosin and kinesin, and there are proteins which are in charge of binding particular small biomolecules and to transport them into the cell or from the cell to another location (*transport proteins*), like hemoglobin which transports oxygen from the lungs to other tissues. Furthermore, *antibodies* are protein whose main function is to bind antigens, or foreign substances in the body, and target them for destruction. Many *hormones*, such as insulin, are proteins in charge of bringing the message released by a cell in one part of the body to

cells in other parts of the organism while, on the surface of the receiving cell there are proteins (*receptor proteins*) which have the role to harvest these messages and to induce a biochemical response in the cell. Finally, most *enzymes* are proteins which bind to a specific substrate and catalyze a given chemical reaction, i.e. they lower the connected rate–limiting free energy of activation. Enzymes carry out most of the reactions involved in metabolism, as well as DNA replication and DNA repair, and the rate acceleration conferred by enzymatic catalysis is often enormous. For example, the uncatalyzed decarboxylation of orotidine 5'–monophosphate has a half–time life of 78 million years. However, when the enzyme orotidine 5'–phosphate decarboxylase is added, the same process takes just 25 milliseconds [36].

To supply with all these functions and since protein action is very specific, thousands of different proteins exist. Nowadays, protein structures are typically obtained by X–ray crystallography or NMR spectroscopy and released into a public repository, the *Protein Data Bank* [37] (pdb), which contains the atomic coordinate entries of more than 75 thousand proteins.

1.1 The amino acids

The chemical structure of an amino acid is formed by an *amino group* NH_2 and a *carboxyl group* $COOH$ which are bound to a central carbon atom, denoted by C_α, to which are also bound an hydrogen atom and a *side chain*, usually called R, which makes the difference between the various amino acids.

The simplest side chain is just an hydrogen atom, corresponding to the simplest amino acid, *glycine*. The other amino acids can exist in either of two optical isomers, called left or right handed amino acids, which are mirror images of each other (see figure 1.1). In nature left–handed amino acids represents the vast majority while right-handed amino acids have been found only in some proteins pruducted by exotic sea–dwelling organism.

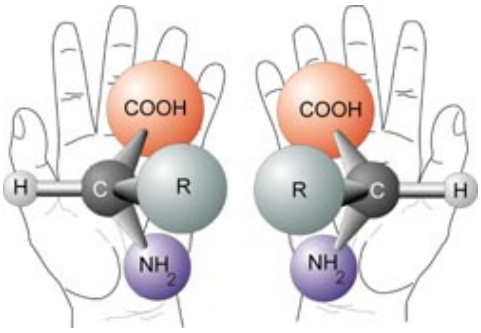


Figure 1.1: Chirality of a generic amino acid. Figure created by NASA.

With the exception of *methionine* and *cysteine*, which contain also a sulfur atom, the amino acids are made of just carbon, hydrogen, oxygen, and nitrogen atoms. Thanks to its sulfur atom, cysteine plays an important structural role in many proteins since it may form a covalent disulfide bond with another cysteine residue thus giving stability to the structure of the protein.

The 20 amino acids can be divided into several groups based on the hydrophilic and hydrophobic properties of their side chains. A hydrophile, from the Greek "hydros" (water) and "philia" (love), is a molecule that is attracted to, and tends to be dissolved by water because of the dipole–dipole (polar molecules) or charge–dipole (charged molecules) interactions with the polar water molecules. On the contrary, a hydrophobe, from the Greek "phobia" (fear), is a non polar molecule which cannot form hydrogen bonds with water. When hydrophobic molecules are dissolved in wa-

ter or in another polar solvent they tend to cluster in order to minimize their surface in contact with water. As it will be discussed further later, these properties are important for protein structure and protein–protein interactions and hydrophobicity is believed to be the main force that drives the protein into a collapsed globular state with the hydrophobic amino acids mostly inside and the hydrophilic ones exposed on the surface.

The polar amino acids are *serine*, *threonine*, *asparagine* and *glutamine*. The first two are small and polar due to an OH group in their side chain. The charged amino acids are subdivided into the positively charged *arginine*, *lysine* and *histidine* and the negatively charged *aspartate* and *glutamate*. Charged amino acids of opposite sign can bind to each other through salt bridges, that in some cases can be essential for maintaining the stability of a protein. The hydrophobic amino acids are *glycine*, *alanine*, *valine*, *leucine*, *isoleucine* and *phenylalanine* while *tyrosine*, *tryptophan*, *proline*, *cysteine* and *methionine* can exhibit both attitudes.

Glycine, given its very small side chain and thus minimal steric constraints, confers particular flexibility to the polypeptide chain. On the contrary, phenylalanine is special because its side chain is a benzene ring and this large planar structure enforces considerable steric constraints on the structure inside the protein, where hydrophobic amino acids are mostly found. Tyrosine and tryptophan have also large planar side chains and tend to be exposed on the surface of a protein, being often associated to its function. Finally proline is the only residue whose side chain is also covalently linked to the nitrogen of the amino group and this particular cyclic structure allows for specific sharp turns of the chain.

1.2 The primary structure

The *primary structure* refers to amino acid sequence of the polypeptide chain. The various amino acids are held together by covalent bonds which form between them during the process of *protein biosynthesis*. The amino acids of the chain are also called *residues* while the covalent bond is called the *peptide bond*.

The primary structure of a protein is determined by a gene corresponding to the protein: since nucleic acids have four *nucleotide* bases (*adenine*, *cytosine*, *guanine* and *thymine* for DNA or *uracil* for RNA) while there are 20 different amino acids, from simple combinatorial considerations it follows that each amino acid must be associated to a sequence of at least three bases. This is actually what happen in nature where each amino acid is associated to a given triplet of nucleotides, called *codon*. Some codons may refer to the same amino acid. Protein synthesis occurs in two major steps: the *transcription* and the *translation*. During the former process an enzyme, called *RNA–polymerase*, copies a given sequence of DNA bases into a RNA fragment called the messenger RNA or *mRNA*. In prokaryotic cells the product of transcription is directly a mature mRNA fragment while in eukaryotic cells the first product needs some post–transcriptional modifications to become mature. Then, in eukaryotic cells, the mRNA translocates into the cytoplasm where *ribosomes* make the translation by using the mRNA as a template to build a particular protein. The ribosome moves along the mRNA "reading" its sequence three nucleotides at a time and matching the codons to the corresponding amino acids thus forming the polypeptide chain. The transport of the various amino acids into the ribosome is

due to short RNA fragments (transfer RNA or *tRNA*) which act also as adaptors between the codons and their specific amino acids.

The peptide bond forms through the condensation of the carboxyl group of one amino acid and the amino group of the next with the release of a water molecule. Thus a polypeptide chain has an amino terminus (N–terminus) and a carboxyl terminus (C–terminus). The length between the C and N atoms that form the peptide bond is about 1.33 Å while the length for each residue in a chain is about 3.8 Å. Peptide bonds can be broken spontaneously in water but this process is extremely slow, due to their high strength. Protein chains are thus chemically and biologically stable unless they are deliberately depolymerized. In the following we will refer to the protein *backbone* as the main chain formed by the repetition of the N, C_α and C atoms of the various residues.

Since the peptide bond is partially double covalent, the $C_{\alpha,i}$, C_i and O_i atoms of the i–th residue and the N_{i+1}, H_{i+1} and $C_{\alpha,i+1}$ atoms of the following have a strong tendency to lie in the same plane. This coplanarity allows for two different conformations: that in which the $C_{\alpha,i}$ and $C_{\alpha,i+1}$ atoms lie at different sides of the line containing the C_i–N_{i+1} bond and the other in which they lie at the same side of that line. Because of steric hindrance, usually the residues in the chain adopt the former configuration, called *trans*, with the only exception of proline, which gives rise also to the latter, named *cis*. This two configurations are by definition associated to two possible values of ω angle defined in figure 1.2 with $\omega = 0$ for the cis isomer and $\omega = \pi$ for the trans isomer.

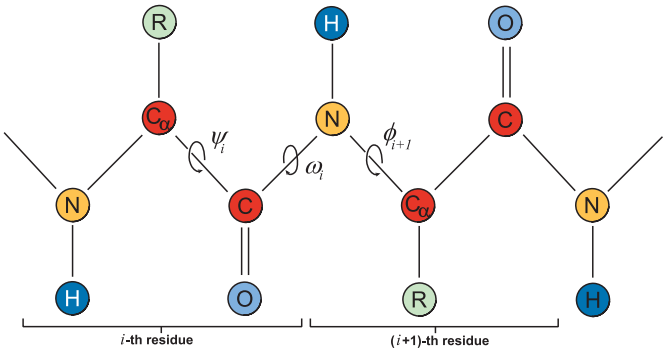


Figure 1.2: Peptide bond between the i–th and $(i+1)$–th residues with angles ψ_i, ω_i and ϕ_{i+1} defined.

Notwithstanding these coplanarity properties, the polypeptide chain is flexible because rotations are still possible around the $C_{\alpha,i}$–C_i (ψ_i angle) and the N_i–$C_{\alpha,i}$ (ϕ_i angle) bonds, conferring two degrees of freedom to each peptide unit. The ϕ_i, ψ_i angles are called *dihedral angles* and, by convention, they take values in between $-\pi$ and π with the zero for ϕ_i corresponding to a configuration with C_i atom staying in the same plane of C_{i-1}, N_i and $C_{\alpha,i}$ atoms and the zero for ψ_i corresponding to a configuration with N_{i+1} atom lying in the same plane of N_i, $C_{\alpha,i}$ and C_i atoms. Because of steric hindrance only certain combinations of ϕ, ψ angles can occur in a real protein [38, 39]. The phenomenological distribution of dihedral angles in the backbone conformation of a real protein is displayed by means of the *Ramachandran*

map. The allowed area for glycine is considerably larger than for other residues, due to the smallest side chain. In contrast, the Ramachandran map shows only a very limited number of possible combinations for proline, whose presence increases remarkably the rigidity of the peptide bond, while the other amino acids instead show a quite similar behavior.

1.3 The secondary structure

The *secondary structure* refers to the regularly repeating local structures stabilized by hydrogen bonds which appear in every protein. In a polypeptide chain surrounded by a polar solvent, each peptide unit behaves like a dipole moment (with module of about 3.5 debye) because the oxygen linked to the carbon and the hydrogen linked to the nitrogen can form hydrogen bonds, being hydrogen bond acceptors and donors respectively. These hydrogen bonds would be satisfied with the solvent molecules but, in native conditions, the protein is forced to assume a folded compact state by the tendency to cluster of the hydrophobic side groups. Thus, in order to have energetically favorable conditions, the different residues along the chain hydrogen bond to each other giving rise to regular local conformations of the backbone, which constitute the elements of the secondary structure. Because of the steric hindrance and the limited directionality of hydrogen bonds, the possible structures are severely constrained. The most common secondary structures are the *α–helices*, the *β–sheets* and the *tight–turns*. Other helices, such as the *3_{10}–helix* and *π–helix*, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely observed in natural proteins except at the ends of α–helices due to unfavorable backbone packing in the center of the helix.

The α–helix structure is a coiled or spiral, normally right–handed, conformation which looks like a spring. In the right handed α–helix the amide hydrogen of an amino acid forms a hydrogen bond with the carbonyl oxygen of the amino acid four residues earlier and this bonding condition is repeated for a stretch of the protein backbone which is usually 10–residues long but some proteins exhibit helices to over forty residues. The pitch and the radius of the α–helix are 5.4 Å and 2.3 Å, corresponding to 3.6 residues per turn. The 3_{10}–helix and the π–helix are similar structures where, respectively, the hydrogen bond is established between amino acids separated by one residue less and one more than in the α–helix. Dihedral angles ϕ and ψ of a perfect α–helix take respectively the values $-57.8°$ and $-47.0°$. This kind of helix is tightly packed and, since there is almost no free space within it, the amino acid side chains stick out towards the outside. Short pieces of left–handed helix sometimes occur with a large content of glycine (which is achiral), but are unfavorable for the other amino acids.

A β–sheet is formed by two or more stretches of amino acids with the backbone chain almost fully extended (the β–strands), which are connected by several hydrogen bonds. Adjacent β–strands can form hydrogen bonds in antiparallel, parallel, or mixed arrangements. As shown in figure 1.3, in an antiparallel arrangement, the successive β–strands alternate directions so that the N–terminus of one strand is adjacent to the C–terminus of the next. This is the arrangement that produces the strongest inter–strand stability because it allows the inter–strand hydrogen bonds between carbonyls and amines to be planar, which is their preferred

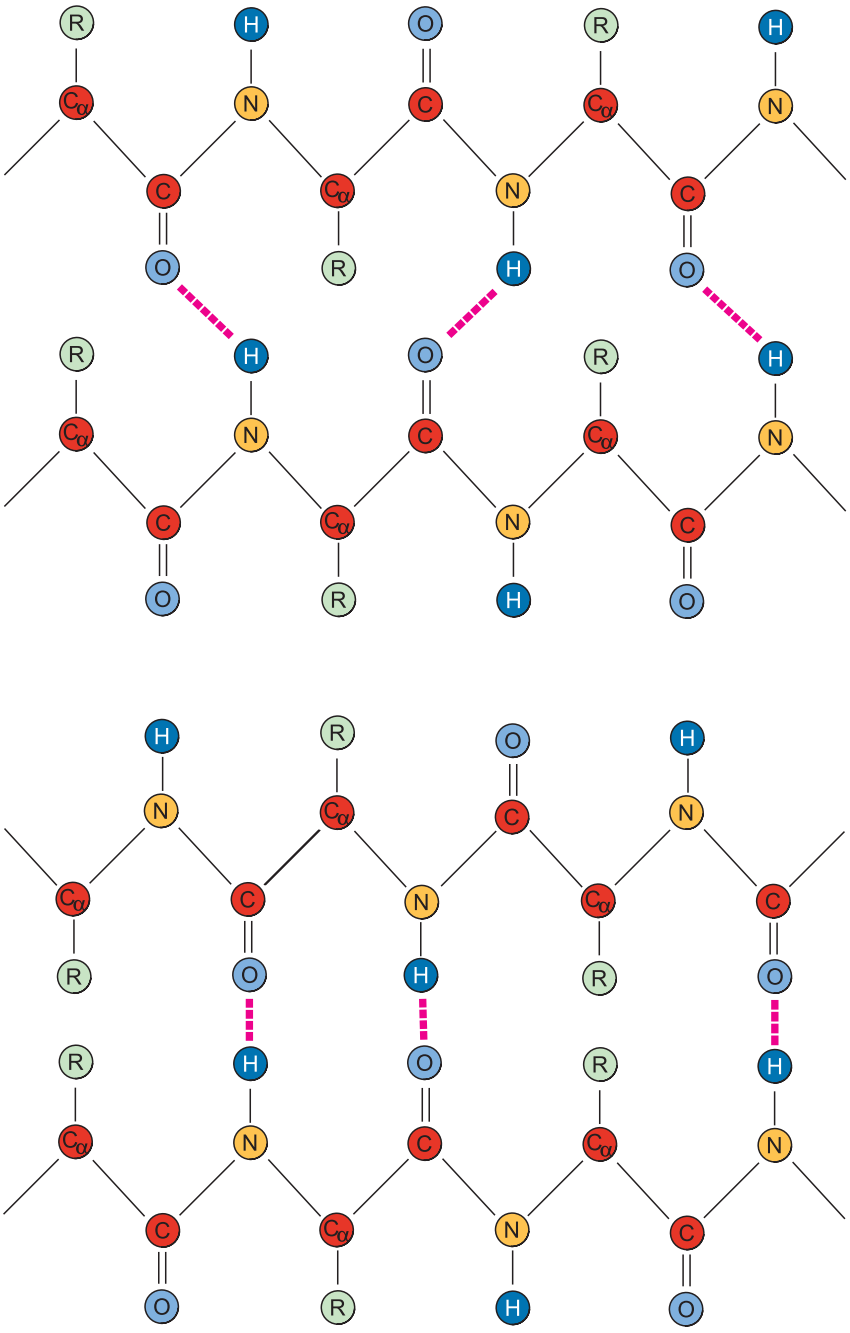


Figure 1.3: Parallel (top panel) and antiparallel (bottom panel) β–sheets. Purple dashed lines represent hydrogen bonds.

orientation. In this configuration ϕ and ψ angles are about $-140°$ and $135°$. In a parallel arrangement, the successive β–strands are oriented in the same direction and this orientation may be slightly less stable because it introduces nonplanarity in the inter–strand hydrogen bonding pattern. ϕ and ψ angles are about $-120°$ and $115°$ in parallel case but large deviations are observed in real proteins. Unlike helical segments, all peptide group hydrogen bond donors and acceptors are satisfied not within but between strands, thus individual β–strands do not have an independent existence.

In general turns are elements of secondary structure of two to five residues where the polypeptide chain reverses its overall direction. If the stretch of residues that allowed the chain to reverse its direction contains more than five residues and does not have a fixed internal hydrogen bonding, one refers to it as an ω–loop. Tight–turns and loops are essential for allowing the polypeptide chain to fold back upon itself to form tertiary interactions. Such interactions are generally long–range and result in compaction of the protein into a globular form. The turn regions are thus generally located on the outside of the globular structure, with helices and/or sheets forming its core. Turns on the surfaces of proteins have a wide range of dynamics, from quite mobile in cases where they form few interactions with the underlying protein surface to quite fixed in the opposite cases.

Amino acids vary in their ability to form the various secondary structure elements: amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine. On the contrary tryptophan, tyrosine, phenylalanine, isoleucine, valine, and threonine prefer to be included in a β–sheet. Finally, as already mentioned, proline is likely to be found in and to be responsible for the tight–turns. Furthermore, proline and glycine are sometimes known as "helix breakers" because they disrupt the regularity of the α–helical backbone conformation. However, these preferences are not strong enough to produce a reliable method to predict the secondary structure of a protein from the amino acids sequence alone.

1.4 The tertiary structure

The *tertiary structure* of a protein is its three–dimensional structure defined by the atomic coordinates. Driven by the hydrophobic force, the polypeptide chain folds upon itself and α–helices and β–sheets pack together to form the tertiary structure of the protein. This structure is then stabilized by salt bridges between polar amino acids and, in some cases, by strong disulfide bonds between cysteine residues. According to the content of secondary structure elements it is possible to distinguish between α–proteins, which are formed only by α–helices like myoglobin, β–proteins, formed only by β–sheets such as the tenth type III module of fibronectin, and α/β–proteins which contains both kinds of secondary structure elements as for example protein G. The tertiary structure of more complex proteins shows often several almost–independent domains, with their own secondary structure, that interact to each other through Van der Waals forces.

Furthermore some proteins show a higher degree of structural complexity which is called the *quaternary structure*. These proteins are formed by more polypeptide chains which fold separately into their own tertiary structure and the quaternary

structure refers to the arrangement of independent tertiary structural units in a multi-subunit complex stabilized through surface interactions, such as formation of the hemoglobin tetramer from myoglobin–like monomers. The subunits that associate may be identical or not and their organization may or may not be symmetric.

Looking at the tertiary structure of a protein one can notice that there is an important difference between the α–helices and the β–sheets: the former are local structure while the latter are not because adjacent strands of a sheet can come from sequentially distant segments of the chain. Thus, given the sequence of the amino acids, it is easier to predict the formation of α–helices than the formation of β–sheets. However, in purely antiparallel sheets, stretches that are sequentially next to each other in the primary structure often form contiguous strands. An example is constituted by the β–hairpin motif in which two antiparallel strands are linked by a turn.

Protein secondary and tertiary structures are not independent of each other, but rather interdependent. It seems likely that this interdependence is the molecular origin of the extraordinary cooperativity of protein structural stability, which is reflected in the observation that, in many proteins, secondary and tertiary structures are lost concomitantly and in an all–or–none manner upon changes in environment that disfavour the folded state, such as higher temperature or solvent additives.

Chapter 2

Protein folding

The physical process by which a polypeptide chain reaches its characteristic three–dimensional structure is named *protein folding*.

In the year 1952 Frederick Sanger determined the sequence of insulin thus proving beyond any doubt that, given a protein, it has a unique sequence of amino acids [40]. From then on, many efforts have been done to understand the principles that lead protein folding but a clear milestone has been set by Christian B. Anfinsen which in 1961 showed that the protein ribonuclease can be reversibly denatured and renatured *in vitro* [41]. This discovery leads Anfinsen to postulate the *thermodynamic hypothesis* of protein folding (also known as the Anfinsen's dogma) which states that the protein amino acid sequence, by itself, contains all the information necessary to define its three–dimensional shape. This amounts to say that, at the environmental conditions at which folding occurs, the native structure is a unique, stable and kinetically accessible minimum of the free energy. Uniqueness requires that the free energy minimum corresponding to the native state is a global minimun and that the sequence does not have any other configuration with a comparable free energy. Stability means that small changes in the surrounding environment cannot give rise to changes in the minimum configuration. This can be pictured as a free energy surface that looks rather like a funnel with the native state in the bottom of it and with the free energy surface around the native state rather steep and high in order to provide stability. Finally, kinetical accessibility refers to the fact that the path in the free energy surface from the unfolded to the folded state must be reasonably smooth or, in other words, that the folding of the chain must not involve highly complex changes in the shape.

Several experiments on thousands of proteins substantiated Anfinsen's hypothesis, thus confirming that protein folding is a spontaneous process and that it does not require any specific cellular machinery or other mysterious biological factors.

Actually there are enzymes that promote disulfide interchange and proline cis–trans isomerization. Furthermore there exist proteins, called chaperones, whose function is to help other proteins to fold and to prevent their aggregation with other proteins. For example GroEL/GroES chaperonin system forms a nano–cavity in which the assisted protein can fold in isolation [42]. But the presence of this complications does not change the previously depicted general picture because all the information necessary to folding remains indeed contained in the amino acids sequence and this makes folding of proteins particularly alluring to study.

Given a particular chain of peptides, the problem of predicting *a priori* the final three–dimensional native structure and how this is reached, is called the *protein folding problem*. Though a great amount of work has been made, the protein folding process remains still not completely understood and in 2005 *Science* named the protein folding problem one of the 125 biggest unsolved problems in science [43]. It is also possible to define an *inverse protein folding problem* as the problem of finding a particular amino acid sequence which folds into a given three–dimensional shape with a given biological function. This latter problem has a great practical importance due to the fact that its solution would allow to design proteins according to their possible use.

This chapter is devoted to discuss the general features of protein folding thermodynamics and kinetics and to introduce many concepts that will be useful later on in this thesis.

2.1 Phases and forces driving the folding

In other words, Anfinsen's dogma defines protein folding as the reversible transition from a disordered ensemble, called the unfolded (or *denatured*) state, to a uniquely folded structure called the folded (or *native*) state.

Proteins may be extracted from cells and studied *in vitro* under different temperatures, pH, denaturant concentration, etc. [8]. At high temperatures and low pH, the proteins unfold while they refold into their native shape upon restoring the native environment.

In the native state, a protein is an exclusively ordered macroscopic system with each of its atoms occupying a definite position, as in a crystal, but differing in that the position of each of the atoms is unique relative to its neighboring atoms. Therefore, a protein represents a new class of macroscopic systems with an aperiodic order. Actually, in the eighties, it has become clear that the native form of a protein is not simply a single state but rather a collection of states that are structurally very similar but are separated by measurable energy barriers [44, 45]. Due to thermal fluctuations, or driven by mechanical forces many proteins can dynamically move in the set of these conformational substates. Such conformational changes have implications on the attitude of a protein to bind other macromolecules [46]. Yet, in a given macromolecule, only a subset of motions, typically involving large scale vibrations or hinge–like movements, is important for biological function [47]: for example activation or inactivation of enzymes relies on their conformational changes and on structural modifications occurring at specific locations.

Most of the folding studies have focused exclusively on the native structure. However, to understand the thermodynamics of folding, both sides of the folding transition must be considered [48]. The unfolded phase is associated with an ensemble of an enormous number of largely unrelated disordered microstates that rapidly interconvert on a time scale which turns out to be of some picoseconds (thus being much faster than the typical time scale of the folding which is of some microseconds or longer). The completely unfolded state of a protein has usually been thought to be properly described as a random coil. Indeed, in good solvent, the radius of

gyration[1], as determined by X–ray scattering or any one of a variety of physical techniques, usually indicates a highly expanded polymer chain [8].

In addition to the previous two, another thermodynamic state can be considered: the *molten globule* phase [49,50]. Under certain conditions, proteins can also exhibit this phase which has some native secondary and tertiary structure but lacks well–packed side chains. A scaling analysis of the thermodynamic properties of proteins of various length suggests that the molten globule is indeed a distinct phase separated by first order transitions from both the native and the unfolded state, although there are interesting exceptions as, for example, α–lactalbumin which has a trasition from the unfolded state to the molten globule phase that appears to be continuous [50].

Finally, it must be mentioned that when a protein fails to fold into the native structure, it can reach a *misfolded state*. These misfolded proteins can aggregate giving rise to the formation of amyloid fibrils which are believed to be the cause of several neurodegenerative and other diseases [2,3].

As already partially discussed in sections 1.3 and 1.4 the stability of proteins is the result of residue–residue and residue–solvent interactions. All these interactions are eventually electrostatic in origin but it is convenient to distinguish them into Van der Waals interactions, salt bridges, disulfide bridges, hydrogen bonds and hydrophobic forces.

Both salt and disulfide bridges are strong interactions between non contiguous residues, the former is the interaction that can occur between opposite charged groups while the latter can occur only between two cysteine residues. However the number of salt bridges and disulfide bonds is generally small. Van der Waals forces are instead quite weak and, furthermore, in aqueous solutions they should occur not only between the protein groups but also between these groups and water and they are usually supposed to be almost identical. If so, they would compensate each other, and their overall contribution to the stabilization of protein structure would be very small, if not zero. The same could be said about hydrogen bonding, since protein residues can form hydrogen bonds not only between themselves but also with water molecules and therefore it is usually assumed that they play a minor role in stabilization[2], although they are important in directing protein folding being the cause of the secondary structure order (see section 1.3).

Therefore, by elimination, one reaches the conclusion that the stabilization of the native structure is primarily due to the hydrophobic effect. It results in the burial of the hydrophobic residues in the core of the protein and it is exemplified by the fact that oil and a polar solvent like water do not mix because oil is not able to form hydrogen bonds with the surrounding water. The hydrophobic effect is possible only in presence of a polar solvent and it is known that proteins do not fold in a nonpolar environment. Thus the hydrophobic interaction occupies, in the life sciences, a position of importance comparable to any of the four fundamental forces in the physical sciences but, unfortunately, the factors which give rise to the hydrophobic effect are complex and still incompletely understood and we do not,

[1] In polymer physics the radius of gyration is used to describe the size of a polymer chain. It is defined as $R_g^2 = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{r}_k - \mathbf{r}_{\text{mean}})^2$ where N is the number of monomers and $\mathbf{r}_{\text{mean}}$ their mean position.

[2] Here it is worth to clarify that also the hydrophobic effect is associated with hydrogen bonding since it is due to a lack of hydrogen bonds, as it will be shown in the following paragraph.

for example, really know its force range.

The hydrophobic effect has an entropic origin. At low temperatures, when a nonpolar substance is brought into a polar solvent, like water, the latter tends to form ordered cages around the nonpolar molecules [51] as shown in figure 2.1.

A possible mechanism for this ordering is that, while a solvent molecule in solvent has a choice to form hydrogen bonds with any of its neighbors, a solvent molecule close to a hydrophobic surface prefers to form hydrogen bonds with other similar molecules rather than to "waste" these bonds by pointing toward the nonpolar solute. If the surface of the nonpolar solute grows, the amount of the ordering of the solvent increases, leading a decrease in the total entropy. Instead enthalpy variations at room temperature are negligible and the conclusion is that the state in which nonpolar compounds are "packed" together is the thermodynamically favoured state, having the smallest free energy. The ordering of the solvent is however expected to melt away at sufficiently high temperature and thus the hydrophobic effect varies substantially with temperature.

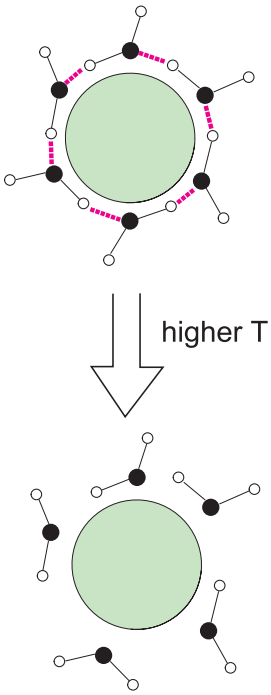


Figure 2.1: Water molecules around a hydrophobic particle at different temperature.

Coming back to proteins, it is now possible to understand the denaturation at high temperature. The loss of the native state stability, under warm unfolding conditions, is the result of the combined effect of the vanishing of the hydrophobic interaction and of the thermal fluctuations of the polypeptide chain, which cause the breaking of the other above mentioned bonds. What is surprising however is the possibility of a melting of proteins at low enough temperatures which is called *cold unfolding* and which has been observed experimentally. A possible reason for this phenomenon is that at low enough temperatures the enthalpy decrease contribution becomes relatively more important cancelling the benefit of increasing entropy by segregating hydrophobic residues and making a swollen state of the chain more stable than a folded structure which excludes the solvent [52].

2.2 Folding thermodynamics

The protein stability is defined as the free energy change ΔG between the folded and unfolded states:

$$\Delta G = \Delta H - T\Delta S\,,$$

where T is the temperature and ΔH and ΔS are respectively the enthalpy and entropy changes between the folded and unfolded states. Proteins become more stable when the free energy difference between the unfolded and the native states increases.

Folding of the random–coil polypeptide chain into a unique conformation entails a tremendous increase in the order of the macroscopic system and should be related to a significant decrease of the entropy of the system. An entropy decrease is thermodynamically unfavourable, so one would expect this entropy effect to be compensated by the energy gained as a result of redistribution of various intramolecular interactions between the protein groups and the environment. The final result is that, for a typical protein, the global stability at room temperature is of the order of about 10 kcal/mol (about 20 k_BT for a protein of 100 residues) which is a small free energy difference, being just the energy of a few hydrogen bonds. If this free energy difference would be splitted between the various residues, as in the case of cooperativity lack, then the folded structure would not be stable. It follows that a protein has an ordered native structure only because it is a cooperative system. The role of the cooperativity also explains why it is not enough for a heteropolymer to have a unique folded conformation in order to be considered protein–like. A polymer with a random sequence could have some lowest energy conformation and, under appropriate temperature and solvent conditions, it could eventually fold to this "native" state. But this transition would be only weakly cooperative thus differing from the folding transition of proteins. Moreover, unlike proteins, the "native state" of a random sequence will be very sensitive to mutations. The mechanisms of protein cooperativity are still not completely understood but they seem to be a peculiarity of the aperiodic structure of proteins [53] because only such a structure seems to provide the complex interlacing of short and long range interactions necessary for cooperativity. As a consequence of the folding cooperativity the overall features of protein folding thermodynamics are quite simple and it has been found[3] that the overwhelming majority of small proteins fold in all–or–none manner. Thus the folding reaction can be well modelled as a two–state transition between a random–coil and the ordered native state.

2.3 Folding kinetics

Besides the prediction of the final structure of a protein starting from its amino acids sequence, the protein folding problem involves also another relevant aspect which is the time that the polypeptide chain needs to reach its final three–dimensional shape. In the late 1960s, Levinthal formulated its famous paradox [4, 5, 54]: if a protein folds by sequentially sampling all possible conformations, given the great amount of degrees of freedom, folding would take an amount of time greater than the life of the universe, even if the conformations are sampled at a rate of picoseconds. For example a chain of 100 residues with 3 degrees of freedom each (i.e. 3 possible combination of the dihedral angles of each residue), with a conformation sampling rate of 10^{12} s^{-1}, would need 10^{27} years to fold. Nevertheless, the fastest proteins take about 10 μs to fold and the slowest ones take just some minutes.

It is likely that only those proteins chains that can fold in a short time were chosen by evolution to function in the living cell but questions remains on which

[3] In general the principal way of exploring the thermodynamics properties is through delicate calorimetry experiments that allow, for example, to measure the specific heat, but also other methods, such as viscosimetry, Circular Dichroism (CD) and nuclear magretic resonance (NMR), have contributed significantly.

kind of mechanisms proteins adopt to overcome Levinthal's paradox. Levinthal himself proposed that a random conformational search does not occur, and the proteins must, therefore, fold through a series of meta–stable intermediate states which follow one another and drive the protein to the native state, i.e. a protein follows a specific folding pathway to reach its native configuration.

In 1973 Ptitsyn proposed that protein folding proceeds through three sequential steps. This model, which has been later called the *framework model*, postulates an initial folding step from the unfolded chain to a pre–molten globule state in which α–helices and β–sheets are already formed but they fluctuate around the native position. This process should be quite fast, as confirmed by many experiments and theoretical studies (for a more detailed discussion see the review by Finkelstein and Galzitskaya and references therein [9]). In the second step these secondary structure elements glue together in the molten globule state and finally, in the last step, also the side chains order to form the completely native protein. To date, all these folding steps and intemediates have been observed experimentally studying the folding at biological conditions.

However, folding can also occur in the zone of thermodynamic equilibrium, where the folded and the unfolded state have almost the same probability to occur, while all the intemediates are unstable and thus do not accumulate. Besides, in 1990s many simple single–domain proteins were found to fold very rapidly as two–state systems without any detectable intermediates in a wide range of conditions [55].

Thus, for some proteins in a wide range of conditions and for the others in the thermodynamic equilibrium zone, the folding kinetics looks like very simple since all the properties of the native state are recovered simultaneously. The idea is thus that the basic features of protein folding could be grasped with the investigation of folding in the transition zone where the unnecessary complications given by the individual behavior of each protein can be partially removed. Here it is possible to indirectly study the folding *transition state ensemble* which is the ensemble of structures corresponding to the saddle point in the free energy which separates the folded and the unfolded states. The folded part of the transition state is called "folding nucleus". The modern perspective is that the folding proceeds via a nucleation and growth mechanism [6] in which only the folding nucleus needs to form its native contacts by "chance" and then the rest of folding process is downhill with the polypeptide chain that attains fast its native structure by accretion. The evidence of a specific nucleus or many nuclei in the early stages of the folding allows to draw a close analogy between this process and the transformation of a vapor into a liquid. The discussed cooperativity of the protein folding is thus similar to that exhibited in first–order phase transitions.

2.3.1 Free energy landscape and reaction coordinates

Energetically the folding of a polypeptide can be seen from an energy landscape point of view. For an N–atom system, the energy landscape is a hypersurface in a $3N$–dimensional space defined by the atomic coordinates with each point on the landscape corresponding to a single set of them. For a protein the number of coordinates is too big to be treated and to extract some useful thermodynamical and dynamical information. Thus the possibility of distinguishing the various phases of

the system and of getting the basic features of protein dynamics is connected to an accurate selection of some suitable macroscopic observable of which the free energy will be a function. These observables are named *reaction coordinates* and, within this picture, the free energy at a given value of the reaction coordinate depends on the enthalpy and the entropy of the ensemble of microscopic states with that value of the reaction coordinate. In recent theoretical literature the number of native residue or the end–to–end length of the protein are often used as reaction coordinates. However, the proper selection of the reaction coordinates is highly nontrivial and still an unresolved issue. Furthermore the protein folding properties should be better described by a multidimensional free energy landscape which depends on more reaction coordinates.

The free energy landscape turns out to be a very useful tool in displaying and conceptualizing the phases of a protein. Given a particular value X of the reaction coordinate x, the quantity $e^{-G(X)/k_BT}$ is proportional to the probability that the microscopical conformation of the polypeptide chain has $x = X$. Thus, the same quantity is also proportional to the time that the system spends with $x = X$. The thermodynamic states can thus be associated with the basins of the local free energy minima and the depth of each minimum shall be related to the temporal persistence of the corresponding phase, i.e. to its stability. At the thermodynamic equilibrium point, the free energies of the native and unfolded basins are equal because of a balance between the few low energy conformation of the former and the enormous number of largely unrelated disordered microstates of the latter.

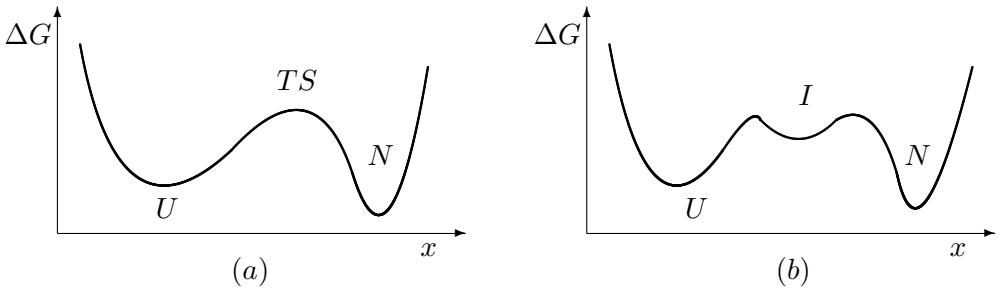


Figure 2.2: Schematic view of the free energy landscape of a two–state folder (a) and of a protein with an intermediate state (b) as a function of the reaction coordinate x. U represents the unfolded state, N the native state, TS the transition state and I an intermediate.

Two common and schematic representation of free energy landscapes as a function of some reaction coordinate x are shown in figure 2.2.

A final remark is necessary: the energy landscape of a protein is not only an inherent property of the protein, but it is to a great extent also influenced by the surrounding environment. Changes in temperature and denaturant concentration of the solution in which the protein is immersed are then reflected in changes in the free energy profile.

2.3.2 Folding Funnels

In the late 1980s and early 1990s, Joseph Bryngelson and Peter Wolynes, inspired by the statistical mechanics of spin glasses, proposed that, in native conditions, proteins have globally "funneled energy landscapes", similar to that shown in figure 2.3, with the energy of the native conformation at the bottom of this funnel [56].

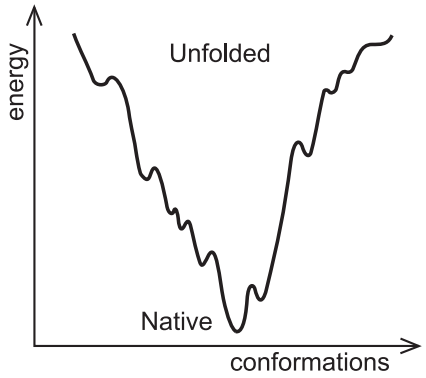


Figure 2.3: Schematic view of the folding funnel.

This "folding funnel" landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather than being restricted to a single mechanism. For most of the proteins, the folding funnel is not smooth but it has many energy barriers and local free energy minima in which proteins can also get "stuck", leading to misfolded states characterized by the presence of favorable but non–native interactions. In order to avoid these "kinetic traps" and to allow a fast kinetics, the degree of ruggedness of the funnel cannot be too great. However proteins have twenty different types of amino acids with different binding affinities, many degree of freedom and steric constraints and a large degree of frustration would then be expected *a priori* from a comparison with classical spin glasses (which have many low–energy states separated by high barriers). To rule out this scenario, the *principle of minimal frustration* has been introduced by Bryngelson and Wolynes themselves. This principle says that natural selection has chosen and improved amino acid sequences so that the undesired interactions between amino acids along the folding pathway are reduced in order to have a smoother funnel landscape and making the acquisition of the folded state a very fast process. This idea is radically implemented by an entire class of physical model for protein folding, called Gō–models [19], in which only native interactions are assumed to exist.

2.3.3 Folding and unfolding rates

For proteins with a two–states behavior, the transition state instability determines the folding (k_{f}) and unfolding (k_{u}) rates, which, according to conventional

theory, follows the Kramers' formula [57]:

$$k_{\mathrm{f}} = k_0\, e^{-(G_{\mathrm{TS}}-G_{\mathrm{U}})/k_BT}, \tag{2.1}$$

$$k_{\mathrm{u}} = k_0\, e^{-(G_{\mathrm{TS}}-G_{\mathrm{N}})/k_BT}, \tag{2.2}$$

where G_{TS}, G_{U} and G_{N} are the free energy of the transition, unfolded and native state, respectively, as shown in figure 2.4 and k_0 would be the transition rate from the folded to the unfolded phase in the absence of the free energy barrier.

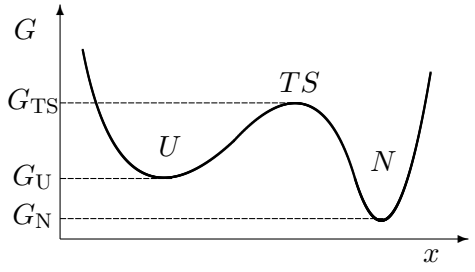


Figure 2.4: Free energy of the native (N), unfolded (U) and transition (TS) state.

Experimentally one measures the relaxation of a mixture of unfolded and folded proteins, after changing the denaturant concentration. The relaxation rate is given by $(k_{\mathrm{f}} + k_{\mathrm{u}})$ as a function of the denaturant concentration. One thus obtains a plot like that in figure 2.5, called "chevron" plot for its characteristic V–shape.

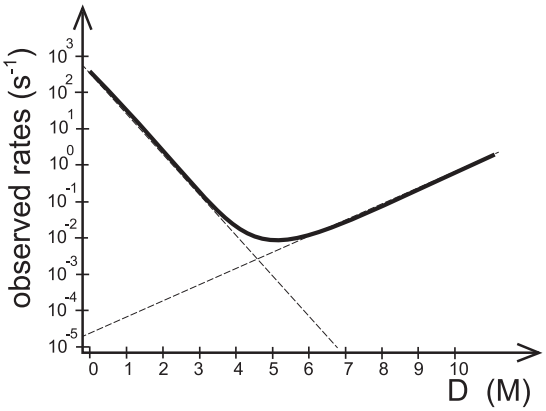


Figure 2.5: Typical "chevron" plot of a protein folding experiment: observed rate as a function of the denaturant concentration D. The linear behavior of the rates allows extrapolation to zero denaturant conditions.

Folding and unfolding rates dominate the observed relaxation rates respectively below and above the denaturation midpoint. This gives rise to the terminology of folding and unfolding arms for the limbs of the chevron. In a two–state model, the logarithm of the folding and unfolding rates is assumed to depend linearly on the denaturant concentration according to two parameter $\mathrm{m_f}$ and $\mathrm{m_u}$, respectively called the folding and unfolding m–values (the slopes of dashed lines in figure 2.5).

One of the few universal features of protein folding kinetics is that the logarithm of the folding rate has a simple linear correlation with the *relative contact order* [58] (RCO) which is the average separation along the sequence of residues that make contacts in the three–dimensional native structure, divided by the total number of residues. Many studies have shown similar good correlation also with other structural parameters of the native state such as for example the *absolute contact order* [55] (ACO, equal to RCO times the number of residues). The important role of the native state topology can be understood by considering that the formation of contacts between residues that are distant along the sequence has a great entropic cost, because it greatly reduces the number of conformations available, and thus it is less likely. Therefore the basic idea is that proteins with a simpler topology are faster to fold.

2.3.4 Transition state ensemble

As already partially discussed, the transition state ensemble is the ensemble of states which correspond to the saddle point of the free energy between the folded and the unfolded state and which contain one or more folding nuclei. These states have, by definition, the same probability of folding or unfolding. The free energy barrier which separates the unfolded and the folded basins has an entropic origin; it arises from incomplete cancellation of the entropy loss upon forming the folding nucleus and the enthalpy gain of making the nucleus. For the folding kinetics, the nucleus acts as a bottleneck composed of a still relatively large number of configurations which are on the edge between the folded and unfolded basins of attraction.

A theoretical way of determining directly the transition state ensemble is using the folding probability method, thereby avoiding ambiguities associated with the choice of a reaction coordinate. The folding probability p_{fold} of a conformation is the probability to fold, leaving from such a conformation, without recrossing back to the unfolded basin. The transition state ensemble is then defined by selecting conformations that have the same probability of reaching first either the native or the unfolded state, that is, those configurations that have $p_{\text{fold}} = 1/2$.

There is also an experimental method to study the nature of the transition state ensemble which is based on the ϕ–value analysis [59, 7] introduced by Fersht. The basic idea entails the substitution, through protein engineering, of amino acids in different positions of a protein with other amino acids. Monitoring the resulting changes in the stability of the native state and in the kinetics of folding, it is possible to find those residues partecipating to the folding nucleus as those residue whose mutations affects the folding rate by changing the transition state stability as strongly as that of the native protein. ϕ–values are defined as follows:

$$\phi = \frac{\Delta \ln k_{\text{f}}}{\Delta \ln K},$$

where $K = k_{\text{f}}/k_{\text{u}}$ is the folding–unfolding equilibrium constant.

According to Fersht's interpretation, a residue which participates in the same interactions in both the native and the transition states would ideally have $\phi = 1$, whereas a residue with $\phi = 0$ is likely to be unstructured in the transition state. The values $\phi \approx 1/2$ are ambiguous and can be interpreted in two different ways:

either the residue is at the surface of the nucleus, making native interactions with only half of the neighboring residues, or it belongs or not to the folding nucleus according to different folding pathways. Finally, it is worth to note that the values $\phi < 0$ and $\phi > 1$, which are in principle inconsistent with the model of a native–like folding nucleus, can occur even if they are extremely rare. They are due to the fact that experimental errors can be high in measuring equilibrium stability as well the folding and unfolding rates in water for the wild–type protein and mutants. The necessity of extrapolating ϕ–values in pure water from measurements made in solutions containing denaturants adds uncertainty to the reported values. When the stability difference between the native and mutant protein are low, experimental error can be very large. Unusual ϕ–values outside the $[0, 1]$ range may arise from these errors rather than illustrating deviations from the conditions assumed by the method [60].

The theoretical and experimental investigation of nucleation mechanism is far from being completed, for example it is still not completely clear how big the folding nucleus must be. Furthermore proteins with different amino acid sequences but similar three–dimensional structure have similar folding nuclei but also this rule is not always true and has several exceptions [9].

Chapter 3

Mechanical unfolding of proteins

In the last fifteen years modern experimental techniques, as *atomic force microscopy* (AFM) [10] or *laser optical tweezers* (LOT) [11], have been developed and have allowed the microscopic manipulation of biomolecules, not only proteins but also RNA and DNA fragments. The main advantage of these techniques is their ability to separate out the fluctuations of a single molecule from the ensemble average behavior observed in traditional biochemical–biophysical experiments. Exerting a mechanical force, it is possible to induce unfolding of the molecule and to measure the binding forces responsible for the stability of biomolecules and/or to explore the unfolding pathways and the possible intermediate states.

Nowadays AFM and LOT have proved to be very useful tools to investigate folding and refolding of protein but one could argue about the the fact that mechanical unfolding kinetic properties may be different from thermal and chemical ones since they are pathway dependent and the pathways may differ in the two cases. Unfortunately experiments have shown that this is the case [12]. Differences arise from the fact that temperature and chemical denaturants have a global effect on an entire protein while the force is applied locally to the termini and thus the protein always starts to unravel at its termini.

Furthermore, cysteine engineering allows to pull the molecule along different pulling directions [61,62] or to study the unfolding properties of a polyprotein where each module is connected to the neighboring ones through different points of force application [62]. Experiments showed that mechanical stability of a protein may be very different (even an order of magnitude) according to the pulling direction. The pronounced effect of pulling direction on protein stability may be explained as follows: if a chain is pulled along the direction of hydrogen bonds, then the unfolding is akin to unzipping, but when force is applied perpendicular to this direction, the unfolding is akin to shearing. The force needed to break hydrogen bonds in the latter case should be larger than in the former one.

Finally, the mechanical manipulation of biomolecules became important also in conjunction with fluctuation relations that describe the behavior of a system driven out–of–equilibrium. In fact, most manipulation experiments are actually performed by switching the system faster than its slowest relaxation rate and therefore in out–

of–equilibrium conditions. Working out–of–equilibrium normally does not allow to obtain equilibrium information from experimental data but there exist relations such as Crooks equality [63] and Jarzynski equality [64] that make it possible to extract equilibrium properties of the system, such as the thermodynamic free energy differences, from non equilibrium experiments based on the measure of the work done on the system.

3.1 Experimental techniques

The atomic force microscope and the optical tweezers allow two different kind of protocols: in the *constant velocity protocol* the distance between the molecule termini is increased at a constant velocity. The increasing force is applied to the protein until it unfolds and the force abruptly drops down. When this is done with an engineered protein composed of several domains which fold in an all–or–none fashion, the outcome of the experiment results in a typical sawtooth pattern in the force–extension curve (see for example reference [65]), where each peak is attributed to the breakage of a folded domain. However, many proteins mechanically unfold via intermediates and, in these cases, the force–extension curve can present a "hump" and/or a specific peak due to the unravelling of some secondary structure element (see for example references [66, 67]). Typical rupture forces are in between 10 and 600 pN [68] where the upper limit is not far from rupturing covalent bonds.

The atomic force microscope and the optical tweezers can be used also in a *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. In this case, instead of the force–extension curve, one studies the time dependence of the end–to–end distance which is typically stepwise (see for example references [69, 12]).

In AFM one terminal of a biomolecule is anchored to a gold substrate while the other terminal is attached to a force sensor and the biomolecule is stretched by increasing the distance between the surface and the force sensor by moving at a constant velocity v either the cantilever or the surface, depending on the experimental apparatus (constant velocity protocol) [65, 68]. The force sensor is a cantilever with stiffness k, and the force f experienced by the molecule is obtained by measuring the cantilever bending δx by a laser and applying Hooke's law $f = k\delta x$. The spring constant of the cantilever tip is typically $k = 10 \div 10^3$ pN/nm and the resolution is generally of few pN, thus AFM is an ideal tool for studying relatively strong inter and intra–molecular interactions. The operating principle of LOT is similar but instead of being anchored to the substrate and the cantilever, the molecule is held by two micro–sized polystyrene or silica beads [68]. The radiation pressure from a focused laser beam is able to optically trap one of these beads. Since the trapping potential is harmonic, the force acting on the bead could again be expressed by $f = k\delta x$ where δx measures the deviation from the trap center and k is the stiffness constant of the trap which has a typical value of $k = 10^{-3} \div 10^{-1}$ pN/nm. Thus the force resolution of LOT is at least 10 times better than AFM, but LOT is also less sensitive to changes in extension and this could mask minor unfolding events like short–lived intermediates. This method entails also the use of DNA molecules as molecular handles to manipulate individual proteins or polyproteins between the

two beads. These handles can have different lengths, they function as spacers between the protein and the beads and keep the interactions between the tethering surfaces to a minimum, thus allowing to study protein folding in the physiologically relevant low–force regime [70].

3.2 Kinetic theory

In these kind of experiments the end–to–end length L of the protein is directly measurable or controlled by instrumentation. L thus results to be a very well defined reaction coordinate to describe the mechanical unfolding process, which makes comparison with theory and simulations easier. The theoretical framework for understanding the effect of external constant force on unfolding rate was first discussed by Bell in 1978 [71] and later extended by Evans and Ritchie to deal with the case in which the force increases linearly with time [72, 73].

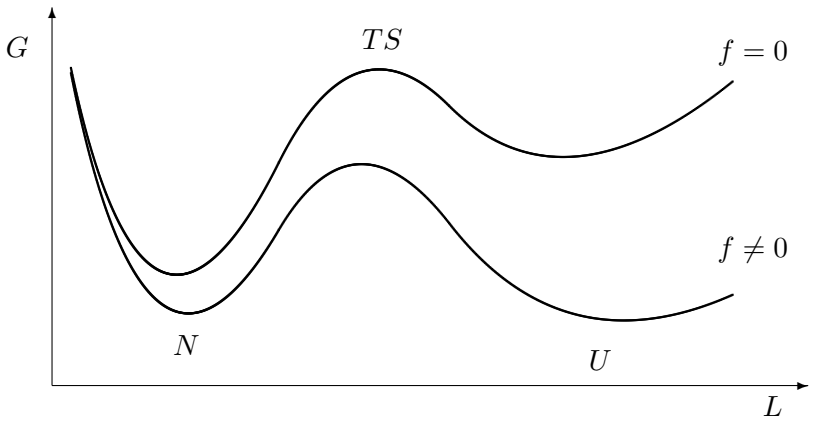


Figure 3.1: Free energy profile of a protein in the absence and in the presence of a constant pulling force f.

The free energy profile as a function of the end–to–end length $G(L)$ is tilted to $G(f;L) = G(L) - fL$ when a constant force is applied to the molecule ends (see figure 3.1). Assuming that the force does not change the distance x_u between the native state (N) and transition state (TS), one obtains that

$$G_{\mathrm{TS}}(f) - G_{\mathrm{N}}(f) = G_{\mathrm{TS}}(0) - G_{\mathrm{N}}(0) - fx_u \ ,$$

where $G_{\mathrm{TS}}(0)$ and $G_{\mathrm{N}}(0)$ are the free energy in the absence of force of the transition state and the native state respectively, and x_u can be regarded as the width of the potential barrier. Inserting the above relation into Kramers' formula 2.1, Bell obtained the following unfolding rate constant:

$$k_{\mathrm{u}}(f) = k_{\mathrm{u}}(0)\, e^{fx_u/k_BT} \ , \tag{3.1}$$

where $k_{\mathrm{u}}(0)$ is the unfolding rate constant in the absence of force. It is worth noting that the unfolding rate grows exponentially with the force, which is the hallmark

of the phenomenological Bell's model. An analogous reasoning results in a similar expression for the refolding rate:

$$k_{\mathrm{f}}(f) = k_{\mathrm{f}}(0)\, e^{-f x_f / k_B T} \,, \tag{3.2}$$

where $k_{\mathrm{f}}(0)$ is the refolding rate constant in the absence of force and x_f is the distance between the transition state and the unfolded state. Using these equations and the force dependence of folding/refolding rates, it is possible to find the position of the transition state in the free energy profile of figure 3.1. Furthermore, using equation 3.1, it is possible to obtain an expression for the distribution of unfolding time conditioned to the constant force f, $P(t\,|\,f)$. Let us consider that the probability $n(t)$ that the protein is still in the native state at time t can be calculated by solving the differential equation

$$\frac{dn(t)}{dt} = -k_{\mathrm{u}}(f) n(t) \,, \tag{3.3}$$

where we have assumed the refolding rate to be negligible. By noting that $P(t\,|\,f)dt = -dn(t)$ and using the boundary condition $n(0) = 1$, it follows that the distribution of unfolding time has a negative exponential behavior:

$$P(t\,|\,f) = k_{\mathrm{u}}(f)\, e^{-t\, k_{\mathrm{u}}(f)} \,. \tag{3.4}$$

Instead, assuming that the force increases linearly with time, $f = rt$ (where r is a constant rate) and making the same Bell's assumption regarding the barrier width x_u, Evans and Ritchie [72,73] obtained an expression for the force distribution and for the most likely rupture force. Using equations 3.1 and 3.3 with $f = rt$, it follows:

$$n(t) = \exp\left\{ \frac{k_B T}{r x_u} k_{\mathrm{u}}(0) \left(1 - e^{r t x_u / k_B T}\right) \right\} \,. \tag{3.5}$$

One can relate the distribution of rupture forces conditional to the rate r, $P(f^*\,|\,r)$, to the survival probability $n(t) = n(f/r)$ by noting that

$$P(f^*\,|\,r)\, df^* = -dn(f^*/r) = -\frac{dn(t)}{dt} dt \,. \tag{3.6}$$

Inserting equation 3.5 into equation 3.6, one obtains:

$$P(f^*\,|\,r) = \frac{k_{\mathrm{u}}(f^*)}{r} \exp\left\{ \frac{k_B T}{r x_u} \left[k_{\mathrm{u}}(0) - k_{\mathrm{u}}(f^*)\right] \right\} \,. \tag{3.7}$$

Finally, maximizing the above expression, it is possible to find the most likely rupture force $f_{\mathrm{rupt.}}$:

$$f_{\mathrm{rupt.}} = \frac{k_B T}{x_u} \ln\left(\frac{r x_u}{k_B T\, k_{\mathrm{u}}(0)} \right) \,. \tag{3.8}$$

In the constant velocity protocol it is often fair to assume $r = kv$ and extensive AFM experiments [74] have confirmed the logarithmic dependence of mean rupture force on the pulling speed, $\langle f^* \rangle \sim \ln v$.

As already mentioned, the major shortcoming of previous arguments is that x_u does not depend on the external force, while a careful look at figure 3.1 shows

that in general this is not the case. Hummer and Szabo [75] proposed a more sophisticated but still analytically tractable procedure. In this theory, by applying the Kramers theory of diffusive barrier crossing [57] to a simple model free–energy surface, it is possible to extract not only $k_{\rm u}(0)$ and x_u but also the free energy difference, $\Delta G_u = G_{\rm TS}(0) - G_{\rm N}(0)$, between the transition state and the native state in the absence of external force. In the limit $\Delta G_u \to \infty$ this theory reduces to the phenomenological approach and predicts that at relatively high pulling speed[1] $\langle f^* \rangle \sim (\ln v)^{1/2}$. At the same time, Dudko *et al.* [76] proposed a model that allows to extract the critical force at which the barrier to rupture vanishes, the free energy of activation and a parameter proportional to the diffusion constant. Their theory predicts that $\langle f^* \rangle \sim (\ln v)^{2/3}$. Thus all these theories are in disagreement and one may suspect that they are valid in different regimes. More recently, Szabo and coworkers [77] have developed an approach that casts the phenomenological and microscopic theories in a unified framework. Assuming a single–well free energy surface, they found a formula for the unfolding rate under a constant force f, that reproduces the above theories and the Bell equation depending on the value of an exponent ν:

$$k_{\rm u}(f) = k_{\rm u}(0) \left(1 - \frac{\nu\, x_u f}{\Delta G_u}\right)^{1/\nu - 1} \exp\left\{\frac{\Delta G_u}{k_B T}\left[1 - \left(1 - \frac{\nu\, x_u f}{\Delta G_u}\right)^{1/\nu}\right]\right\} . \qquad (3.9)$$

In the case $\nu = 1$ the above expression reduces to the Bell expression 3.1. If $U_0\,(x)$ is the potential energy along the pulling direction x in the absence of external force, the values of $\nu = 1/2$ and $\nu = 2/3$ correspond respectively to assume a cusp potential

$$U_0\,(x) = \begin{cases} \Delta G_u \left(\frac{x}{x_u}\right)^2 , & \text{for } x < x_u\ , \\ -\infty\ , & \text{for } x \geq x_u\ , \end{cases} \qquad (3.10)$$

or a linear–cubic potential

$$U_0\,(x) = \frac{3\,\Delta G_u}{2}\frac{x}{x_u} - 2\,\Delta G_u \left(\frac{x}{x_u}\right)^3 . \qquad (3.11)$$

According to this theory, when $f = rt$, the distribution of rupture forces is

$$P(f^*\,|\,r) = \frac{k_{\rm u}(f^*)}{r} \exp\left\{\frac{k_B T}{r x_u}\left[k_{\rm u}(0) - k_{\rm u}(f^*)\left(1 - \frac{\nu\, x_u f^*}{\Delta G_u}\right)^{1/\nu - 1}\right]\right\} , \qquad (3.12)$$

and the mean rupture force is

$$\langle f^* \rangle \cong \frac{\Delta G_u}{\nu x_u}\left\{1 - \left[\frac{k_B T}{\Delta G_u}\ln\frac{k_B T\, k_{\rm u}(0)\, e^{(\Delta G_u/k_B T)+\gamma}}{r x_u}\right]^{\nu}\right\} , \qquad (3.13)$$

[1] Qualitatively, Hummer and Szabo found three pulling regimes predicted by their theory. In an intermediate range of pulling velocities, in which experiments are typically conducted, the average force depends approximately linearly on the logarithm of the pulling velocity. Below, the average rupture force becomes linearly dependent on the pulling speed, and in the above regime, the force at rupture grows as $(\ln v)^{1/2}$.

where $\gamma \simeq 0.577$ is the Euler–Mascheroni constant. If the free energy barrier ΔG_u is large compared to $k_B T$, then the contribution of γ is negligible and in the phenomenological limit $\nu \to 1$, equations 3.7 and 3.8 are recovered. It is appealing to analytically continue ν in the above expressions to all ν and thus ν can be used as an additional fitting parameter to find the best agreement with experiments.

In the derivation of both phenomenological and microscopic models, the adiabatic assumption, equation 3.3, has been implicitly used. This is valid if the pulling rate r is low enough, so that the system ruptures when the activation energy is still large. At extreme pulling speeds or external forces, the adiabatic approximation breaks down and the above theories become inapplicable [75]. If this adiabatic approximation is indeed valid and $f(t) = rt$, then the product $r \ln n[t(f)]$ as a function of f is independent of r [78]. In this case, the following relation between the constant force experiments (measuring $k_\mathrm{u}(f)$) and constant speed experiments (measuring $P(f \,|\, r)$) has been established [77],

$$k_\mathrm{u}(f) = \frac{r\, P(f \,|\, r)}{1 - \int_0^f P(f' \,|\, r) df'} \,. \tag{3.14}$$

Chapter 4

The Wako–Saitô–Muñoz–Eaton model and its generalization for mechanical unfolding

Although no method exists that can reliably predict the three–dimensional structure of a protein from its amino acids sequence, many plausible models have attempted to describe protein folding. Some of them, called phenomenological models, generally aim to solve the Levinthal's paradox. Examples of this kind of models are the nucleation–growth models [6] and the framework model [13] already discussed in section 2.3, the diffusion–collision model [14] and the jigsaw puzzle model [15].

However, more recently, many coarse–grained [19,20,21,22,23] and all–atom [16, 17,18,79] models have been developed and have made possible to simulate the folding of proteins at a greater level of details. The success of these simulations came partly from the enhanced power of computers and partly from a better knowledge of the general principles of protein folding. Knowledge that, on the other hand, have been greatly improved by these models.

In principle, molecular dynamics all–atom simulations can be used to obtain information concerning the detailed protein folding kinetics. Nevertheless, there are strong limitations given by the computational power available and by the fact that the force fields used are not known to a sufficient extent. As a consequence, nowadays simulations no longer than few μs can be performed. However all–atom simulations have been quite successful to study the function of proteins involved in biological short–time processes [80] or to study unfolding under special conditions, such as for example high temperature, that reduce the time involved by several orders of magnitude. Other studies deal with very small fragments of the polypeptide chain, so that the conformation space is small enough to manage simulations on the right time scale.

To overcome the problem connected to all–atom simulations, models with a simplified description of the protein can be used [23]. These models usually reduce the number of degrees of freedom of the polypeptide chain by coarse–graining, i.e. by gathering together groups of atoms in fewer particles which interact through specific interactions. Many levels of coarse–graining can be adopted but one of the usual choices is to describe each amino acid as a single particle centered around

its C_α atom. Another simplification relies on the use of sampling strategies, such as Monte Carlo techniques, that allow an exploration of protein conformational space that is faster and more efficient than that provided by molecular dynamics. Next, one can still decide whether to use a continuous (off–lattice) or a lattice based representation. A very interesting type of lattice model have been introduced by Gō and Taketomi [19] in 1978. They did not attempt to predict the native structure from the amino acid sequence but, instead, they try to elucidate the main features of folding assuming a target structure and by assuming that only amino acid pairs which interact in the ground state interact at all. This assumption biases the folding towards the native state by eliminating the frustration due to non–native interactions and, by imposing a sufficiently large energy gap between the unfolded and the folded state, it allows to reproduce the two–state behavior.

An alluring case of intrinsically cooperative simple model is represented by the Wako–Saitô–Muñoz–Eaton (WSME) model. This model was first introduced by Wako and Saitô in 1978 [24, 25] and then forgotten for some time until Muñoz, Eaton and coworkers proposed it again in the late 1990s [26, 27, 28].

Equilibrium thermodynamics of this model can be solved exactly [29]. The model has remarkable equilibrium properties [81] and it successfully describes the kinetics of protein folding [82, 83, 84, 85]. More recently a generalized version of the model has been proposed, which permits to reproduce the general features of mechanical unfolding [30, 86] and, through Monte Carlo simulations, to obtain (for some already widely studied proteins and RNA fragments) unfolding pathways which are consistent with results of experiments and/or of simulations made with more detailed models [87, 88, 89, 90]. The model has also been used with success to study folding equilibrium and kinetics and to mimic mutations of a small ankyrin repeat protein [91] and a modified version of it has been applied to the study of the conformational dynamics of photoactive yellow protein [92]. Furthermore, the model have even found applications in the study of amyloid aggregation [93] and in problems of strained epitaxy [94].

4.1 The model

The WSME model is a Gō–like model which aims to describe the protein folding process assuming to know the native structure of the protein and that only native interactions exist. It is also an Ising–like model in the sense that residues can have only two states, native or non–native, with energy favoring the native state and entropy favoring the non–native state.

In this model, a given N residues protein is described by a sequence of N binary variables m_k, with k numbered from the N–terminus to the C–terminus. The variable m_k is equal to 1 if the k–th residue is in the native configuration while it is equal to 0 otherwise. For the k–th residue, to be in the native configuration means that the respective dihedral angles, ϕ_k and ψ_k, have a value equal (or at least close) to the value that they have in the native structure. As discussed in section 1.2, the pair (ϕ_k, ψ_k) cannot assume any value on the torus $(-\pi, \pi] \times (-\pi, \pi]$ but it is limited by the allowed regions of the Ramachandran map. In any case, the disordered state allows a much larger area than the native conformation and the model deals with this fact by assigning an entropic cost q_k to the ordering of k–th residue.

Two residues interact only if they and all residues between them are native and only if they are in contact in the native structure, i.e. they have at least a pair of atoms which are closer than a given threshold length[1] in the native structure available in the Protein Data Bank [37]. Then, if residues i and j are in contact in the native structure, a negative energy $-\varepsilon_{ij}$ and a contact matrix element $\Delta_{ij} = 1$ are associated to them, while $\Delta_{ij} = 0$ if the two residues are not in contact. An usual choice for the parameters ε_{ij} is to define them, according to the prescription of Muñoz and Eaton [28], as $\varepsilon_{ij} = k\varepsilon$, where ε is a parameter that must be determined through comparison with experiments, k is an integer such that $5(k-1) < n_{at} \leqslant 5k$, and n_{at} is the number of pairs of atoms in contact between i-th and j-th amino acids. In the following this prescription will be used but it must be noticed that it is not the only one. For example one can choose the ε_{ij} to be proportional to the number of atoms in contact between i-th and j-th residue ($\varepsilon_{ij} = n_{at}\varepsilon$) or they can be even treated as random variables [95].

The Hamiltonian of the model, or, more precisely, its configuration dependent free energy, reads:

$$H(m) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\varepsilon_{ij}\Delta_{ij}\prod_{k=i}^{j}m_k - k_BT\sum_{i=1}^{N}q_i(1-m_i)\,, \tag{4.1}$$

where $m = \{m_k\}$ is a given configuration and T is the absolute temperature. The meaning of the product $\prod_{k=i}^{j} m_k$ is to require that two residues i and j (which have $\Delta_{ij} = 1$ and $\varepsilon_{ij} > 0$) contribute to the lowering of the free energy only if they and all the residues between them are native. The presence of this kind of terms makes the WSME model extremely cooperative, allowing a good mimicking of the all–or–none transition typical of protein two–state behavior. The form of equation 4.1 also clearly shows that the WSME model is a one–dimensional model with long–range, many–body interactions.

4.1.1 Exact Solution of the model

This section shows how to obtain the exact solution of the WSME model following the treatment of Bruscolini and Pelizzola [29] that is based on a transfer matrix approach. The equilibrium problem can be stated as the problem of evaluating the partition function

$$\mathcal{Z} = \sum_{m_1=0,1}\sum_{m_2=0,1}\cdots\sum_{m_N=0,1} e^{-\beta H(\{m_k\})} = \sum_{m\,\in\,\{0,1\}^N} e^{-\beta H(m)}\,, \tag{4.2}$$

where $\beta = 1/k_BT$, as a function of the free parameters of the model. The ensemble averages of the most interesting observables are indeed related to manipulations of the total free energy $F = -k_BT\ln\mathcal{Z}$. As it is clear from equation 4.2, $\mathcal{Z}$ is defined by adding together a number of Boltzmann weights which grows exponentially in the number N of residues as 2^N. This makes the calculation of $\mathcal{Z}$ not trivial when N is already in the order of few tens. However the WSME model has remarkable mathematical properties which makes the computation of $\mathcal{Z}$ treatable even at a greater number of degrees of freedom N.

[1] in the following a threshold length of 4 Å will be always assumed.

It is here convenient to introduce the concept of *native stretch* which will be repeatedly used in the rest of this thesis. A native stretch from residue i to residue j is defined as a sequence of native residues delimited by the two non–native residues i and j. Using the quantity

$$S_{ij} = (1 - m_i) \left(\prod_{k=i+1}^{j-1} m_k \right) (1 - m_j) \, , \tag{4.3}$$

which is equal to 1 if the sequence of residues from i to j is a native stretch and is 0 otherwise, and setting the boundary conditions $m_0 = m_{N+1} = 0$ and $q_0 = q_{N+1} = 0$ it is possible to rewrite the Hamiltonian 4.1 as

$$H\left(\{S_{ij}\}\right) = -\sum_{i=0}^{N} \sum_{j=i+1}^{N+1} h_{ij} S_{ij} \, , \tag{4.4}$$

where

$$h_{ij} \equiv -\chi_{ij} - k_B T q_i \, , \tag{4.5}$$

and

$$\chi_{ij} = \sum_{r=i+1}^{j-2} \sum_{s=r+1}^{j-1} \varepsilon_{rs} \Delta_{rs} \, , \tag{4.6}$$

is minus the energy of the native stretch from i–th to j–th residue. The original one–dimensional problem has thus been mapped to a two–dimensional one, where the state of the system is defined by the values of the variables S_{ij} which select the native stretches. This new variables can be associated to the triangular–shaped portion of the square lattice defined by $0 \leq i < j \leq N + 1$.

As shown in figure 4.1 the new variables are not all independent. Once given $S_{ij} = 1$, it follows that $S_{ir} = 0$ for $\forall\, r \neq j$, $S_{rj} = 0$ for $\forall\, r \neq i$, $S_{rs} = 0$ for $\forall\, r, s \,/\, i < r < s < j$. Thus, the condition $S_{ij} = 1$ determines the state of not only row j but also of the k–th row, with $i < k < j$. In this way the feasibility of the transfer matrix approach becomes evident. In fact, looking at figure 4.1, row j can assume only j states according to where it has a filled circle (or a square). Let's define the state of row j with a vector v_k^j, where $k = 0, 1, \ldots, j-1$ points at the position of the filled circle (or the square). In components $(v_k^j)_l = \delta_{l,k+1}$. The transfer matrix from row j to row $j-1$ is defined by its action on vector v_k^j:

$$Q_j^{j-1}(\lambda)\, v_k^j = v_k^{j-1} \qquad \text{for} \quad 0 \leq k < j-1 \, , \tag{4.7}$$

$$Q_j^{j-1}(\lambda)\, v_{j-1}^j = \sum_{k=0}^{j-2} \lambda^{(j-k-2)}\, w_{k,j-1}\, v_k^{j-1} \, , \tag{4.8}$$

where $w_{k,j-1} = \exp\left[-\beta h_{k,j-1}\right]$. In the second equation, each term of the sum indicates that a native stretch from k to $j-1$ is introduced with its Boltzmann weight $w_{k,j-1}$. λ is a dummy variable whose exponent takes into account the number of native residues of the new native stretch. $Q_j^{j-1}(\lambda)$ is a $(j-1) \times j$ matrix with elements

$$\left[Q_j^{j-1}(\lambda) \right]_{lm} = \delta_{lm} + \delta_{mj}\, \lambda^{(m-l-1)} \, \exp\left[-\beta h_{l-1,m-1}\right] \, . \tag{4.9}$$

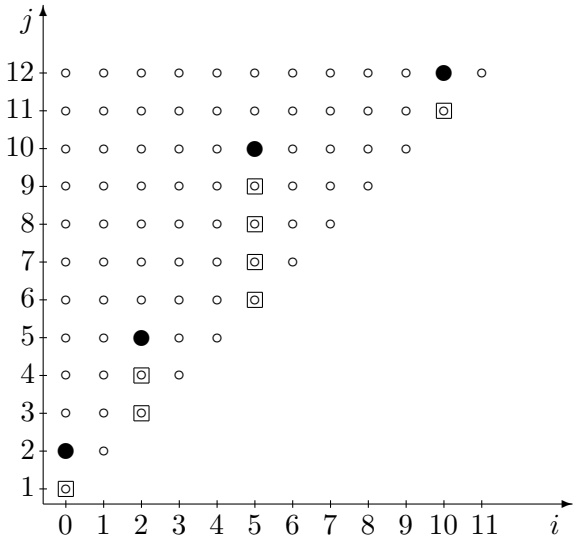


Figure 4.1: Two–dimensional description of the configuration of a protein with $N = 11$ residues and $m_2 = m_5 = m_{10} = 0$ and $m_k = 1$ for the other residues. Filled circles represent $S_{ij} = 1$, empty circles $S_{ij} = 0$. Squares show that a filled circle in line j determines the state of all lines k with $i < k < j$.

Now it is possible to calculate the partition function of the model as $\mathcal{Z} = \mathcal{Z}(\lambda = 1)$, where

$$\mathcal{Z}(\lambda) = Q_2^1(\lambda)\, Q_3^2(\lambda) \ldots Q_{N+2}^{N+1}(\lambda)\, v_{N+1}^{N+2} = \sum_{j=0}^{N} Z_j\, \lambda^j \ . \tag{4.10}$$

The terms Z_j in the above expression are the contribution to the partition function coming from those configurations with a fixed number j of native residues. It is thus possible, from these terms, to obtain the free energy landscape as a function of the number of native residues as $F(j) = -k_B T \ln Z_j$. If we now rewrite λ as e^{μ}, its meaning becomes more clear; in fact μ can be seen as a chemical potential.

4.2 WSME model and mechanical unfolding

In this section it is shown how the WSME model can be modified to deal with the problem of protein mechanical unfolding. To achieve this goal, the new Hamiltonian is defined as the sum of the interaction term of the old Hamiltonian 4.1, and of a potential energy term $V(L)$ which depends on the end–to–end length L of the protein and which takes into account the presence of a pulling external force

$$H = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varepsilon_{ij} \Delta_{ij} \prod_{k=i}^{j} m_k + V(L) \ . \tag{4.11}$$

Thus, a necessary step towards the generalization of the model has to handle with an appropriate description of the end–to–end length. This is done following the idea that, given a particular configuration of the protein, its end–to–end length can be obtained as the vector sum of the lengths of its various native stretches.

Given a native stretch from i–th to j–th residue ($S_{ij} = 1$), its length l_{ij} is obtained from the three–dimensional structure of the native state as the distance between the C_α atoms of residues i and j. Again the boundary conditions $m_0 = m_{N+1} = 0$ are set; if a native stretch ($S_{ij} = 1$) starts with $i = 0$ or (and) ends at $j = N + 1$, one considers the position of the nitrogen at the N–terminus or (and) of the carbon at the C–terminus. In the limiting cases $j = i+1$ and $i = 0$, $j = N+1$ the length l_{ij} represents respectively the distance between two consecutive amino acids and the end–to–end length of the protein. The direction, with respect to the force direction, of the stretch is instead determined by introducing a new unit–vector variable $\vec{\sigma}_{ij}$. The end–to–end length of the molecule reads

$$\vec{L}(\{\vec{\sigma}_{ij}\}) = \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} l_{ij}\, S_{ij}\, \vec{\sigma}_{ij} \ . \tag{4.12}$$

It is easy to verify that the number of native stretches of a given configuration $\{m_k\}$ is equal to 1 plus the number of m_k that are equal to zero (m_0 and m_{N+1} excluded), that is $1 + \sum_{i=1}^{N}(1 - m_i)$.

Coming to the issue of the stretch orientations, in the following, with the aim of keeping the model as simple as possible, only two possibilities are considered: parallel or antiparallel to the direction of the external force. A spin variable σ_{ij} is thus associated to the rigid stretch from i-th to j-th residues, taking the values $+1$ and -1 for the above two preferences respectively.

It is maybe interesting to note, before going on, that in the original version of the model, the binary variables $\{m_k\}$ were not associated with the residues but with the bonds connecting them and the length l_{ij}, for example, was the distance between the midpoint of the C_{i-1} and N_i atoms and the midpoint of the C_j and N_{j+1} atoms. The principal reason for which it is better to associate the binary variable with the residues is that the maximum end–to–end length, which is the length of the completely unfolded, fully extended configuration[2],

$$L_{\max} = \sum_{i=0}^{N} l_{i,i+1} \ , \tag{4.13}$$

matches with the experimental value, while in the case of the original choice it is shorter of about 15–20%. In fact the distance between two consecutive C_α atoms of the chain remains the same for any configuration of the dihedral angles while the distance between the midpoint of the C_{i-1} and N_i atoms and the midpoint of the C_j and N_{j+1} atoms depends on the angles ϕ_i and ψ_i. Thus the right bending point on which to build a polypeptide chain configuration are the C_α atoms.

As discussed in the previous chapter, experiments usally deal with two different protocols to pull a protein: a constant velocity protocol or a constant force one. In the rest of this chapter the latter protocol will be taken into account and it will be shown as in this case it is still possible to solve the model exactly. In this case, the coupling to the external constant force f is expressed through the potential

[2]One could argue about the fact that $L_{\max}$ is the maximum of the end–to–end length of a protein described through this model. However simple considerations involving triangle inequality allow to conclude that this is actually the case.

$V(L) = -f\,L$. Denoting with $m = \{m_k\}$ a given state of the residue variables and with $\sigma = \{\sigma_{ij}\}$ a given state of the spin variables, the Hamiltonian reads

$$H(m,\sigma;f) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\varepsilon_{ij}\Delta_{ij}\prod_{k=i}^{j}m_k - f\sum_{i=0}^{N}\sum_{j=i+1}^{N+1}l_{ij}\,S_{ij}\,\sigma_{ij}\ . \tag{4.14}$$

Since the end–to–end length contributes linearly to the Hamiltonian and the variables σ_{ij} do not interact among themselves, it is possible to obtain an effective Hamiltonian which has the same structure of the Hamiltonian 4.1 of the initial model and therefore the equilibrium thermodynamics is exactly solvable also in this case. In fact one can perform the sum on the σ variables in the partition function

$$\mathcal{Z}(f) = \sum_{m\,\in\,\{0,1\}^N}\ \sum_{\sigma\,\in\,\mathcal{O}(m)} e^{-\beta H(m,\sigma;f)} = \sum_{m\,\in\,\{0,1\}^N} e^{-\beta H_{\rm eff}(m;f)}\ , \tag{4.15}$$

where $\mathcal{O}(m)$ represents the set of the "active" σ_{ij} on a given configuration m, i.e. those σ_{ij} with i and j such that $S_{ij} = 1$, and

$$H_{\rm eff}\,(m;f) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\varepsilon_{ij}\Delta_{ij}\prod_{k=i}^{j}m_k - \frac{1}{\beta}\sum_{i=0}^{N}\sum_{j=i+1}^{N+1}\ln\left[2\cosh\left(\beta f l_{ij}\right)\right]S_{ij}\ . \tag{4.16}$$

Interestingly, in the case of force $f = 0$ the last expression reduces to equation 4.1 with $q_k = \ln 2$ for every k. Here, the particular value of q_k depends on the number of orientations per stretch. The greater entropy of the unfolded state relative to the folded state is indeed encoded in this number.

4.2.1 Exact solution through a recursive algorithm

In principle the generalized version of the WSME model just described could be solved exactly rewriting the Hamiltonian in the form 4.4 with

$$h_{ij} \equiv -\chi_{ij} - f\,l_{ij}\,\sigma_{ij}\ , \tag{4.17}$$

and substituting the Boltzmann weight $w_{k,j-1}$ in equation 4.8 with

$$w_{k,j-1} = 2\,\cosh\left(\beta f l_{k,j-1}\right)\,e^{\beta\chi_{k,j-1}}\ . \tag{4.18}$$

But here an alternative and computationally more efficient way to solve the equilibrium thermodynamics of the model is proposed [30,86]. The method is based on an efficient recursive algorithm. To this purpose, let us start to define the end–to–end length L_n up to the first $n \le N$ residues

$$L_n = \sum_{i=0}^{n}\sum_{j=i+1}^{n+1} l_{i,j}\,\sigma_{i,j}\,S_{i,j}\ , \tag{4.19}$$

where it is assumed $m_{n+1} = 0$, and the partial Hamiltonian $H_n = E_n - f\,L_n$, where the term E_n represents the interaction energy up to the first n residues

$$E_n = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\varepsilon_{i,j}\Delta_{i,j}\prod_{k=i}^{j}m_k\ . \tag{4.20}$$

Denoting with $\mathcal{O}_n(m)$ the set of the "active" σ_{ij} on a configuration m of the first n residues, the corresponding partition function is

$$Z_n = \sum_{m \in \{0,1\}^n} \sum_{\sigma \in \mathcal{O}_n(m)} e^{-\beta E_n + \beta f L_n} \ . \tag{4.21}$$

Let us also define the auxiliary quantities

$$A_n^i = \sum_{m \in \{0,1\}^n} \sum_{\sigma \in \mathcal{O}_n(m)} \left[(1 - m_{i-1}) \prod_{k=i}^{n} m_k \right] e^{-\beta E_n + \beta f L_n} \ , \tag{4.22}$$

with $0 \leq n \leq N$ and $1 \leq i \leq n+1$. It is easy to check that for $i = n+1$:

$$\begin{aligned} A_n^{n+1} &= \sum_{m \in \{0,1\}^n} \sum_{\sigma \in \mathcal{O}_n(m)} (1 - m_n)\, e^{-\beta E_n + \beta f L_n} = \\ &= 2\cosh(\beta f\, l_{n,n+1}) \sum_{m \in \{0,1\}^{n-1}} \sum_{\sigma \in \mathcal{O}_{n-1}(m)} e^{-\beta E_{n-1} + \beta f L_{n-1}} = \\ &= 2\cosh(\beta f\, l_{n,n+1})\, Z_{n-1} \ , \end{aligned}$$

and that for $1 \leq i \leq n$:

$$\begin{aligned} A_n^i &= \sum_{m \in \{0,1\}^n} \sum_{\sigma \in \mathcal{O}_n(m)} (1 - m_{i-1}) m_i \dots m_n\, e^{-\beta E_n + \beta f L_n} = \\ &= 2\cosh(\beta f\, l_{i-1,n+1}) \exp[\beta\, \chi_{i-1,n+1}] \sum_{m \in \{0,1\}^{i-2}} \sum_{\sigma \in \mathcal{O}_{i-2}(m)} e^{-\beta E_{i-2} + \beta f L_{i-2}} = \\ &= 2\cosh(\beta f\, l_{i-1,n+1}) \exp[\beta\, \chi_{i-1,n+1}]\ Z_{i-2} \ . \end{aligned}$$

Using the telescopic identity $\sum_{i=1}^{n+1} (1 - m_{i-1}) \prod_{k=i}^{n} m_k = 1$ it is then possible to show that $Z_n = \sum_{i=1}^{n+1} A_n^i$, which, together with the previous expressions and the initial condition $Z_{-1} = 1$, gives the recursive algorithm

$$\begin{cases} A_n^i = 2\cosh(\beta f\, l_{i-1,n+1})\ e^{\beta\, \chi_{i-1,n+1}}\ Z_{i-2} \ , \\ \\ Z_n = \sum_{i=1}^{n+1} A_n^i \ . \end{cases} \tag{4.23}$$

This make possible to compute easily the total partition function $\mathcal{Z}(f) = Z_N$ even for long proteins, involving a number of operations which grows polynomially in N, precisely as N^2.

4.2.2 Order parameters and power expansions

By suitably manipulating the recursive scheme 4.23 it is moreover possible to calculate averaging observables. As an example, let us consider the problem of computing the average fraction of native residues $\langle M \rangle$ with $M = \frac{1}{N} \sum_{i=1}^{N} m_i$. The

following relation holds:

$$\begin{aligned}\langle M \rangle &= \frac{1}{\mathcal{Z}} \sum_{\{conf.\}} \left(\frac{1}{N} \sum_{i=1}^{N} m_i \right) e^{-\beta H} = \\ &= \frac{1}{N} \frac{1}{\mathcal{Z}} \frac{\partial}{\partial \lambda} \left(\sum_{\{conf.\}} e^{-\beta H + \lambda \sum_{i=1}^{N} m_i} \right) \Bigg|_{\lambda=0} = \\ &= \frac{1}{N} \frac{1}{\mathcal{Z}(\lambda=0)} \left(\frac{\partial}{\partial \lambda} \mathcal{Z}(\lambda) \right) \Bigg|_{\lambda=0} . \end{aligned} \tag{4.24}$$

To compute the partition function $\mathcal{Z}(\lambda)$ it is then sufficient to opportunely modify equations 4.21 and 4.22 to obtain the new recursive scheme

$$\begin{cases} A_n^i(\lambda) = 2\cosh\left(\beta f \, l_{i-1,n+1}\right) \, e^{\beta \chi_{i-1,n+1} + \lambda(n-i+1)} \, Z_{i-2}(\lambda) \, , \\ \\ Z_n(\lambda) = \sum_{i=1}^{n+1} A_n^i(\lambda) \, . \end{cases} \tag{4.25}$$

In general, the average value of an observable is achievable each time it is possible to express the observable as a quantity that, as the interaction energy E_n or the end–to–end length L_n, can be built adding the residues step by step.

Furthermore, if the considered quantity can assume only a finite set of values, by expanding rather than deriving with respect to λ, it is possible to prepare a scheme that allows to express the free energy landscape as a function of a reaction coordinate represented by the quantity itself. Here the case of the free energy landscape as a function of the end–to–end length ($\lambda = \beta f$) is presented. In fact, given the finite resolution of the atomic coordinates of the structure deposited in the Protein Data Bank (which is 10^{-3} Å), it is fair to round the distances l_{ij} to the same resolution. Thus the number of end–to–end length possible values is finite and it is allowed to do the following power expansions:

$$\mathcal{Z}(f) = \sum_{L=-L_{\max}}^{L_{\max}} z(L) \, e^{\beta f L} \, , \tag{4.26}$$

$$Z_n(f) = \sum_{L=-L_{n,\max}}^{L_{n,\max}} z_n(L) \, e^{\beta f L} \, , \tag{4.27}$$

$$A_n^i(f) = \sum_{L=-L^\star_{n,\max}}^{L^\star_{n,\max}} a_n^i(L) \, e^{\beta f L} \, , \tag{4.28}$$

where $L_{\max}$ is given by equation 4.13, $L_{n,\max}$ is the maximal length up to n residues and $L^\star_{n,\max}$ is the maximal length up to n residues but subjected to the same constrains of A_n^i, i.e. $L^\star_{n,\max} = L_{i-2,\max} + l_{i-1,n+1}$. The importance of the finite resolution of the microscopic lengths l_{ij} is understood by observing that the values that the end–to–end length can assume are not known *a priori* and thus the algorithm has to span all the possible values in the interval $[-L_{\max}, L_{\max}]$. The

set of these values has thus to be finite. Substituting the above expressions in the recursive scheme 4.23, it follows:

$$\begin{cases} a_n^i(L) = e^{\beta\,\chi_{i-1,n+1}}\ [z_{i-2}(L - l_{i-1,n+1}) + z_{i-2}(L + l_{i-1,n+1})]\ , \\ \\ z_n(L) = \sum_{i=1}^{n+1} a_n^i(L)\ , \end{cases} \tag{4.29}$$

with initial conditions $z_{-1}(L) = 1$ for $L = 0$ and $z_{-1}(L) = 0$ for $L \neq 0$. The final result of scheme 4.29, $z_N(L)$, corresponds to the constrained zero–force partition function $\mathcal{Z}(L; f = 0)$, L being a given value of the protein end–to–end length,

$$\mathcal{Z}(L; f = 0) = \sum_{m \in \{0,1\}^N} \sum_{\sigma \in \mathcal{O}(m)} \delta_{L,L(m,\sigma)}\, e^{-\beta H(m,\sigma;f=0)} = z_N(L)\ , \tag{4.30}$$

where $\delta_{L,L(m,\sigma)}$ is the Kronecker delta selecting those configurations with end–to–end length equal to L. The corresponding free energy landscape at zero–force is given by $F(L) = -k_B T \ln \mathcal{Z}(L; f = 0)$. Being the potential energy $V(L)$ a function of L, the total free energy of the model as a function of the end–to–end length turns out to be

$$G(L) = -k_B T \ln \mathcal{Z}(L; f = 0) + V(L)\ . \tag{4.31}$$

The validity of the above relation holds not only for the constant force case but also in case of more complicated potentials $V(L)$. Furthemore, the knowledge of $G(L)$ allows also to compute the ensemble average of a generic observable g that is a function of the end–to–end length, $g = g(L)$

$$\langle g(L) \rangle = \frac{\sum_{L=-L_{\max}}^{L_{\max}} g(L)\, e^{-\beta G(L)}}{\sum_{L=-L_{\max}}^{L_{\max}} e^{-\beta G(L)}}\ . \tag{4.32}$$

As a final remark, let us stress again that the two techniques just described are general (not strictly connected to the fraction of native contacts or to the end–to–end length) and can be combined. Let us for example consider two quantities R and P that are functions of the molecule configuration (m, σ) and can have only a finite set of values. Let's also assume that R and P could be built step by step in the discussed recursive scheme, i.e. that the quantities R_n and P_n exist for each n, $0 \leq n \leq N + 1$. Introducing two coupling parameters λ and μ one can modify the Hamiltonian $H \to H + \lambda R + \mu P$ and write a recursive scheme similar to 4.25. At this point, one can decide to power expand with respect of the parameter μ, thus obtaining the partition function $Z(\lambda, P)$, and then to derive Z with respect to λ, thus obtaining the ensemble average of R conditional to a particular value of P. Alternatively it is possible to power expand with respect to both λ and μ; the resulting partition function $Z(R, P)$ allows to get the two–dimensional free energy landscape in function of the reaction coordinates R and P.

4.2.3 Kinetics of the model

In a typical pulling experiment a controlled force is applied to one of the free ends of a protein and the induced elongation is measured. As described in section 3.2, the

common interpretation is that the so caused unfolding of the molecule, is hindered by kinetic barriers associated with the strongest linkages, which serve to stabilize the molecular structure. The breaking of a contact can thus be viewed as the overcoming of such a barrier [72, 96]. In order to discuss this scenario, after the equilibrium behavior, the kinetics of the model must be considered.

The nonequilibrium unfolding kinetics can be studied by Monte Carlo (MC) simulations. The MC method is basically a prescription to draw an ergodic trajectory into the space of the possible configurations. In stationary conditions, this trajectory visits each configuration with a frequency proportional to its Boltzmann weight, as required, in equilibrium conditions, in the canonical ensemble. One of the most famous way to obtain such a trajectory is given by the so called Metropolis algorithm, which here is adapted to the WSME model for mechanical unfolding. The applications to real proteins that will be considered in the next chapter, involve a kinetics in which a single–residue flip concerning the variables m is followed by a single–spin flip on the variables σ.

Following the the main ideas in reference [97], let us start by considering that the configuration (m, σ) goes in the configuration (m', σ'): $[(m, \sigma) \to (m', \sigma')]$ by flipping the k–th residue, with $k = 1, \ldots, N$. Recalling the quantity S_{ij} (equation 4.3), with $i < k < j$, it is easy to check that, for any choice of the state $m \in \{0, 1\}^N$, there exists a unique pair (i, j) of indices such that $S_{ij} = 1$ if $m_k = 1$ and $S_{ik} = S_{kj} = 1$ if $m_k = 0$. We assume that the flip of the k–th residue is allowed only if in the original configuration $m_k = 0$ and $\sigma_{ik} = \sigma_{kj}$ (going into a configuration with $m'_k = 1$, $\sigma'_{ij} = \sigma_{ik}$ and $\sigma'_{ik} = \sigma'_{kj} = 0$[3]), else if $m_k = 1$ (going to $m'_k = 0$, $\sigma'_{ik} = \sigma'_{kj} = \sigma_{ij}$ and $\sigma'_{ij} = 0$). For what follows it is useful to define the quantity $\Theta(m, \sigma; m', \sigma') = 1$ if the above conditions hold and $\Theta(m, \sigma; m', \sigma') = 0$ otherwise. Furthermore, following Metropolis prescription, (m', σ') will follow (m, σ) in the trajectory only if it is verified one of the two following circumstances:

- the energy of the new configuration is lower than or equal to the energy of the original configuration, $H(m', \sigma') \leq H(m, \sigma)$,
- $H(m', \sigma') > H(m, \sigma)$ and a random number x, generated with uniform probability in the domain $[0, 1]$, is such that $x \leq \exp\{-\beta[H(m', \sigma') - H(m, \sigma)]\}$.

Then we proceed with a spin flip, $(m', \sigma') \to (m', \sigma'')$, by choosing a variable σ'_{ij} with uniform probability among the $1 + N(1 - M)$ stretch orientational variables, setting $\sigma''_{ij} = -\sigma'_{ij}$ and accepting this spin flip if it is verified one of the two following circumstances:

- $H(m', \sigma'') \leq H(m', \sigma')$,
- $H(m', \sigma'') > H(m', \sigma')$ and a random number y, generated with uniform probability in the domain $[0, 1]$, is such that $y \leq \exp\{-\beta[H(m', \sigma'') - H(m', \sigma')]\}$.

Repeating this procedure, we generate a sequence of configurations which constitutes the Metropolis trajectory. Such a trajectory, also called Markov chain, is a stochastic process since its evolution depends on the choices of the residue and of

[3] With a slight abuse of notation we have set equal to zero the variables σ_{ij} which are not "active", i.e. do not belong to the set $\mathcal{O}(m)$, in a given configuration.

the native stretch to flip and on the random number x and y. Each MC step is thus defined by a stocastic matrix $W = W_1 W_2$ with W_1, associated with the residue flip, equal to

$$W_1\left((m,\sigma)\to(m',\sigma')\right) = \begin{cases} \dfrac{1}{N} & \text{if } H(m',\sigma') \le H(m,\sigma)\ , \\ \dfrac{1}{N} e^{-\beta[H(m',\sigma')-H(m,\sigma)]} & \text{if } H(m',\sigma') > H(m,\sigma)\ , \end{cases} \tag{4.33}$$

if $\Theta(m,\sigma;m',\sigma') = 1$ and $W_1\left((m,\sigma)\to(m',\sigma')\right) = 0$ otherwise. W_2, associated with the spin flip, is instead equal to

$$W_2\left((m,\sigma)\to(m,\sigma')\right) = \begin{cases} [1+N(1-M)]^{-1} & \text{if } H(m,\sigma') \le H(m,\sigma)\ , \\ \dfrac{e^{-\beta[H(m,\sigma')-H(m,\sigma)]}}{1+N(1-M)} & \text{if } H(m,\sigma') > H(m,\sigma)\ , \end{cases} \tag{4.34}$$

Since we followed the Metropolis prescription, the Boltzmann distribution corresponding to the Hamiltonian H is in detailed balance with these matrices and thus also with W, hence it is an invariant distribution of W.

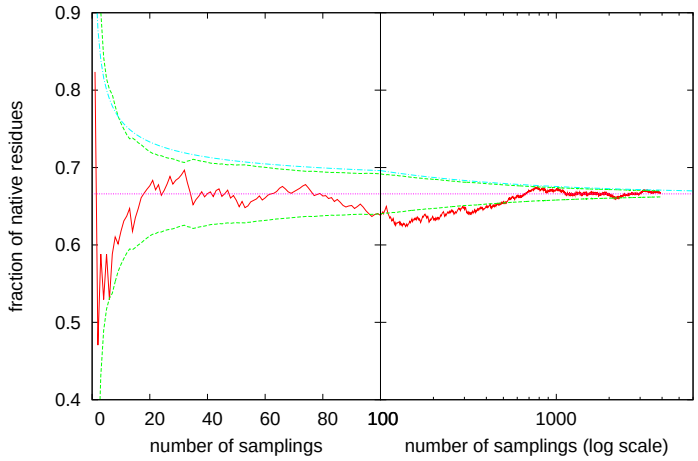


Figure 4.2: Monte Carlo estimate of the fraction of native residues as a function of the number of samplings of different configuration of the protein at the unfolding temperature (red line). Purple line represents the exact value. Green lines are an estimate of standard deviation around the exact value. Cyan line is the expected scaling $\sim 1/\sqrt{n}$ of the standard deviation.

Besides to use the MC sampling to study the kinetic behavior in non–equilibrium conditions, as will be done in the next chapters, MC can also be use to study the equilibrium behavior. Figure 4.2 reports a MC estimation of the fraction of native residues of the final hairpin of protein G, as a function of the number of samplings n of configurations at $f = 0$ and at the unfolding temperature (defined as the temperature at which $\langle M\rangle = 2/3$ and obtained with the exact equilibrium solution). To ensure measures independence, each sampling comes from a different

MC trajectory. In each trajectory the simulation starts with the protein in the completely folded state and $5 \cdot 10^6$ MC step are performed before the sampling. It is possible to note that the estimate of $\langle M \rangle$ tends to the exact value in the great n limit and that the standard deviation follows the expected scaling law $\sim 1/\sqrt{n}$ [98].

A final remark is due. A time–dependent Hamiltonian H_t, as in the case of a constant velocity protocol, in the above expressions results in a time–dependent stochastic matrix W_t and hence in a non–homogeneous Markov process. The explicit presence of the time however does not change the fact that, if π_t is the Boltzmann distribution related to H_t, then π_t is invariant for W_t.

Chapter 5

WSME model and mechanical unfolding of proteins

In recent years, the mechanical properties of biopolymers under mechanical loading have been the subject of an intense research activity, both experimental and theoretical, in the last two decades. For a recent review, see [68]. Innovative single molecule experimental techniques, mainly based on atomic force microscopy (AFM) and optical tweezers [99, 100, 66, 62, 101], have been used to investigate the response of biopolymers to controlled forces, while theoretical and computational models at different levels of coarse graining have been proposed and investigated [79, 30, 86, 102, 87, 103, 104]. These works have both helped to understand experimental results and, on the other hand, they also studied the molecules under conditions otherwise not accessible to the experimental techniques.

In such a context, the WSME model has already been shown [30,86] to reproduce the general features of mechanical unfolding experiments, like the force dependence of the average unfolding time in a constant force protocol, or the rate dependence of the unfolding force in a constant rate protocol, together with the corresponding probability distributions. The same model turned out to predict the correct values for the unfolding lengths of a titin domain [30,86] and of ubiquitin [87]. Moreover, it has been used to investigate the unfolding pathways of ubiquitin [87] and of a 236–base RNA fragment [88], and the resulting pathways turned out to be consistent with both experimental and computational results, where more detailed molecular models were used.

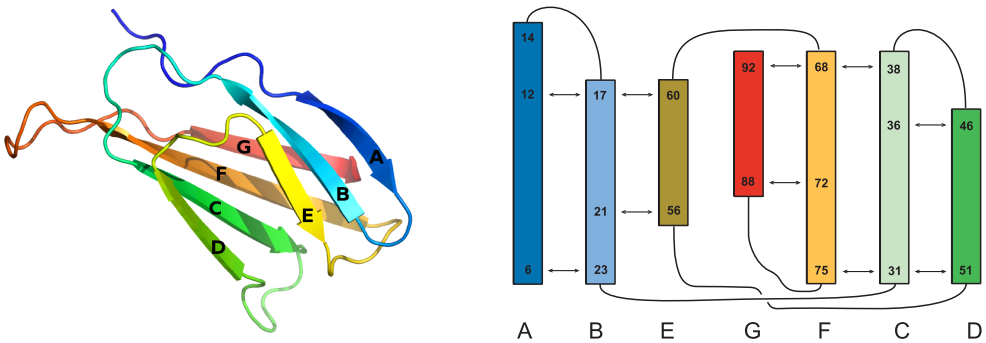
This chapter is devoted to the study of mechanical unfolding of two real proteins, the tenth type III domain from fibronectin and the green fluorescent protein, using the WSME model. Mechanical properties and unfolding pathways of model proteins, studied both at constant force and at constant pulling velocity, will be presented and compared with experimental results and previous simulations. It will be shown that, notwithstanding its simplicity, this model can capture important details, such as rupture forces and intermediate states, that are consistent with experiments and simulations based on more detailed models. An advantage of this model is that it is not computationally demanding and thus it permits to investigate the unfolding behavior at pulling velocities and forces which are comparable to those used in experiments, while, usually, more detailed models have to consider velocities and

forces that are orders of magnitude greater. The same reason makes possible to generate a large set of unfolding events, which is important when studying a system with multiple unfolding pathways.

5.1 A Fibronectin domain

Among the various molecules studied, fibronectin is a particularly important one, due to its role in mediating a wide variety of cellular interactions with the extracellular matrix and in playing important functions in tissue elasticity, cell adhesion, migration, growth and differentiation [105,106]. It is important for processes such as wound healing and embryonic development and altered fibronectin expression and degradation has been associated with a number of pathologies, including cancer and fibrosis [107]. Fibronectin is a giant multimodular protein which usually exists as a dimer composed of two nearly identical subunits linked covalently at their C–termini by disulfide bonds. Each monomer consists of three types of repeating units called FnI, II and III. Approximately 90% of fibronectin monomer sequence is composed by 12 type I repeats, two type II repeats and 15–17 type III repeats.

In the following we will focus on the 10th type III module (FnIII_{10}) which is known to be crucial for cell adhesion through the binding of its RGD motif (residues Arg78–Gly79–Asp80) to transmembrane integrin receptors [108]. In fact it has been proposed that a stretching force may influence the adhesion properties by causing full or partial unfolding of the FnIII_{10} module, and thereby deformation of the RGD motif [109]. The secondary structure of FnIII_{10} (figure 5.1) consists of two antiparallel β–sheets forming a β–sandwich. The β–strands are usually denoted with letters from A (the strand closest to the N terminal) to G (the C terminal one). The two sheets are made of strands ABE and DCFG, respectively, and the RGD motif is in the loop separating strands F and G.



(a) (b)

Figure 5.1: Native structure of FnIII_{10} (pdb ID 1ttf) with β–strands labeled A–G in sequence order. (a) Sketch generated by PyMOL. (b) Order of β–strands with residues belonging to them and residue–residue hydrogen bonding.

The mechanical unfolding of FnIII_{10} has been studied both experimentally [110, 67] and by computer simulations [109, 111, 112, 113, 104]. Single AFM experiments

have shown that $FnIII_{10}$, with a rupture force of about 80 pN, has, together with $FnIII_{10}$, a low mechanical stability compared to other fibronectin type III domains though it is significantly more thermostable than other domais which on the contrary have a great mechanical stability, such as for example $FnIII_1$ [110]. Furthermore, AFM experiments showed that $FnIII_{10}$ can unfold according to different pathways [67]. Apparent two–state transitions were observed, as well as unfolding through intermediate states. Experiments on suitable mutants suggested the possible existence of different intermediate states [67], which is also consistent with some simulations.

Paci and Karplus [111], using steered molecular dynamics simulations, found two unfolding pathways, both proceeding through partially unfolded intermediate states lacking two of the seven native β–strands. The missing strands were A and B in one case, and A and G in the other. In a more recent study, using steered molecular dynamics, Gao *et al.* [113] found three different pathways. In this study it has been shown that strand A may separate first, later followed by detaching of strand B and finally by complete unfolding. Alternatively the unfolding may proceed passing through an intermediate with strands A and G detached or it can visit an intermediate state in which only strand G unravels before complete unfolding. Other simulations predicted simpler [109, 112] or more complex [104] scenarios. In particular, Mitternacht *et al.* [104], using an implicit–water all–atom model and Monte Carlo simulations, found that five different intermediate states and many unfolding pathways, which visit these intermediates, are possible. Two intermediates lack only one β–strand from the native structure, this being either strand A or strand G, while the other three lack two strands: A and B, F and G or A and G.

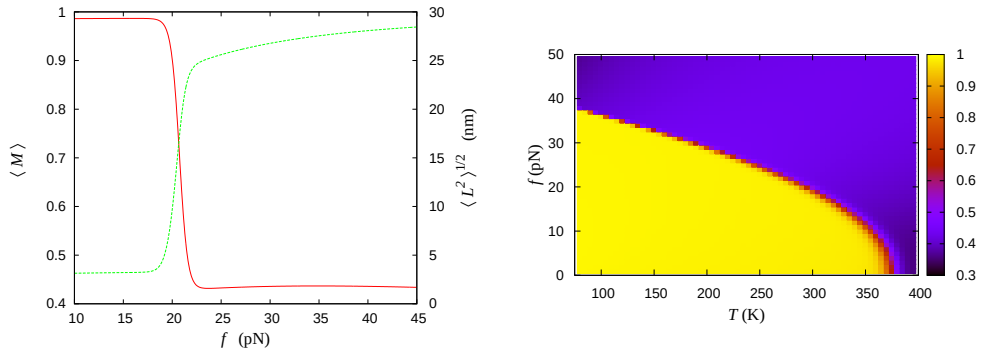
In the case of $FnIII_{10}$, particular interest has been put in exploring the biologically relevant low force regime [105, 114, 115], which is thought to be close to the equilibrium unfolding force and cannot be explored by simulations of more detailed, and more computationally expensive, molecular models. Our model can probe forces close to the equilibrium unfolding force, whose value we use to set our force unit. Such value is unfortunately not exactly known. Erickson [114] estimates the equilibrium unfolding force to be at most 5 pN, on the basis of an order of magnitude calculation. On the other hand, an estimate close to 20 pN was reported in [104].

In the rest of this section results obtained through the modified WSME model of section 4.2 will be presented but the reader must keep in mind that in this section we have assumed that the binary variables $\{m_k\}$ are not associated with the residues but with the bonds connecting them (see discussion in section 4.2). Having chosen the value of 20 pN, in order to set the force unit, we have obtained results that give unfolding forces in very good agreement with the AFM experiments.

5.1.1 Equilibrium properties

Equilibrium studies and Monte Carlo simulations have been done setting the temperature to $T = 0.768\,T_m$, where T_m is the equilibrium unfolding temperature at zero force. Since experimentally $T_m = 375$ K [116], we have $T = 288$ K. The force unit is then set in such a way that the equilibrium typical unfolding force at

$T = 288$ K is 20 pN. Since an experimental measurement of this quantity is missing, it has been chosen on the basis of the estimates reported in reference [104].



(a) (b)

Figure 5.2: (a) Average fraction of native bonds $\langle M \rangle$ (red line) and end–to–end length $\sqrt{\langle L^2 \rangle}$ (green line) as a function of the pulling force at temperature $T = 288$ K. (b) Average fraction of native bonds $\langle M \rangle$ as a function of the temperature T and the pulling force f.

As mentioned before, the equilibrium thermodynamics can be solved exactly in our model. Thus it is possible to follow the macroscopic state behaviour of the protein at different pulling forces. In figure 5.2a the average fraction of native bonds $\langle M \rangle$ and end–to–end length $\sqrt{\langle L^2 \rangle}$ are plotted as functions of the force f. This definition of the length is appropriate to describe the system also in the small force regime, where instead $\langle L \rangle$ vanishes due to the fluctuations of the molecule length between positive and negative values. In fact, at zero force, a given configuration of the variables m_i with end–to–end length module equal to L has an equal probability to have end–to–end length L or $-L$. The introduction of the force breaks this $\mathbb{Z}_2$ simmetry. When the force increases a quite sharp transition to an elongated state occurs, showing that the global minimum of the free energy landscape corresponds to either the native state or to the unfolded one. It is moreover possible to obtain a phase diagram by computing the value of the average fraction of native bonds as a function of both the temperature and the force (see figure 5.2b).

As shown in section 4.2.2, expanding the partition function $\mathcal{Z}(f)$ in powers of $e^{\beta f}$, one obtains the equilibrium energy landscape of the protein as a function of the reaction coordinate L. In presence of a constant force, the free energy landscape is tilted and is given by $G(L) = -k_B T \ln \mathcal{Z}(L; f = 0) - fL$. Figure 5.3 shows also the landscape for various forces: at zero force there is just one minimum at about 3.5 nm corresponding to the folded state. By increasing the force three more minima appear: two of them (end–to–end lengths of about 6 and 13 nm) are always local minima, and will be later associated to intermediate states, while the third one, corresponding to the fully unfolded state, becomes the global minimum when the force exceeds 20 pN. It is worth to note how the free energy landscape of figure 5.3 reproduces well the features of the landscape obtained by Mitternacht *et al.* [104] using the extended Jarzynski equality [64, 117].

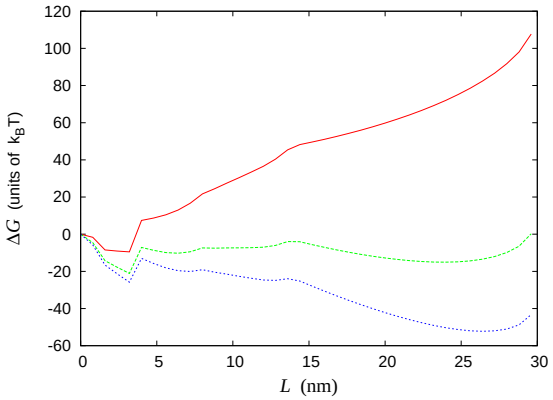


Figure 5.3: Free energy landscape at $T = 288$ K and for forces $f = 0$ pN (red line), $f = 20$ pN (green line) and $f = 28$ pN (blue line). $\Delta G = G(L) - G(0)$.

5.1.2 Unfolding pathways

The nonequilibrium unfolding kinetics have been studied by Monte Carlo (MC) simulations. More precisely, in the framework of a master equation approach [82], we choose transition rates according to the Metropolis algorithm. Rigorously speaking, this choice cannot be derived from an underlying microscopic dynamics of the molecule. Nevertheless, it has been shown [30, 86, 87, 88] that it reproduces many quantitative and qualitative aspects of folding and unfolding of real molecules under an external force. A single MC step consists of a single–bond flip on the variable m_j, chosen with equal probability among the N peptide bond variables, followed by a single–spin flip on the variable σ_{ij}, also chosen with uniform probability among the $1 + N(1 - M)$ stretch orientational variables (see section 4.2.3 for a detailed discussion).

Simulations have been run with nine values of the force (122 pN, 98 pN, 81 pN, 65 pN, 53 pN, 46 pN, 40 pN, 36 pN, 28 pN) and six constant pulling velocities (0.03 μm/s, 0.05 μm/s, 0.1 μm/s, 0.3 μm/s, 0.5 μm/s, 1 μm/s). The time unit will be specified later. For each value of the force or of the velocity 100 different unfolding trajectories have been considered.

In order to trace unfolding pathways, the weighted fraction of native contacts has been used as order parameter:

$$\varphi_s = \frac{\sum_{i=r_1(s)}^{r_2(s)-2} \sum_{j=i+1}^{r_2(s)-1} \varepsilon_{ij} \Delta_{ij} \prod_{k=i}^{j} m_k}{\sum_{i=r_1(s)}^{r_2(s)-1} \sum_{j=i+1}^{r_2(s)} \varepsilon_{ij} \Delta_{ij}} \,, \tag{5.1}$$

where s is the string of bonds we are analyzing and $r_1(s)$, $r_2(s)$ its first and last peptide units. As an example the string containing strands A and B has $r_1(AB) = 6$ and $r_2(AB) = 23$. A straightforward generalization is necessary for order parameters of strings of non-consecutive strands (i.e. C-F and B-E). φ_s turns out to be a better order parameter than the fraction of native bonds used in previous works [30, 86, 87] because of its greater stability with respect to fluctuations. When discussing the folded or unfolded character of an individual β–strand, appropriate order parameters can be identified on the basis of the secondary structure. As an

example, strand F appears in a β–sheet between strands C and G, which suggests to use φ_{CF} and φ_{FG} as order parameters for strand F.

Force Clamp

In the force clamp protocol ($H = H_0 - f\,L$, where f and L are the external pulling force and the molecule end–to–end length and H_0 is the Hamiltonian at zero force) the molecule is first equilibrated in absence of force, then at time $t = 0$ the force instantaneously jumps to a non–vanishing constant value, which ranges between 28 and 122 pN. It is again worth to note that the forces used are much closer to the equilibrium unfolding force, and hence to *in vivo* conditions, than most previous works. In fact, the smallest force probed by Karplus and Paci [111] was 69 pN, and they did not observe any unfolding event at this force, while Gao *et al.* [113] used forces not smaller than 400 pN. Only in the all–atom Monte Carlo simulations by Mitternacht *et al.* [104] unfolding events at constant forces as small as 50 pN could be observed.

The unfolding trajectories that have been found, can be grouped in four classes according to their main features, i.e. their end–to–end length plateaus (if they exist) and the order parameters behaviour for the whole molecule and its various pairs of β–strands. At large forces we observe simple 2–state trajectories, while at smaller forces various intermediates are obtained. A scheme of the possible pathways is shown in figure 5.4.

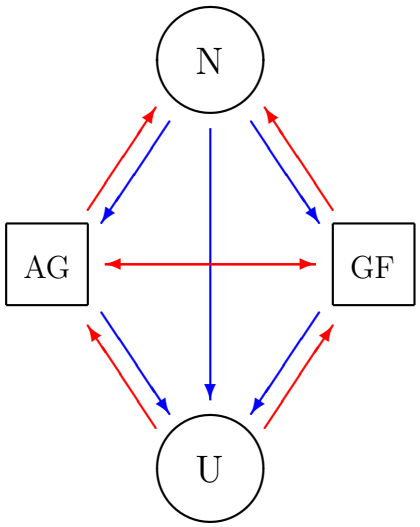
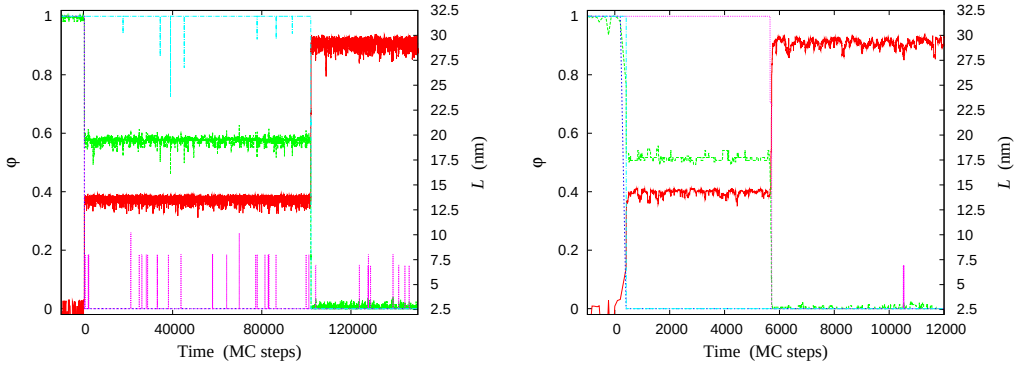


Figure 5.4: Unfolding pathways scheme of FnIII_{10} pulled by a constant force. Transitions denoted by red arrows have been observed only at low forces (40, 36 and 28 pN). Oblique red arrows represent refolding transitions.

In trajectories exhibiting intermediate states it turns out, as already pointed out in previous papers [109, 112], that strand G is always the first to break away. In cellular environment such behaviour seems to be connected to the function of the RGD motif Arg78-Gly79-Asp80 [109, 113]. When the module is fully folded, the RGD motif is available for adhesion, while if strand G is pulled and detached from the remainder of the module, the RGD motif gets closer to the surface of the module and is not functional.

Strand G detachment may be rapidly followed by complete unfolding or by an intermediate state. A possibility is that strand A detaches almost at the same time of strand G, while the remaining part of the molecule stays folded for a certain time before complete unfolding. This kind of unfolding pathway will be labeled with *AG*, its intermediate end–to–end length is about 13.5 nm. It may happen that instead of strand A, strand F detaches together with G, such unfolding pathway (intermediate end–to–end length $\sim$ 14 nm) will be labeled *GF*. Other possibilities occur only in

the biologically relevant regime of low forces. It is believed [104] that such relevant forces, *in vivo*, are of the same order of magnitude as the equilibrium unfolding force (~ 20 pN), though forces as low as 5 pN have been suggested [114] as typical unfolding forces. The low force unfolding pathway is a mixture of the previous two: strands A and G are the first to unfold, then, before the molecule completely unfolds, A refolds and F unfolds. This may happen reversibly many times in a single trajectory with consecutive folding (unfolding) of strand A and parallel unfolding (folding) of strand F. Such trajectories will be labeled *mixed* AG–GF because the molecule is fluctuating between two different intermediates (AG and GF). These intermediates have almost the same end–to–end length, and therefore cannot be distinguished in a simple free energy landscape, as illustrated in figure 5.2, where a single, broad minimum is observed at $L \simeq 13$ nm.



(a) *Unfolding pathway*: AG

(b) *Unfolding pathway*: GF

(c) *Unfolding pathway*: no intermediates

(d) *Unfolding pathway*: mixed AG–GF

Figure 5.5: Typical MC trajectories: end–to–end length (red line) and a few order parameters as functions of time, with $f = 65$ pN (figures a, b, c) and $f = 28$ pN (d). Green line: weighted fraction of native contacts, whole FnIII_{10}. Blue line: weighted fraction of native contacts between strands G and F. Purple line: weighted fraction of native contacts between strands A and B. Cyan line: weighted fraction of native contacts between strands C and F.

Figure 5.5 shows four typical trajectories, three at 65 pN constant force (trajectories a, b and c) and one (trajectory d) in the low force regime at 28 pN constant

Table 5.1: Relative frequencies of unfolding pathways at constant force. 100 trajectories for each value of the force.

	AG	GF	No intermediates	Mixed AG–GF
28 pN	0	0	0	1
36 pN	0.07	0	0	0.93
40 pN	0.96	0	0	0.04
46 pN	0.93	0.07	0	0
53 pN	0.67	0.32	0.01	0
65 pN	0.59	0.31	0.1	0
81 pN	0.41	0.34	0.25	0
98 pN	0.43	0.23	0.34	0
122 pN	0.2	0.06	0.74	0

force. Each simulated trajectory stops 10^5 MC steps after the protein reaches the threshold value $L_{\rm u} = \frac{1}{2}L_{\rm max}$. An exception is the case $f = 28$ pN where we take $L_{\rm u} = \frac{2}{3}L_{\rm max}$, because of the larger length fluctuations and in order to prevent the trajectory ending before a complete unfolding event takes place. During thermalization, before turning the force on at time $t = 0$, the length of the polypeptide chain fluctuates around $L = 0$, since different orientations of the molecule are equally likely. Then, at time $t = 0$, a "*waiting phase*" starts, which can be easily seen in figure 5.5d. This waiting phase corresponds to a metastable state which is characterized by an end–to–end length ~ 3.5 nm corresponding to the elongation in the native state. The order parameters which have not been plotted go to zero only when the protein reaches the fully elongated configuration (end–to–end length ~ 29 nm). The first rise in the end–to–end length to the intermediate value is always associated to the drop in order parameters connected to at least two different pairs of strands, with the pair made by β–strands G and F always involved. The other pairs involved in first unfolding event can be the pair of strands A and B (in this case the molecule goes in the AG intemediate, figure 5.5a) or the pair of strands C and F (the molecule goes in the GF intemediate, figure 5.5b). Alternatively, all the order parameters drop to zero almost simulaneously and the molecule completely unfolds without any detectable intermediate state (figure 5.5c). Figure 5.5d reports a *mixed* AG–GF trajectory obtained at force $f = 28$ pN, slightly larger than the equilibrium unfolding force: after the long waiting phase there are two different intermediate states before the complete unfolding. Despite the large fluctuations in the order parameters associated to the pairs C–F and A–B, it is still possible to see their general behaviour and to recognize the first intermediate state as AG and the second as GF. We stress that the GF and AG intermediates have very similar end–to–end length and fraction of native contacts φ_s (for the whole chain), making them indistinguishable in simple, one–dimensional, free energy landscapes: indeed, they are lumped together in the broad minimum at $L \simeq 13$ nm in the landscape of figure 5.2.

Table 5.1 shows the frequencies of various unfolding pathways. Predictably, as the force increases, the trajectories without any intermediate state become dom-

inant and we expect them to be the only escape route at even higher forces, as already observed in previous all–atom simulations [104]. At $f = 28$ pN, because of long life times and great fluctuations, all the trajectories are of *mixed AG–GF* type. Furthermore, at such a low force, the molecule can completely refold after it partially unravelled. This can happen many times before complete unfolding and the resulting trajectories look like a Greek fret with the end–to–end length going alternately up and down as in figure 5.6.

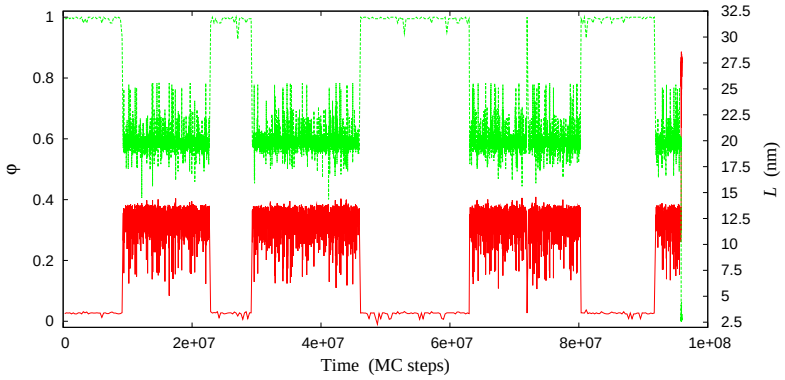


Figure 5.6: A typical trajectory at $f = 28$ pN showing subsequent partial unfolding, refolding events. Red line: end–to–end length, green line: fraction of native contacts, whole FnIII_{10}.

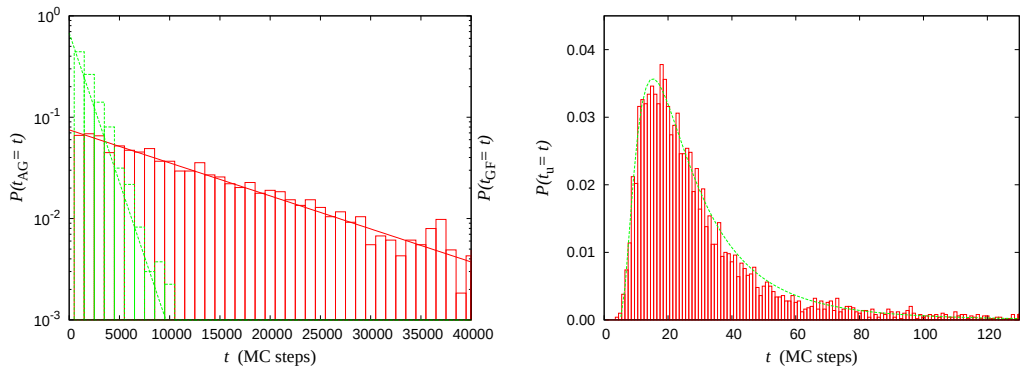
Both the waiting and intermediate states (as the whole unfolding process) are characterized by time lengths varying in a wide range of values for different applied forces and, because of stochasticity, for different trajectories. In table 5.2 the mean life times at various constant forces are reported. The times τ_{AG} and τ_{GF} have been obtained by an average of the times occurring between the first and the second jump in the end–to–end length, which have been defined using the respective threshold values $L = 7.5$ nm and $L = 22.5$ nm. These averages have been calculated only for those trajectories which exhibit the corresponding unfolding pathway, while τ_{AG} and τ_{GF} at force $f = 28$ pN and τ_{GF} at force $f = 40$ pN are not reported in the table because of vanishing frequencies of the corresponding trajectories, as shown in table 5.1. The mean waiting phase time τ_{ws} is the average over the 100 trajectories of the time at which the end–to–end length becomes longer than the threshold value $L = 7.5$ nm. For forces $f = 98$ and 122 pN it does not make sense to define a waiting phase life time, since the protein starts to unravel as soon as the external force is applied at $t = 0$. Finally, the unfolding mean time τ_u is the average on all the trajectories of the unfolding time, i.e. the time at which the molecule reaches the unfolding length previously defined.

The probability distributions of intermediate life times for $f = 81$ pN have been plotted in figure 5.7a, where it can be seen that both distributions can be fitted to the negative exponential function $P(t_s) = (1/\tau_s) \exp\{-t_s/\tau_s\}$ (see equation 3.4) where s is AG or GF, t_s is the intermediate life time of s and $\tau_s = \langle t_s \rangle$ its average. Since AG has a longer life than GF, and being the unfolding time the sum of the waiting phase time and of the intermediate state time, one can naively conclude that if the protein follows the GF pathway, it will reach the unfolded state earlier. For

Table 5.2: Unfolding time (τ_u), waiting phase life time (τ_{ws}), AG intermediate life time (τ_{AG}) and GF intermediate life time (τ_{GF}) at different constant forces. Values are in MC steps and are approximated averages on 100 different trajectories at each force.

	τ_u	τ_{ws}	τ_{AG}	τ_{GF}
28 pN	$2.2 \cdot 10^7$	$8.7 \cdot 10^6$		
36 pN	$3.0 \cdot 10^6$	$8.8 \cdot 10^5$	$4.3 \cdot 10^5$	
40 pN	$8.8 \cdot 10^5$	$2.8 \cdot 10^5$	$6.1 \cdot 10^5$	
46 pN	$2.9 \cdot 10^5$	$6.9 \cdot 10^4$	$2.4 \cdot 10^5$	$1.4 \cdot 10^4$
53 pN	$1.1 \cdot 10^5$	$2.2 \cdot 10^4$	$1.2 \cdot 10^5$	$9.4 \cdot 10^3$
65 pN	$2.7 \cdot 10^4$	$3.6 \cdot 10^3$	$3.8 \cdot 10^4$	$2.9 \cdot 10^3$
81 pN	$6.6 \cdot 10^3$	$1.1 \cdot 10^2$	$1.5 \cdot 10^4$	$1.2 \cdot 10^3$
98 pN	$1.9 \cdot 10^3$		$4.1 \cdot 10^3$	$7.4 \cdot 10^2$
122 pN	$1.5 \cdot 10^2$		$5.7 \cdot 10^2$	$1.9 \cdot 10^2$

the same reason and since at $f = 81$ pN the dominant contribution to the unfolding time comes from τ_{GF} and τ_{AG} one can argue that at this force the exponential function fits well the unfolding times distribution too [118, 86]. Furthermore, at very high forces a lognormal distribution of unfolding times has been proposed [118]. Figure 5.7b shows this behaviour at force $f = 150$ pN and the corresponding fit to $P(t_u) = [1/\sqrt{2\pi}\sigma\,(t_u - t_0)] \exp\left\{-\ln^2[(t_u - t_0)/m]/2\sigma^2\right\}$.



(a) (b)

Figure 5.7: (a) Histograms of the intermediate life times for AG pathway (red line) and GF pathway (green line) at force $f = 81$ pN. Data obtained from 3600 different trajectories. The lines are exponential fits. (b) Histogram of the unfolding times at force $f = 150$ pN. Data obtained from 5000 different trajectories and the bin size of the histogram is 1. The fit is to a lognormal distribution.

In figure 5.8 the average unfolding time τ_u as a function of force f is reported. Three regimes are clearly distinguishable. In the high force regime the unfolding time saturates to a constant plateau, as observed for several other proteins [68].

A fit to the Arrhenius' law 3.1 in the low force regime (from 25 to 60 pN) and in the intermediate force regime (from 60 to 115 pN) permits to find the values of the unfolding length x_u relatively to the two regimes. Fits yield $x_u = 0.13 \pm 0.01$ nm between 60 and 115 pN and $x_u = 0.34 \pm 0.01$ nm in the low force regime. This latter value compares well, given the extreme simplicity of our model, with the experimental results $x_u = 0.38$ nm [110]. By comparing our estimated zero–force unfolding time τ_0 with the corresponding experimental value $\tau_{\texttt{exp}} = 50$ s [110], we find out that a single MC step in our model corresponds to about 25 ns.

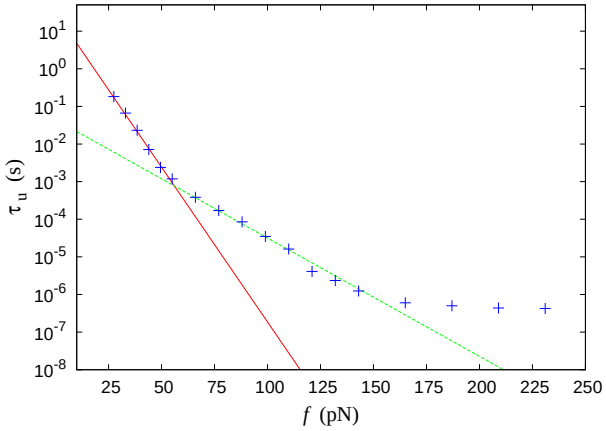


Figure 5.8: Mean unfolding time τ_u as a function of the force f applied to the molecule (average over 100 different trajectories). The red line is a fit to the Arrhenius' law in the range of forces from 25 to 60 pN. In this range we find from the fit $x_u = 3.4 \pm 0.1$ Å. The green line is a fit from 60 to 115 pN, $x_u = 1.3 \pm 0.1$ Å.

Furthermore, looking at data in table 5.2 it is possible to try to interpret the three different force ranges in figure 5.8. In the highest force range there is neither an intermediate state, nor a waiting phase and the unfolding time corresponds mainly to the MC time needed for completing the unfolding where every MC move that unravels the molecule and thus increases the length is accepted and every move that reduces the length is refused, that is, an extremely biased random walk, corresponding to the scenario proposed in [119]. Lowering the force the contributions of τ_{GF} and τ_{AG} to the global unfolding time become important while the waiting phase, if it exists, is still quite short. Finally, in the lowest force interval, also the waiting phase gives its contribution and this matches with the larger slope of the fit line.

Constant Velocity

In this paragraph another manipulation strategy is considered. In the constant velocity protocol the potential $V(L)$ of equation 4.11 is time–dependent and harmonic with the form:

$$V(L) = \frac{k}{2}\left(L_0 + vt - L\right)^2 \tag{5.2}$$

where k is a spring constant, v is the pulling velocity and L_0 is the initial equilibrium elongation. To study the FnIII_{10} unfolding behavior in constant velocity

protocol, MC simulations have been carried out at six different pulling velocities (0.03, 0.05, 0.1, 0.3, 0.5, 1 μm/s) and with a spring constant $k = 30$ pN/nm and an initial length $L_0 = 3.2$ nm. Once again, these conditions are much closer to experimental ones than most previous simulations. In constant velocity simulations, Vogel *et al.* [109] used $v = 50$ m/s (with a spring constant of ~ 4 nN/nm), Klimov and Thirumalai [112] considered $v = 6$ mm/s or faster, while experimental pulling speeds [110, 67] were 0.4 and 0.6 μm/s (with spring constants of 45–50 pN/nm) and in vivo pulling speeds are believed to be even smaller [114]. Only the all–atom Monte Carlo simulations by Mitternacht *et al.* [104] could probe constant pulling speeds in the same range as those considered here (with a spring constant of 37 pN/nm).

In figure 5.9 there is a sketch of the possible unfolding pathways scheme in the constant velocity case. Consistent with our constant force results and with previous simulations [109, 112, 104], at each value of v most of the trajectories start with the detachment of strand G, giving rise to an intermediate corresponding to the shallow minimum around 6 nm in the free energy landscape of figure 5.2b. In few runs strand A is the first to unravel but, before any other strand unravels, it refolds, with the consequence that the unfolding of the molecule visits in any case the intermediate *G*. Then the unfolding continues through a phase in which strand A is gradually unzipped and when this unzipping is completed the molecule reaches the intermediate *AG* (end–to–end length $\sim$ 13.5 nm). As in the constant force case a *mixed AG–GF* behavior has been found: some trajectories do not stay in the *AG* intermediate till the complete unfolding but they may jump from *AG* to *GF* intermediate (end–to–end length $\sim$ 14 nm) and back. Table 5.3 reports the relative frequencies of various unfolding pathways. It is worth noting that, since statistical fluctuations are greater at low pulling rates, the number of *mixed AG–GF* trajectories, and the number of trajectories in which strand A unravels before strand G, grow as pulling velocity decreases. Typical trajectories are reported in figure 5.10, where it is also possible to appreciate that the force applied to the fibronectin domain generally increases with time, except for the abrupt rupture of the native state and of the intermediate state *AG*, and during the unzipping of strand A.

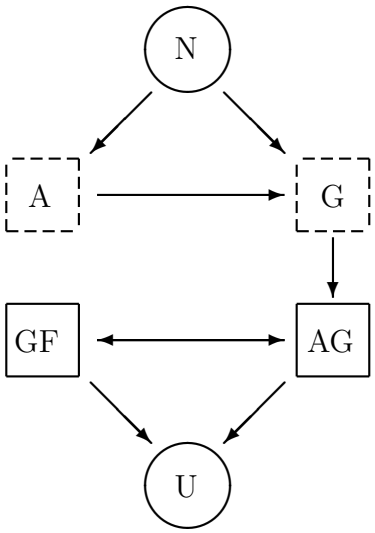
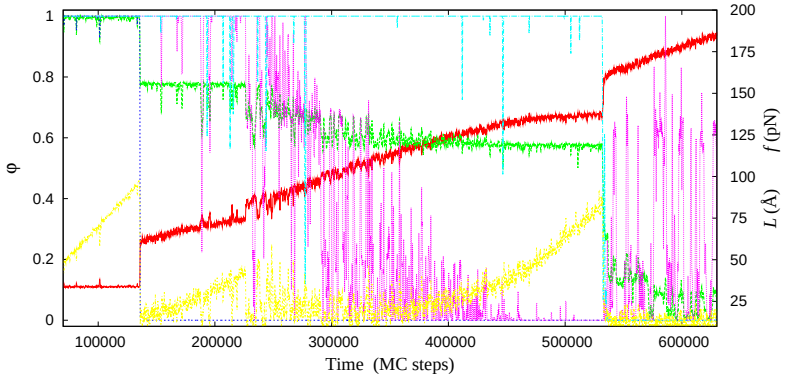


Figure 5.9: Unfolding pathways scheme of FnIII_{10} at constant velocity. Intermediate states in the full square boxes have a rupture force remarkably higher than those in dashed boxes.

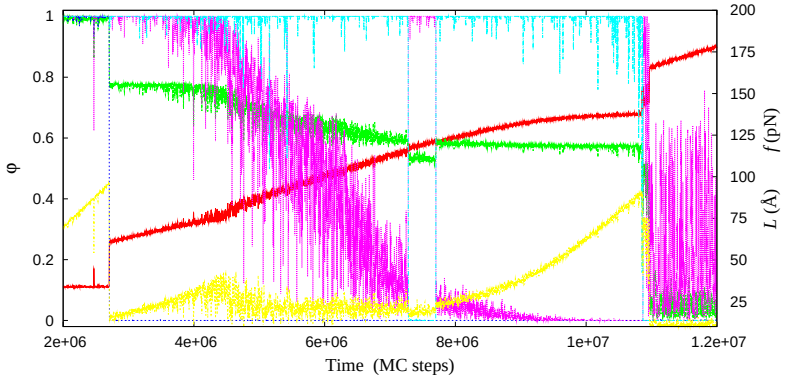
The average rupture forces of the native state and of the intermediate states have been reported in table 5.4. At the pulling speed considered, the average rupture force for the native state ranges between 80 to 100 pN, which is in remarkable agreement with the AFM results. Fernandez and coworkers reported 75 pN when pulling at 0.6 μm/s [110] and 100 pN at 0.4 μm/s [67]. Simulations of Mitternacht

Table 5.3: Relative frequencies of unfolding pathways. 100 trajectories for each value of the velocity.

	G		A → G	
	AG	mixed AG–GF	AG	mixed AG–GF
1 μm/s	0.82	0.15	0.03	0.00
0.5 μm/s	0.76	0.20	0.03	0.01
0.3 μm/s	0.49	0.39	0.09	0.03
0.1 μm/s	0.11	0.85	0.01	0.03
0.05 μm/s	0.08	0.87	0.01	0.04
0.03 μm/s	0.07	0.79	0.00	0.14



(a) *Unfolding pathway*: G → AG



(b) *Unfolding pathway*: mixed AG–GF

Figure 5.10: MC time evolution of the end–to–end length (red line), force (yellow line) and of a few order parameters with a constant velocity of 1 μm/s (a) and 0.03 μm/s (b). Other colours as in figure 5.5. Each point in the graph is a mean over a bin of 10 (a), and 2000 (b), points in the corresponding trajectory. Bins been used to reduce fluctuations in the plot.

Table 5.4: Average rupture forces (force unit: pN).

	N → G	G → AG	AG → U	GF → U
1 μm/s	98.5 ± 6.4	40.8 ± 2.6	99.6 ± 9.9	77.3 ± 7.7
0.5 μm/s	96.1 ± 6.1	42.2 ± 2.6	96.5 ± 7.3	77.5 ± 4.0
0.3 μm/s	94.5 ± 6.9	43.4 ± 2.4	92.4 ± 7.8	76.5 ± 5.8
0.1 μm/s	89.0 ± 6.5	45.4 ± 1.8	86.5 ± 7.9	69.9 ± 5.8
0.05 μm/s	87.8 ± 5.1	46.1 ± 1.6	81.9 ± 8.4	67.6 ± 5.9
0.03 μm/s	87.3 ± 5.8	46.9 ± 1.7	81.5 ± 9.7	66.7 ± 5.4

et al. [104] led to values from 88 pN at 0.03 μm/s to 114 pN at 0.1 μm/s. In the same work, the average rupture forces of intermediate states A and G were reported to range between 40 and 80 pN while the rupture forces of intermediates with two strands detached has been found to be higher, ranging between 107 and 216 pN for intermediate FG, between 115 and 198 pN for intermediate AG and between 283 and 318 pN for intermediate AB. Li *et al.* [67] reported an average unfolding force of the intermediate states of about 50 pN. In this work, pulling on suitable mutants, two kind of intermediates were inferred on the basis of experimental results, namely G and AB. In our model we did not observe intermediate AB, while intermediate G has an average rupture force between 40 and 50 pN. The other intermediates we observed, AG and GF, are more stable, with average unfolding forces around 70–100 pN in accordance with the simulations of Mitternacht *et al.* [104].

Furthermore it is worth noting that, except for intermediate G, the obtained average rupture forces increase with the pulling speed, as predicted by theories [72, 73, 96, 76, 77] and verified in experiments [74]. An explanation for the different behavior of the rupture force of intermediate G is that unravelling of strand A is an unzipping event rather than an abrupt rupture and then statistical fluctuations, which are greater at lower pulling velocity, have more influence on the value of the rupture force.

The distribution of the unfolding forces is well fitted by the theoretical result [72] (see equation 3.7):

$$P(f) = \frac{1}{t_0 r} e^{\beta f x_u} \exp\left[-\frac{k_B T}{r x_u t_0}\left(e^{\beta f x_u} - 1\right)\right] , \tag{5.3}$$

where t_0 is the average unfolding time in the absence of force. Such an equation corresponds to the rupture force probability distribution of a single molecular bond subject to a force that increases linearly with a rate r. In figure 5.11 the unfolding force histogram at $v = 0.5$ μm/s has been plotted. The fit is to equation 5.3, with $a = r \cdot t_0$ and x_u as fitting parameters, and gives $x_u = 0.8$ nm, which is larger than the value found for the constant force set-up. However, it must be kept in mind that the above theoretical result was derived for a force which is linear in time with a slope r, while here the force is associated to a harmonic potential which moves at constant velocity v.

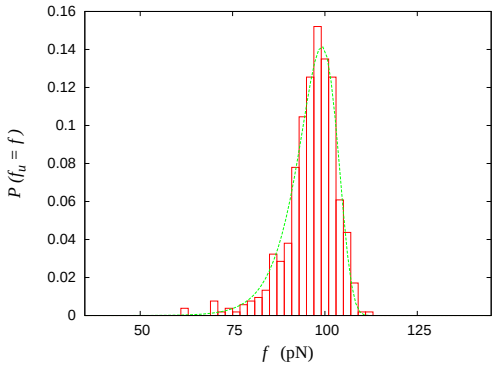
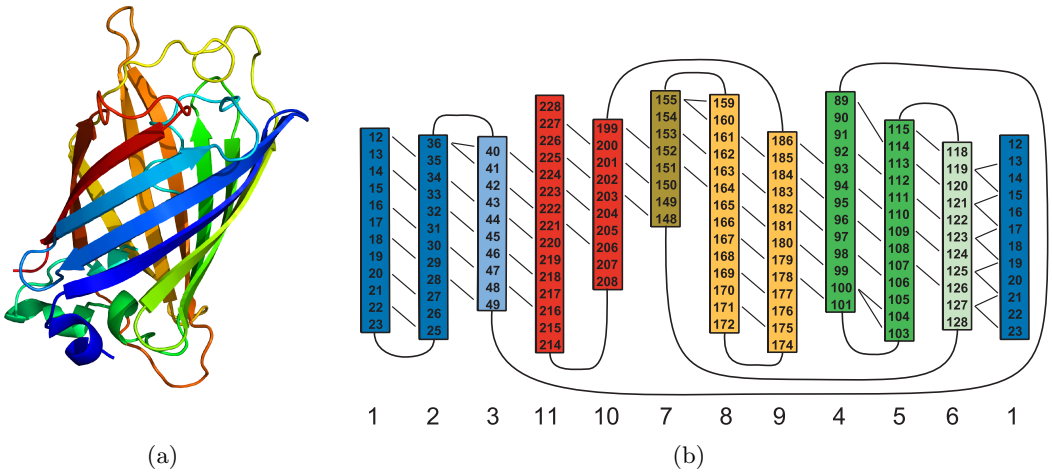


Figure 5.11: Distribution of the rupture forces of the native state at pulling velocity $v = 0.5$ μm/s. Data obtained from 500 different trajectories; bin size of the histogram is 2. The fit is to eqution 5.3.

5.2 The Green Fluorescent Protein

One of the most interesting proteins studied is the Green Fluorescent Protein (GFP) from the jellyfish *Aequorea victoria*, which exhibits bright green fluorescence when exposed to light with a suitable wavelength. Thanks to its fluorescence properties, GFP has been extensively used as a marker of gene expression and protein localization, as an indicator of protein–protein interactions and as a biosensor [120,121]. In the year 2008, O. Shimonura, M.Chalfie and R. Tsien were awarded the Nobel prize in chemistry for their research on GFP (see [122] for a recent review about it). Wild GFP is a 238–residues long protein constituted by 11 β–strands arranged in a cylinder–like structure, called β–barrel, which has a diameter of 2.4 nm and a height of 4.2 nm (see figure 5.12). A short α–helix is present at the beginning of the chain and other short α–helices are along the barrel axis.



(a) (b)

Figure 5.12: Native structure of GFP (pdb ID 1B9C) with β–strands labeled 1–11 in sequence order. (a) Sketch generated by PyMOL. (b) Order of barrel β–strands with residues belonging to them and residue–residue hydrogen bonding.

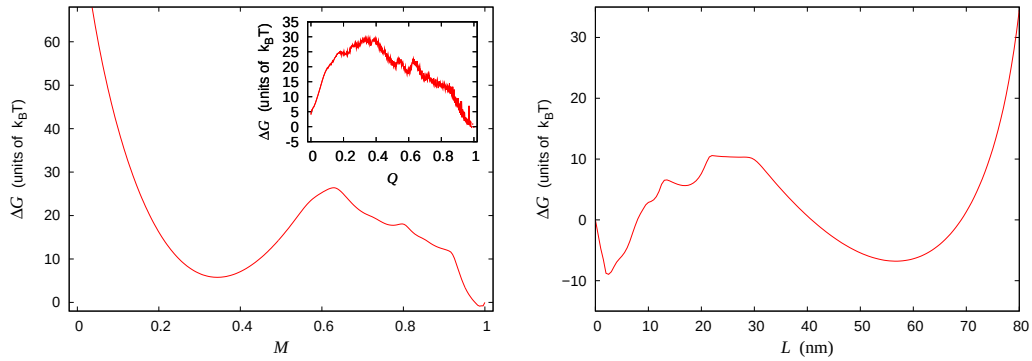
The chromophore structure is obtained by an autocatalytic post–translational cyclization and oxidation process around residues 65–66–67 which are embedded in the barrel [121,123]. The role of the can–like structure is both to protect the chromophore and to help its formation by restricting its available confomational space and, indeed, it is commonly believed that GFP fluoresces only when its structure is almost intact [124,125]. Li *et al.* [125], using deletion analysis, indicated the domain 7–229 as the minimal native structure to have fluorescence.

Furthermore GFP has been the subject of mechanical experiments and numerical simulations [66,62,126,127,90] aimed at characterizing its response to external force and the structure of its intermediate states. The final goal of such studies is a full characterization of the GFP response to mechanical stress, so as to pave the way to its use as a molecular force sensor. Dietz and Rief [66] found that unfolding of GFP always starts with unravelling of the N–terminal α–helix, which is revealed by a very smooth "hump–like" transition with a short contour length increase of 3.2 nm in the force–extension traces. This intermediate state has a rupture force of about 104 pN but, before complete unfolding, the molecule visits another intermediate state which lacks also one of the β–strands. Combining experiments on two engineered mutants and simulations, Mickler *et al.* [126] showed that the β–strand that breaks first is the N–terminal one (β_1) in 72% of cases and the C–terminal one (β_{11}) in remaining 28% of cases. The same authors also showed that in the former case a third intermediate state exists, which has three β–strands detached, namely β_1, β_2 and β_3. Finally Dietz *et al.* [62], using cysteine engineering [61], pulled the GFP module along precisely controlled directions obtaining fracture forces widely varying from 100 to 600 pN according to the pulling direction.

5.2.1 Equilibrium properties

Figure 5.13a shows the computed free energy profile as a function of the fraction of native residues M and of native contacts Q at the denaturation temperature $T = 356$ K [128]. Inspection of these plots indicates that at this temperature: (i) when the protein is in its native state, all the native contacts are formed, and almost all the residues are in the native configuration; (ii) in the unfolded state, no native contacts are formed, and 1/3 of the residues are in the native configuration; (iii) the transition state corresponds to $Q \sim 0.3$–0.4, while at $Q \sim 0.5$–0.7 there are some hints of the possible existence of intermediate states; (iv) the unfolding barrier is of the order of 25 k_BT in both cases. These results for the free energy profile as a function of Q can be compared to the result obtained by Andrews *et al.* [103] by weighted–histogram analysis of molecular dynamics data. The qualitative picture is very similar, although some differences can be observed. The molecular dynamics results show that some fluctuations in native contacts are allowed in both the native and unfolded states, a feature which is missing in our result due to the extreme cooperativity of the model. Moreover, the unfolding barrier is predicted by Andrews *et al.* to be around 15 k_BT: this is consistent with the observation that our model predicts systematically higher energy barriers and unfolding forces as will be discussed later.

In figure 5.13b the free energy profile as a function of the end–to–end length is reported for a typical case, where the equilibrium unfolding force $f = 35.9$ pN is



(a) (b)

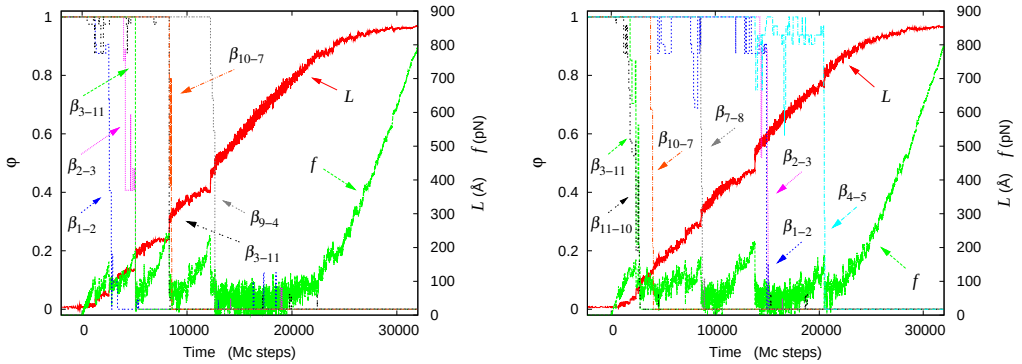
Figure 5.13: (a) Free energy landscape as a function of the fraction of native residues M at $T = 356$ K. Inset: free energy landscape as a function of the fraction of native contacts Q at $T = 356$ K. (b) Free energy landscape as a function of the end–to–end length L with $T = 293$ K and force $f = 35.9$ pN.

applied to the molecule ends. Besides the native and the unfolded minima we can see three other local minima (or bends which become actual minima at different values of the force) around 11, 18 and 25 nm. As it will be shown in detail in the next section, these local minima and bends correspond to intermediate states effectively populated in the MC simulations. Analysing the equilibrium probability $0 \leq \langle m_k(L) \rangle \leq 1$ that the $k-$th residue is native–like when the molecule total elongation is L (data not shown), it has been found that such bends correspond to the following structures: β_1 and β_{11} (for $L \simeq 11$ nm), $\beta_{10}\beta_{11}$ ($L \simeq 18$ nm) and $\beta_1\beta_2\beta_3$ ($L \simeq 25$ nm). Here and in the following, $\beta_k \cdots \beta_n$ denotes an unfolded structure of the GFP, where β–strands from k to n are not in a native–like conformation, i.e. they are unfolded (in all these structures the N–terminal α–helix is also not in a native–like conformation).

5.2.2 Unfolding pathways

Let us start to consider simulations at constant velocity, mimicking the effect of an AFM cantilever, which is retracted at a velocity v. The force is applied to the molecule ends and the value of the spring constant has been set to $k = 30$ pN/nm. The behavior of the protein at velocities $v = 0.3$ μm/s, 1 μm/s, 2 μm/s and 3.6 μm/s, has been investigated. In most of the trajectories considered, the N–terminal α–helix is the first secondary structure element to unravel. This event is typically associated with very small signals in the end–to–end length trace almost masked by fluctuations, at odds with the clear jumps we observe in the end–to–end length for the detaching of β–strands (see figure 5.14). This is analogous to what occurs in the experiments where the unfolding of the helix is associated to a very smooth "hump–like" transition with a short contour length increase of 3.2 nm in [66] and by a small jump in the root mean square distance as a function of time in [126].

It has been found that, at all velocities considered, in less than 10% of the trajectories β_{11} is the first strand to unravel, while the remaining trajectories follow



(a) *major unfolding pathway* (b) *minor unfolding pathway*

Figure 5.14: Typical unfolding trajectories of a GFP module under constant velocity pulling ($v = 0.3$ μm/s). Length L, force f and weighted fractions $\varphi_{\beta_{i-j}}$ of strand–strand contacts as functions of time for two typical cases: major (a) and minor (b) unfolding pathway, see text.

the major unfolding pathway found in experiments. Figure 5.14 shows the behavior as a function of time of the end–to–end length L, of the force and of several weighted fractions of native contacts between adjacent β–strands $\varphi_{\beta_{i-j}}$ according to equation 5.1.

Major unfolding pathway. Inspection of figure 5.14a, corresponding to the major unfolding pathway, provides clear evidence that there are three main unfolding events. (i) A drop in the number of contacts between strands β_1 and β_2, signalling the unfolding of β_1 (actually the α–helix has already unfolded, as discussed above, but the corresponding weighted fraction of native contacts φ_α is not reported in the figure for the sake of clarity). The length of the corresponding intermediate state is in the range 10–12.5 nm, where the free energy profile of figure 5.13b shows a bend. (ii) A drop in the number of contacts involving strands β_2 and β_3, signalling the unfolding of these strands. The corresponding intermediate length is around 20 nm, where the free energy profile has a local minimum. (iii) A drop in the number of contacts involving strands β_{10} and β_{11}, signalling the unfolding of these strands. The corresponding intermediate length is in the range 30–37 nm: inspection of figure 5.14b suggests that for such an elongation the molecules already lies in the basin of the unfolded minimum. At this point the molecule can be considered as unfolded notwithstanding a last rupture event could be seen.

Minor unfolding pathway. Figure 5.14b corresponds to the minor unfolding pathway and it shows that, in this case, the first strand to unravel is β_{11} followed by β_{10}. Mickler *et al.* [126] traced the unfolding pathway only up to the β_{11} intermediate because the subsequent event is the flattening of the barrel but, after the barrel flattens, there is at least another rupture event as the last force jump in figure 1b of reference [126] shows. It is reasonable to assume that this event is related to the breaking of native–like contacts between the beta strands, which were not ruptured

during the flattening of the barrel. Our model, which lacks a fully three–dimensional representation, cannot describe the flattening of the barrel, while it can describe with a high time resolution the breaking of the beta strand contacts, which here yield in a few distinct steps.

It is now possible to put the local minima and bends of the free energy landscape of figure 5.13b (which is a thermodynamic equilibrium property of the system) in correspondence with intermediates found in our simulations and in experiments (which are performed in non–equilibrium conditions). Some of these features of the free energy profile are indeed barely visible, but the equilibrium probabilities $\langle m_k(L) \rangle$, introduced in the previous section to give a structural interpretation of the various minima and bends, are perfectly consistent with the nonequilibrium m_k values obtained from the simulations, which allow to identify the structures of the nonequilibrium intermediates.

In reference [66] the authors observed two intermediates with separation values from native configuration of 3.2 and 10 nm. The first one, is an intermediate with only the N–terminal α–helix detached that profile of figure 5.13b does not show, while the second is an intermediate with the N–terminal α–helix detached and a β–strand detached which corresponds to the bend at 11 nm (9.2 nm away from native state) in figure 5.13b. The authors of reference [126] reported the existence of another intermediate (N–terminal α–helix and first, second and third β–strands detached) with a distance of 26.3 nm from the native state (16.3 nm from the previous second intermediate) which is clearly associated to our dip at 25 nm, corresponding to the intermediate state $\beta_1\beta_2\beta_3$. The 18 nm intermediate ($\beta_{10}\beta_{11}$) instead has no analogue in experiments.

5.2.3 Pulling along different directions

Let us now consider simulations where the points of force application are not the molecule ends, so that the direction of the force with respect to the molecule is varied. Table 5.5 reports, for different directions (specified through the application point residue numbers), the mean unfolding forces, where unfolding is defined as unravelling of the first β–strand. Since most of these directions were considered in experiments [62], at least at $v = 3.6$ μm/s, it is interesting to compare the results obtained through the WSME model, to the experimental ones. The obtained unfolding forces are systematically larger than the experimental values, with the largest discrepancies (a factor 2 to 3) occurring for directions 3–212 and 132–212. However, it is interesting that in spite of the simplicity of the model, which lacks a fully three–dimensional representation, the orders of magnitude for the rupture forces are correct and many qualitative aspects are reproduced. In particular, by analyzing the experimental data one finds that the unfolding force increases with the following order: (i) pulling along the end–to–end direction (it must be noted that the rupture force along this direction was measured for $v = 0.3$ μm/s instead of 3.6 μm/s as most other directions); (ii) 3–212 and 132–212 directions, the corresponding rupture forces are equal within the experimental error; (iii) 182–212 and 3–132 directions, the corresponding rupture forces are equal within the experimental error (though the latter was measured for $v = 2$ μm/s); (iv) 117–182 direction.

This hierarchy is respected by our results: we find that the rupture force in-

Table 5.5: Unfolding forces at different velocities for different directions. Experimental values (* from reference [66] and † from reference [62]) in parentheses.

	Unfolding force (pN)		
Direction	$v = 0.3$ μm/s	$v = 2$ μm/s	$v = 3.6$ μm/s
end–end	140 ± 3	177 ± 7	184 ± 13
	$(104 \pm 40)^*$		
182–end	196 ± 7	226 ± 6	244 ± 7
3–212	244 ± 12	298 ± 12	317 ± 20
			$(117 \pm 19)^\dagger$
132–212	251 ± 7	266 ± 3	273 ± 6
			$(127 \pm 23)^\dagger$
132–end	306 ± 12	360 ± 20	381 ± 26
182–212	365 ± 2	390 ± 7	409 ± 15
			$(356 \pm 61)^\dagger$
3–132	383 ± 16	471 ± 49	535 ± 80
		$(346 \pm 46)^\dagger$	
117–182	467 ± 3	501 ± 11	512 ± 11
			$(548 \pm 57)^\dagger$

creases when we consider the pulling directions as ordered above, the only exception being for 3–212 and 132–212, whose unfolding forces are not equal (we obtain a smaller force for the latter), and the same holds for 182–212 and 3–132 (we obtain a larger force for the latter).

Table 5.6 reports the potential width values x_u corresponding to the rupture of the first β–strand for different directions. These were obtained through a fit of the most probable unfolding force $f_{\text{rupt.}}$ as a function of velocity to the Evans–Ritchie theory [72, 73], which gives (see equation 3.8)

$$f_{\text{rupt.}} = \frac{k_B T}{x_u} \ln\left(\frac{\tau_0 x_u}{k_B T} r\right) , \tag{5.4}$$

where τ_0 is the unfolding time at zero force. It must be kept in mind that in the Evans–Ritchie theory the force grows with a constant rate $r = k \cdot v$ and hence its applicability to the present case (harmonic potential whose center moves at constant velocity v) is only approximate.

The obtained potential widths are consistent with experimental ones only in a few cases (end–end, 3–132) but, once again, this might be attributed to the fact that WSME model lacks a fully three–dimensional representation. Furthermore, it must also be observed that the Evans–Ritchie theory is built on the assumption that x_u is independent of the applied force, and this can be another source of error in the determination of x_u. This assumption was relaxed in more recent theories [77, 129] which yield generalizations of equation 5.4, which predict that the $f_{\text{rupt.}}$ versus $\ln v$ plot is nonlinear, with the slope being an increasing function of v, as observed in many experiments. Indeed our data show some nonlinearity, but this is too small to apply these theories, probably because our velocities span only one order of

Table 5.6: Potential width x_u obtained from a fit to Eq. 5.4. Experimental values between parentheses.

direction	x_u (nm)	direction	x_u (nm)
end–end	0.21 ± 0.04	132–end	0.14 ± 0.01
	(0.28 ± 0.03)		
182–end	0.22 ± 0.03	182–212	0.24 ± 0.04
			(0.14 ± 0.002)
3–212	0.14 ± 0.01	3–132	0.11 ± 0.03
	(0.45 ± 0.01)		(0.125 ± 0.005)
132–212	0.46 ± 0.06	117–182	0.22 ± 0.01
	(0.32 ± 0.005)		(0.12 ± 0.003)

magnitude. Previous applications of these theories [77, 30, 86, 129] were done on data sets with velocities spanning 4–5 orders of magnitude, such that the nonlinear effects were much more important.

Pulling the GFP molecule along the same directions but with a constant force protocol led to results consistent with the simulations at constant velocity. In particular it has been found that pulling end–to–end or along directions 3–212 and 182–end, it is possible to unfold the molecule at the relatively low force of 100 pN, in a reasonable MC simulation time of few days for each trajectory. GFP is instead stiffer to pull along the other directions. In increasing order of stiffness we have: end–to–end, 3–212 and 182–end (minimal unfolding force: 100 pN), 3–132 and 132–212 (190 pN), 132–end (220 pN), 182–212 (220 pN) and, finally, 117–182 (340 pN). Furthermore, consistent with the fibronectin case and the idea that unfolding pathways become more deterministic with increase in constant pulling force [104], it has been found that, in general, the number of possible pathways decreases with increase in force and that at very high forces (600 pN or more), independently of the pulling direction, no intermediate states can be detected.

5.2.4 GFP polyprotein as a force sensor

As it has been discussed above, the equilibrium properties of the GFP at constant force, can be obtained exactly, for any pulling direction. Exploiting this result, one can design a polyprotein where each module is connected to the neighboring ones through different points of force application, as illustrated in figure 5.15. Such a molecule can easily provide the value of the applied force in a wide range of values, and thus can be used as a force probe.

For example Dietz *et al.* [62] already proposed a copolymer with mixed linkage geometries GFP(3,212)(132,212), made up of several GFP modules, where a module linked by its (3,212) residues to the main structure was alternated with a module linked by its (132,212) residues. Such a molecule can be easily obtained by using the cysteine engineering method discussed in reference [61], which allows one to construct polyproteins with precisely controlled linkage topologies: the points of force application to each module correspond to the position of the linking cysteines in the folded tertiary structure.

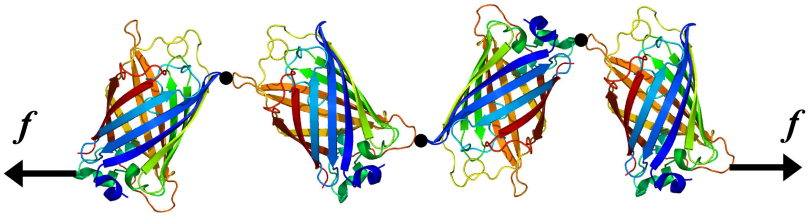


Figure 5.15: Sketch of a polyprotein made of various modules connected between them through different residues.

In order to understand the general behavior of the model polyprotein under a constant force we first investigate the response to a constant force of a single GFP module. The corresponding equilibrium unfolding forces are reported in table 5.7.

Table 5.7: Equilibrium unfolding force for different directions.

direction	unfolding force (pN)	direction	unfolding force (pN)
end–end	35.9	182–end	65.0
3–212	38.9	117–190	67.3
3–182	42.6	102–190	71.2
117–end	50.8	117–182	78.1
3–132	56.4	132–182	96.7

At equilibrium a force applied to the free ends of the polyproteins will have the same value throughout the whole chain. Thus, the different modules will unfold at different values of the force, according to the hierarchy shown in table 5.7, and thus the luminescence will be different for different values of the force. If we assign a value 1 (in an arbitrary scale) to the maximum possible luminescence, where each module is emitting green light, a luminescence of 0.5 will correspond to a configuration, and thus to a force, where half of the modules are unfolded (non intact structure). Given that each module with a different linkage has a different unfolding force, we obtain a curve like the one shown in figure 5.16, relating the luminescence of the polyprotein to the force applied to its free ends, where the force ranges from 35.9 to 96.7 pN. It is worth to note that interface interactions and aggregation effects between neighboring units in polyproteins similar to the one we propose, have been ruled out by experimental investigations [62].

It is worth noting that in principle more modules, with different linkages, can be added, and this would give a more precise determination of the force. Once the polyprotein here proposed has been engineered, a curve like the one shown in figure 5.16 can be very easily obtained in an optical tweezers experiment at constant force as those discussed, for example, in reference [99]. This approach would also allow one to calibrate the device.

Although unfolding studies of GFP along different directions were already performed [62,126], those previous studies considered the dynamic–loading set up, with a constant retraction speed of the AFM cantilever. On the contrary we investigate here for the first time the unfolding at constant force of GFP. The unfolding force

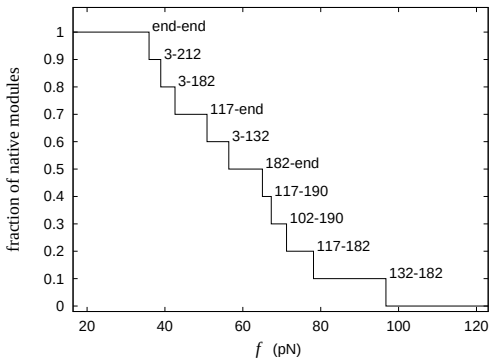


Figure 5.16: Fraction of native–like modules as a function of force at $T = 293$ K. Each "step" corresponds to the unfolding of a different module in the polyprotein and thus to a decrease in the luminescence by a "unit".

of a molecule under dynamic loading depends not only on the molecular features, but also on the force rate, and thus a force probe based on those data must be able to measure at the same time the loading rate and the rupture force. The constant force probe proposed here instead does not exhibit this drawback.

Chapter 6

Protein folding in the cell: Confinement of proteins

Protein folding in the cell occurs in a heterogeneous environment, a perturbation that may alter both the thermodynamics and kinetics of folding relative to the observations made in dilute conditions. Indeed, in the past the majority of experiments on protein folding have been carried out in diluted solutions but in the last two decades it has become clear that these experiments do not take into account two issues which arise *in vivo* and whose relevance on thermal stability and equilibrium rates is not negligible. Namely, crowding and confinement [130, 131, 132, 133].

Crowding refers to the fact that the interior of the cell contains a large number of macromolecules such as lipids, carbohydrates, nucleic acids and proteins themselves. Actually, no single macromolecular species occurs at high concentration but, taken together, the macromolecules occupy about 30% of cells internal volume [130]. This fraction could even reach 40% in *Escherichia Coli* [134].

Confinement is instead a mere limitation in the volume available to the polypeptide chain. For example, the earliest environment encountered by a nascent polypeptide chain is the ribosome exit tunnel and many newly synthesized proteins rely on assistance by molecular chaperones to reach their native states efficiently and at a biologically relevant timescale. By enclosing newly synthesized or stress–denatured polypeptides in Anfinsen–like cages, molecular chaperones protect them from misfolding and aggregating in the highly crowded cellular environment. In *E. Coli*, approximately 250 different proteins interact with GroEL chaperone upon synthesis and more than 80 of them have an obligate dependance on encapsulation into the GroEL hydrophilic cavity [135, 42]. It has been shown that confinement inside the GroEL–GroES could result in a significant acceleration of folding as compared to folding in free solution [136]. However, for the sake of completeness, it is worth to note that GroEL role is not limited to offer a passive cavity to the substrate but it may also have a more active role in the process of protein folding [137].

Understanding better the role of crowding and confinement is a necessary step towards an improved knowledge of how protein folding works in, and is modified by, cellular environment. The present chapter is inserted into this context. A further generalization of the WSME model is proposed, which can handle proteins confined between two walls and whose equilibrium thermodynamics can be still solved ex-

actly. The confined WSME model has been used to study thermodynamics and kinetics of three ideal structures and three simple proteins in confining conditions. The ideal structures are a 10 residues ideal α–helix, a 2–stranded and a 3–stranded ideal β–sheets, each with 7 residues per strand. Real structures are a 3–helix bundle, protein G and its C–terminal β–hairpin. The chapter is organized as follows: first of all, a brief review of the results obtained in the field is presented, then the confined WSME model is described and finally sections 6.3.1 and 6.3.2 focus on the confinement–induced changes of thermodynamics and kinetics respectively.

6.1 Enhancement of thermal stability, increase in folding rates and other aspects of confinement

Studying protein folding properties in a crowded environment is experimentally possible simply by adding high concentrations of macromolecules to solutions, but this approach has problems because of specific interactions which arise between proteins and crowding agents and because crowding promotes protein–protein aggregation [130]. Based on the idea that the main effect of crowding is the reduction of volume available to the protein due to steric constraints, theoretical studies and simulations have shown that crowding may be quantitatively mapped onto confinement as long as crowding agents can be modelled as hard spheres and the volume fraction occupied by them does not exceed 10% [33]. Thanks to this mapping, experimental and theoretical studies on confinement may give many hints also for crowding effects.

However the above conditions often do not hold in the cell interior because of too high concentration of agents or presence of macromolecules–protein attractive interactions. In addition, gradients in macromolecule concentrations may exist [138] and, from a more general point of view, crowding is a dynamic phenomenon in nature whereas confinement is a static one. Thus, the mapping is not close enough to draw a completely satisfactory analogy between crowding and confinement.

An experimental procedure to mimic the effects of confinement (and, to some extent, of crowding) is the encapsulation of proteins within pores of silica gels [139, 140] or glasses [31] or polyacrylamide gels [32]. These experiments reported, for most of the considered proteins, an increase in thermal stability when they are confined into nanopores. Melting temperature (T_{f}) shift is even dramatic in the cases of α–lactalbumin and RNase A, being as large as about 30 K [139,31]. On the other hand effects of crowding seem to be more controversial and recent experiments suggested that crowding influence on stability is modest [138,141].

The commonly accepted reason for the increase in stability is the change in conformational entropy induced by confinement [142,143,144,145,146,147]. Encapsulating the protein in a given volume disallows the most expanded configurations of the denatured state ensemble and so indirectly favours more compact structures and, among them, the folded state. The same argument explains also why confinement should lead to an increase in folding rates (k_{f}) as long as the nanopore size is large enough to contain the folded state and to permit chain reconfigurations around it [142,143,144,145,146,147,148].

Let us consider a confined polymer chain in three dimensions, with d_{c} the number

of subtracted dimensions. From polymer physics we know that the free energy F required to confine the chain, follows a simple power law dependance on the number of monomers N and on the size of the cage R [149, 150]:

$$\frac{F}{k_B T} \simeq N^{\delta} \left(\frac{R}{a}\right)^{-\gamma} , \tag{6.1}$$

where a is the monomer length. For an ideal gaussian chain confined between two walls ($d_c = 1$), in a cylinder ($d_c = 2$) or in a spherical cavity ($d_c = 3$), $\gamma = 2$ and $\delta = 1$, while for an excluded volume chain with $d_c = 1, 2$, $\gamma = 5/3$ and $\delta = 1$. Finally, for an excluded volume chain with $d_c = 3$, $\gamma = 15/4$ and $\delta = 9/4$.

As shown by Takagi *et al.* [146], this scaling law can be used to deduce a similar law for the folding temperature of a protein. At the folding temperature T_f, the free energy of the native G_N and the denatured G_U states are equal by definition: $G_N = G_U$. The native state has only negligible conformation entropy, and thus it is fair to set the native entropy at zero. In the unfolded state, it is instead fair to neglect the enthalpy contribution: if S is the entropy of the unfolded state, in bulk conditions $G_U = -T_f S$, while upon confinement G_U increases of a quantity $\Delta G_U \sim T(R/R_0)^{-\gamma}$, where R_0 is a constant length and it has been assumed that the confinement to a characteristic length R affects only the denatured state (as in figure 6.1). Since the energy of the folded state remains the same, it follows that $\Delta T_f \sim R^{-\gamma}$.

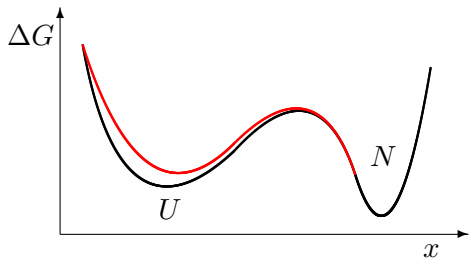


Figure 6.1: Sketch of the free energy profile modification from bulk conditions (black line) to confinement conditions (red line). x is a generic reaction coordinate.

A similar argument allows to find a scaling law for the folding rates [147]. To a first approximation, the folding rates k_f can be estimated from high–friction Kramers' kinetics [57] by using $k_f \propto D \exp[-(G_{TS} - G_U)/k_B T]$, where D is an effective diffusion coefficient and G_{TS}, G_U are the free energies of the transition state and the unfolded state respectively. Now, if we assume that the diffusion coefficient and the free energy of the transition state do not change with confinement, and that the free energy of the folded state increases of a quantity $\Delta G_U \sim k_B T(R/R_0)^{-\gamma}$, it follows that $\Delta \ln k_f \sim (R/R_0)^{-\gamma}$.

These assumptions become less accurate for proteins confined to very small cavities since, in this case, also the free energy of the transition state and of the native state may undergo some modifications.

Unfortunately experiments have not yet been able to prove such scaling law [32, 140] but coarse–grained–model–based simulations have reproduced the expected behavior [146, 147]. Using a Gō–model α–carbon representation of proteins and Langevin simulations in a cylindrical cage, Takagi *et al.* [146] found $\gamma = 3.25 \pm 0.09$. Best and Mittal [147] simulated confinement of protein G and a 3–helix bundle in different geometries and reported that for $d_c = 1, 2$ both values $\gamma = 2$ and $\gamma = 5/3$ are a good estimate of the behavior of the two proteins, but they also remarked that it is hard to distinguish which value fits best the simulations because least square fitting of power laws can produce biased estimates of parameters for small samples. For spherical confinement the same authors reported a behavior which is stronger than $\gamma = 2$ but much weaker than the expected value for the excluded volume chain ($\gamma = 15/4$). It is therefore still not completely clear whether protein–folding thermodynamics and kinetics follow a polymer–like scaling behavior under confinement and what γ value will be relevant.

Furthermore many other aspects are still not completely clear and need deeper investigation. In particular Takagi *et al.* [146] have shown that acceleration of folding by confinement is more prominent for proteins with a greater relative contact order but a comprehensive study of how γ exponent depends on the protein topology and for which values of R the above scaling law remains valid is still missing. Furthermore very few work has focused on how confinement affects the nature of the transition state ensemble. Cheung and Thirumalai [151] have studied into details the changes, upon crowding and confinement, in the transition state ensemble of a three–stranded β–sheet WW domain, showing that it does not change significantly except that the average width of its configuration decreases with respect to bulk conditions. Understanding if these results hold for most of the proteins and, if not, which other factors may arise, would be necessary for a complete characterization of confined–induced folding. Finally, confinement and crowding may induce modifications in the folding pathways of proteins which fold passing through intermediate states. Nothwithstanding the relevance of the problem, up to my knowledge, it has still to be addressed in the case of confined proteins and only Pincus and Thirumalai [152] have investigated mechanical unfolding pathways in crowded environment.

6.2 Confined WSME model

Before introducing confinement into the model, let us remember that the zero–force partition function, constrained at a given end–to–end length value L (equation 4.30),

$$\mathcal{Z}(L; f = 0) = \sum_{m \in \{0,1\}^N} \sum_{\sigma \in \mathcal{O}(m)} \delta_{L,L(m,\sigma)}\, e^{-\beta H(m,\sigma;f=0)} \,, \tag{6.2}$$

can be recursively calculated building up the protein residue by residue and evaluating at each step n the partition function $z_n(L)$, where n is the number of residues

at that step (see section 4.2.2 for details). The corresponding recursive scheme is:

$$\begin{cases} a_n^i(L) = e^{\beta \chi_{i-1,n+1}} \left[z_{i-2}(L - l_{i-1,n+1}) + z_{i-2}(L + l_{i-1,n+1}) \right] , \\ \\ z_n(L) = \sum_{i=1}^{n+1} a_n^i(L) , \end{cases} \tag{6.3}$$

where $\chi_{i,j}$ is minus the energy of the native stretch from i–th to j–th residue and the initial conditions are $z_{-1}(L) = 1$ for $L = 0$ and $z_{-1}(L) = 0$ for $L \neq 0$. The absolute value of the possible end–to–end lengths of a protein cannot be greater than $L_{\max} = \sum_{i=0}^{N} l_{i,i+1}$, which corresponds to the length of the molecule in the completely unfolded, fully extended configuration. Thus, because of finite resolution of amino acids coordinates in the pdb file (which is 10^{-3} Å), L belongs to a finite set of values in the range $[-L_{\max}, L_{\max}]$.

Let us also remember that the set of all possible lengths $\{l_{ij}\}$ is obtained directly from the three dimensional structure deposited in the Protein Data Bank (pdb) as the distances between the various pairs of central carbon atoms $\{C_{\alpha_i}, C_{\alpha_j}\}$. Besides l_{ij}, two more lengths associated to the stretch from the i–th to the j–th residue will be important in the following. These are the maximum $p_{ij}^{\max}$ and the minimum $p_{ij}^{\min}$ among the distances between C_{α_i} and the projections of each C_{α_k} $(i \leq k \leq j)$ on the straight line from C_{α_i} to C_{α_j}. Note that, as shown in figure 6.2 (axis $x2$), $p_{ij}^{\max} \geq l_{ij}$ and $p_{ij}^{\min} \leq 0$.

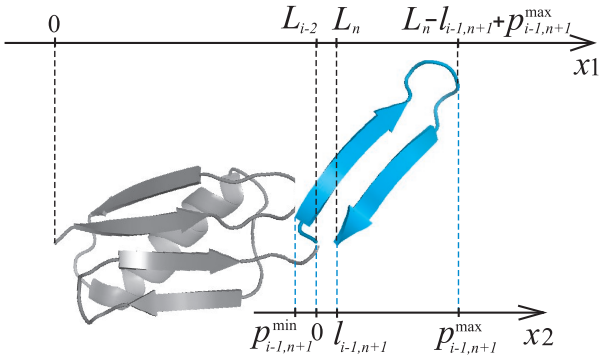


Figure 6.2: Sketch of a configuration with residue $m_{i-1} = 0$. Axis $x1$ shows relevant lengths of entire molecule. Axis $x2$ shows relevant lengths of native stretch from $(i-1)$–th to $(n+1)$–th residues.

Now, consider again the recursive scheme of equation (6.3) and set the starting point of the molecule in the middle of the cage. In order to confine the protein into a cage of size $2R$ with inert walls, when adding a native stretch from $(i-1)$–th to $(n+1)$–th residues (which are respectively at the distances L_{i-2} and L_n from the N–terminus), one has to require that every residue of this stretch lies inside the cage. This issue may be solved by considering also the lengths $p_{i-1,n+1}^{\max}$ and $p_{i-1,n+1}^{\min}$ of the native stretch (see axis $x1$ of figure 6.2) and inserting appropriate

step functions in the recursive scheme:

$$\begin{cases} a_n^i(L) &= e^{\beta\chi_{i-1,n+1}} \times \\ & \times \Big[z_{i-2}(L - l_{i-1,n+1})\, \mathcal{F}(R,\, L,\, l_{i-1,n+1},\, p^{\max}_{i-1,n+1},\, p^{\min}_{i-1,n+1}) + \\ & + z_{i-2}(L + l_{i-1,n+1})\, \mathcal{F}(R,\, L,\, -l_{i-1,n+1},\, -p^{\min}_{i-1,n+1},\, -p^{\max}_{i-1,n+1}) \Big] \ , \\ \\ z_n(L) = \sum_{i=1}^{n+1} a_n^i(L) \ , \end{cases} \tag{6.4}$$

where the function $\mathcal{F}$ is defined as

$$\mathcal{F}(R,\, L,\, l,\, x,\, y) \equiv \theta(R - L + l - x)\, \theta(R + L - l + y) \ ,$$

and θ is the Heaviside step function:

$$\theta(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

Translational freedom must also be taken into account. To this end, for a given configuration, instead of considering simply the end-to-end length, it would be better to consider as the relevant length the distance between the two farthest residues of that configuration. We call it the configuration effective length. Fixing in the center of the cage the N–terminus excludes from the partition functions $z_n(L)$ the contribution of some of the configurations which have an effective length shorter than $2R$ (for example in fig. 6.3a configuration $a1$ has an effective length shorter than configuration $a2$ but the former is forbidden while the latter is allowed).

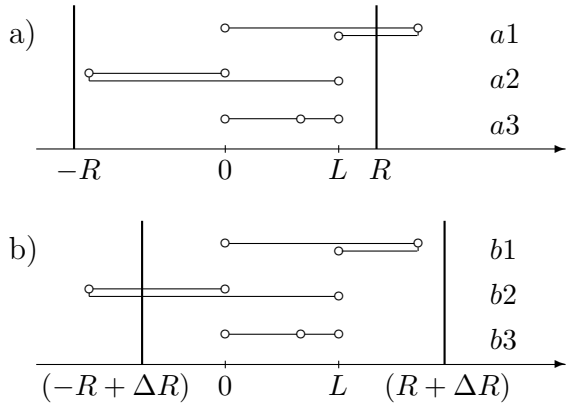


Figure 6.3: Three different configurations which would give a contribution to the partition function constrained at length L without any cage. With cage a only configurations 2 and 3 contribute. In b only configurations 1 and 3 contribute.

Thus, to take into account all the configurations with an effective length shorter than the cage size, the partition function has to be computed for different positions of the cage relative to the N–terminus. The final partition function will be the sum of various partition functions at different cage positions. Note that some configurations

will appear many times in such a scheme (for example state $a3$ of fig. 6.3a) as a consequence of their greater translational freedom.

To obtain the final partition function one has to repeat this procedure considering all the possible positions of the cage relative to the N–terminus, i.e. to start with the range $[-2R, 0]$ and to move the cage with a step ΔR equal to the resolution of the $\{l_{ij}\}$ until the final range $[0, 2R]$ is reached. To speed up computations we rounded the lengths to a resolution of 10^{-1} Å . For the 3–helix bundle we checked that this assumption does not modify the results through a comparison with results obtained at the resolution of 10^{-3} Å.

6.3 Results and discussion

As already mentioned in the introduction of this chapter, the confined WSME model has been used to study the effects of confinement, between two inert hard walls, of six different protein structures. Three are real structures: a 3–helix bundle (pdb code 1PRB), protein G (pdb code 2GB1) and its final hairpin. The other three structures are an ideal α–helix of ten residues (radius 2.3 Å, pitch 5.4 Å, $\varepsilon_{ij} = 1$ if $j = i + 4$ and $\varepsilon_{ij} = 0$ otherwise), a 2–stranded and a 3–stranded antiparallel β–sheets with 7 residues in each strand (the 3–stranded sheet is drawn in figure 6.4). In the following, code 'a010' refers to the ideal α–helix, 'b207' and 'b307' to the

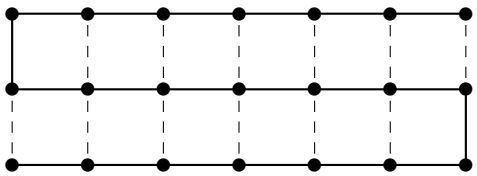


Figure 6.4: Ideal antiparallel β–sheet with 3 strands. Distance between two consecutive residues is 3.8 Å. Dashed lines represent native contacts. For them $\varepsilon_{ij} = 1$, while $\varepsilon_{ij} = 0$ in other cases.

two β–sheets which have respectively 2 and 3 strands and 'GB1h' refers to the final hairpin of protein G. Thermal stability of the considered proteins has been studied exploiting the property of the model to be exactly solvable at equilibrium, while Monte Carlo simulations have been used to study folding rates behavior.

6.3.1 Equilibrium

To study the equilibrium response to confinement of the six structures, thermodynamic quantities as the Helmholtz free energy, the specific heat and the average fraction of native residues have been computed at different cage sizes R. For each structure the distance $2R$ between the walls has been varied in a range from about the mininum effective length of the completely unfolded state to twice the maximum length of the completely unfolded state, i.e. from 4 Å (the distance between two subsequent amino acids is about 3.8 Å) to $2L_{\max}$.

Let us denote with $L_{\text{N eff.}}$ the effective length of the native state (values are reported in table 6.1). It is possible to make a naive distinction between two different

confinement regimes: (i) one, for $2R > L_{\text{N eff.}}$, which disallows the most expanded conformations of the non–native basin but not the folded state, and (ii) the strong confinement regime, for $2R < L_{\text{N eff.}}$, which forbids also the fully native state.

Table 6.1 also shows the effective length of the unfolded state $L_{\text{U eff.}}$. This is obtained through a Monte Carlo simulation at the unfolding temperature as the average effective length over the configurations belonging to the unfolded basin. Details about Monte Carlo moves will be given in the next section.

Table 6.1: Native state end–to–end length (L_{N}), effective length of the native state ($L_{\text{N eff.}}$), maximum length of the fully unfolded state ($L_{\max}$) and effective length of the unfolded state ($L_{\text{U eff.}}$) for the six different structures.

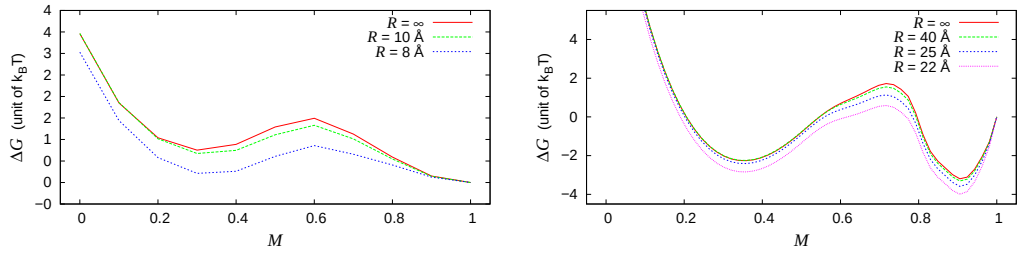
	a010	b207	b307	GB1h	1PRB	2GB1
L_{N} (Å)	14.3	3.8	24.0	6.5	40.0	27.8
$L_{\text{N eff.}}$ (Å)	14.3	3.8	24.0	6.6	40.0	29.1
$L_{\max}$ (Å)	34.2	49.4	76.0	63.8	201	212
$L_{\text{U eff.}}$ (Å)	14.1	21.8	27.3	24.1	40.5	46.3

Since, without confinement, for a given set of binary variables $\{m_k\}$, the model admits $2^{\sum_{i=1}^{N}(1-m_i)}$ configurations and this number grows exponentially with the amount of non–native residues, one may expect that confinement in a cage of size R, with $L_{\max} > 2R > L_{\text{N eff.}}$, gives a reduction of conformational entropy which affects more the non–native basin than the native one. Besides, one has to consider translational freedom whose role is to further stabilize the most compact configurations irrespectively of the fact that they belong or not to the native basin. Thus a structure with $L_{\text{N eff.}} > L_{\text{U eff.}}$, as in the case of the ideal α–helix , does not undergo any stabilization of the folded state. Figure 6.5 shows the free energy landscapes for the three real structures and for the ideal α–helix at different confinement sizes (for a better comparison the free energy of the completely folded state has always been set to zero). For the final hairpin of protein G, confinement increases the free energy of both the native and non–native basin: both native and non–native basins are destabilized but the latter is more affected. On the contrary, for the 3–helix bundle both native and non–native basins are stabilized, with a slightly greater stabilization upon confinement for the native state. For protein G, only the non–native basin is destabilized by confinement. Finally, it is possible to see that the native ideal α–helix is destabilized by confinement at $R = 10$ Å and $R = 8$ Åwith respect to the non–native state. Indeed, for the ideal α–helix, it has been found that such behavior already from radius of confinement lower than $R = 15$ Å and no enhancement in the unfolding temperature could be detected for greater values of R.

The increased stability of the native state relative to the unfolded state should result in higher unfolding temperature according to [145, 146]:

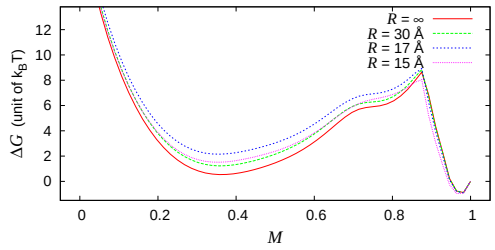
$$\frac{T_{\text{f}} - T_{\text{f}}^0}{T_{\text{f}}^0} \propto \left(\frac{2R}{L_{\text{N eff.}}}\right)^{-\gamma} \tag{6.5}$$

where here, and from now on, we denote with T_{f}^0 the unfolding temperature without

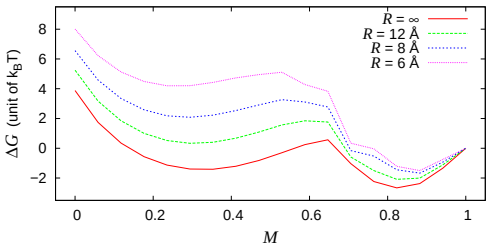


(a) *protein structure:* a010

(b) *protein structure:* 1PRB



(c) *protein structure:* 2GB1



(d) *protein structure:* GB1h

Figure 6.5: Free energy profile in function of the fraction of native residues M at various confinement radius R for the ideal α–helix (a), the 3–helix bundel (b), protein G (c) and its final hairpin (d). Free energy of completely native state ($M = 1$) have been setted to zero.

confinement. For each protein, we have determined T_f as the temperature at which the average fraction of native residues is such that $(M - M_\infty)/(M_0 - M_\infty) = 0.5$, where $M_\infty = 1/3$ is the value of M at infinite temperature and $M_0 \approx 1$ is its value at zero temperature.

Except for the ideal α–helix, other proteins exhibit an enhancement in their thermal stability to a different extent depending on their structure: the increase in unfolding temperatures is of few percents for the 3–stranded β–sheet, 3–helix bundle and protein G, while for the two β–hairpins $T_\mathrm{f} \simeq 6.6\, T_\mathrm{f}^0$ (ideal 2–stranded β–sheet) and $T_\mathrm{f} \simeq 2.7\, T_\mathrm{f}^0$ (final hairpin of protein G). Such drastically different behavior is due to the very short effective lengths of native states of the two hairpins and to the limitation of the model which projects the positions of all residues on a single direction and loses information on the real three–dimensional structure. For the 3–helix bundle and for protein G, the increases in unfolding temperature correspond respectively to about 1.5 K and 9.3 K.

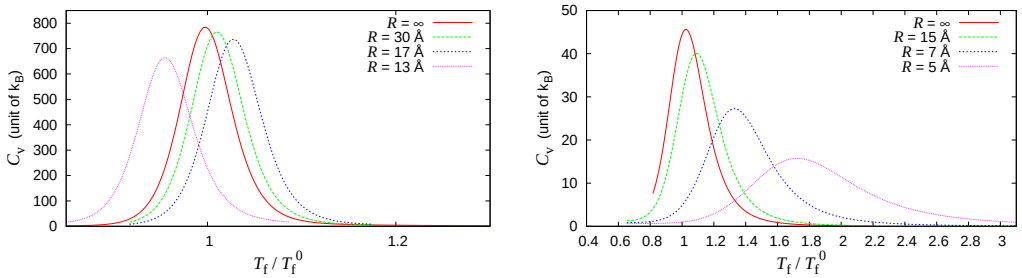
Values R_I^eq of the cage radius for which, at equilibrium, unfolding temperature reaches its maximum and the extent of enhancement are reported in table 6.2.

The enhancement in thermal stability can be appreciated in figure 6.6 where we reported the specific heat as a function of temperature. The top panel also shows well another feature of the unfolding phase transition in confined environment which is a decreased cooperativity with confinement [146].

A fit to equation 6.5 of unfolding temperatures as a function of R (figure 6.7) yielded exponents γ reported in table 6.2. All values are in between 1.50 (3–helix bundle) and 2.35 (final hairpin of protein G). Remarkably, in this range we find also

Table 6.2: Values of R for which unfolding temperature reaches its maximum ($T_{\mathrm{f}}^{\max}$) and the extent of enhancement. Values of γ from fits to equation 6.5 and fit ranges. Fits in ranges from $L_{\mathrm{U\ eff.}}/2$ to $L_{\max}$ for 'b207' and 'GB1h' result in exponents γ'.

	b207	b307	GB1h	1PRB	2GB1
$R_{\mathrm{I}}^{\mathrm{eq}}$ (Å)	2	17	3	25	17
$T_{\mathrm{f}}^{\max}$	$6.55\,T_{\mathrm{f}}^{0}$	$1.013\,T_{\mathrm{f}}^{0}$	$2.73\,T_{\mathrm{f}}^{0}$	$1.005\,T_{\mathrm{f}}^{0}$	$1.03\,T_{\mathrm{f}}^{0}$
γ	2.14 ± 0.03	1.57 ± 0.05	2.35 ± 0.03	1.50 ± 0.05	1.65 ± 0.04
fit range (Å)	$[4, 50]$	$[18, 76]$	$[4, 64]$	$[26, 201]$	$[18, 212]$
γ'	1.72 ± 0.06		1.60 ± 0.07		
fit range (Å)	$[10.9, 50]$		$[12.05, 64]$		



(a) 2GB1 (b) GB1h

Figure 6.6: Specific heat $C_V = \frac{1}{k_B T^2} \frac{\partial^2 Z}{\partial \beta^2}$ as a function of the temperature at various confinement radius R for protein G and its final hairpin.

the theoretical values of γ for an excluded volume chain confined in a slit or in a cylinder ($\gamma = 5/3$) and for a gaussian chain in a slit, a cylinder or a sphere ($\gamma = 2$).

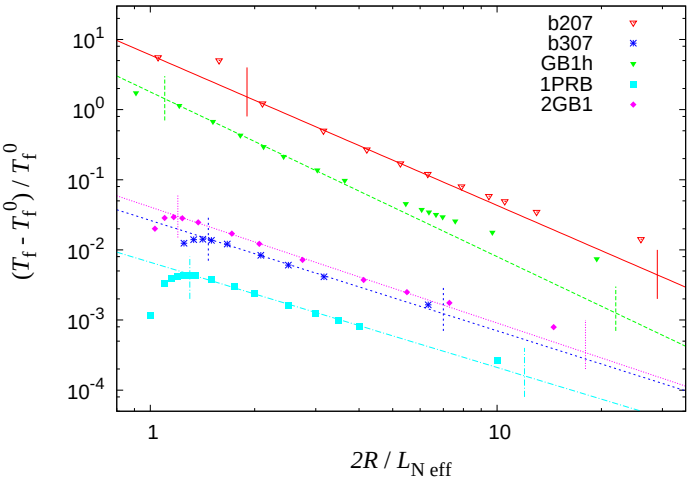


Figure 6.7: Shift in unfolding temperature as a function of confining cage radius R. Fits to equation 6.5 in ranges reported in table 6.2. The vertical lines represent the ranges spanned by fits.

Furthermore, a more careful analysis of data in figure 6.7 suggested us to fit, in the case of the β–hairpins, also in a more limited range of R values going from $L_{U\ eff.}/2$ to L_{max} (figure 6.8). In this very low confinement regime $\gamma = 1.72$ for the ideal hairpin and $\gamma = 1.6$ for the final hairpin of protein G.

Notwithstanding the simplicity of the model and its unidimensionality, the obtained results follow the general trend of previous experimental studies [140, 31, 32, 42] and simulations [145, 146, 147]: provided the native state is compact, when reducing the space available to a given protein, unfolding temperature T_f grows until a certain confinement size which depends on the protein. If the confinement size is further decreased, unfolding temperature decreases. Furthermore, the results also support the theoretical prediction [142, 146, 147] that enhancement depends on the confinement size R by the scaling law $\Delta T_f \sim R^{-\gamma}$. For the five structures which show enhanced thermal stability, we found that exponents γ lie in between the upper and lower values of 2.35 and 1.5.

6.3.2 Kinetics

The folding kinetics have been studied by Monte Carlo (MC) simulations in which a 2–components ternary variable (m_k, s_k) have been associated to each residue k. If $m_k = 1$, $s_k = 0$ while if $m_k = 0$, $s_k = \sigma_{kj} = \pm 1$ is the direction of the native stretch from the k–th to the j-th residue. A single MC step consists in choosing a residue k with uniform probability among the N residues and changing (m_k, s_k) variable with equal probability to any of its other two states. This move is alternated with a 0.1 Å translation of the entire protein to the left or to the right with equal probability. Few remarks are necessary: suppose to have a native stretch from the

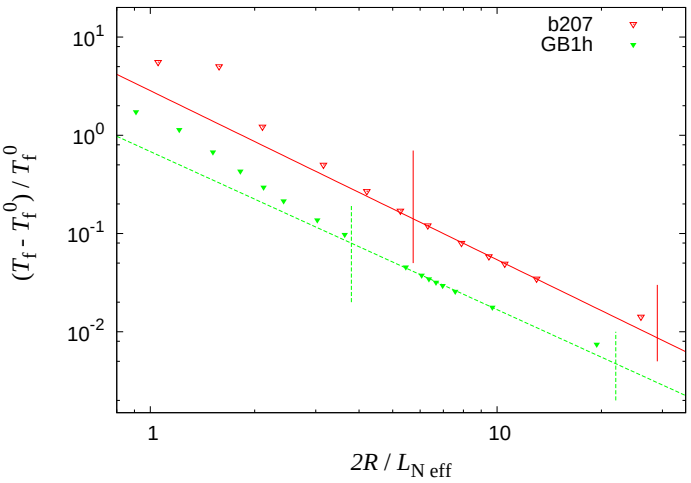


Figure 6.8: Shift in unfolding temperature as a function of confining cage radius R for 'b207' and 'GB1h'. Fits to equation 6.5 in ranges $L_{U\ eff.}/2$ to L_{max}. The vertical lines represent the ranges spanned by fits.

i–th to the j-th residue and to transform the variable (m_k, s_k), $i < k < j$, from $(1, 0)$ to $(0, s_k = \pm 1)$. The direction of the new native stretch from the k–th to the j–th residue will be determined by s_k while the new native stretch from i to k will inherit the direction of the old one from i to j. If instead the state of k–th residue is moved from $(0, \pm 1)$ to $(1, 0)$, two native stretches merge into one with direction equal to the direction of the first old native stretch. At each MC step confinement requirements must be checked.

The acceleration of folding has been estimated [147] to follow the scaling law:

$$\ln\left(\frac{k_f}{k_f^0}\right) \propto \left(\frac{2R}{L_{N\ eff.}}\right)^{-\gamma} , \tag{6.6}$$

where k_f^0 denotes the folding rate in the $R \to \infty$ limit. The idea is that, since confinement makes the extended unstructured conformations inaccessible, the increase in k_f is due to the restricted search for the native state among the remaining configurations. Furthermore one has to consider that the increase in folding temperature also biases the system towards the native state.

In our Monte Carlo simulation we determined folding rates as the inverse of mean first passage times by using 10^4 folding trajectories. First passage time is defined as the time at which, starting from a random unfolded configuration, the weighted fraction of native contacts,

$$\varphi = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \varepsilon_{ij}\Delta_{ij}\prod_{k=i}^{j} m_k}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \varepsilon_{ij}\Delta_{ij}} , \tag{6.7}$$

catches up with the threshold 0.9, which ensures the protein has reached the folded state and has not got stuck in some intermediate. Temperature has been set to $0.9\,T_f^0$ in order to speed up the simulations.

Table 6.3: Values of R for which the folding rate reaches its maximum k_f^{max} at $T = 0.9\,T_f^0$ and the extent of enhancement. Values of γ from fits to equation 6.6 in the reported ranges.

	b207	b307	GB1h	1PRB	2GB1
R_I^{kin} (Å)	19	23	12	19	18
k_f^{max}	$1.13\,k_f^0$	$1.13\,k_f^0$	$1.46\,k_f^0$	$1.50\,k_f^0$	$2.35\,k_f^0$
γ	1.42 ± 0.20	1.53 ± 0.33	1.54 ± 0.11	1.71 ± 0.08	1.67 ± 0.07
fit range (Å)	$[19, 50]$	$[23, 76]$	$[14, 64]$	$[22, 201]$	$[20, 212]$

Again, among the six different protein structures studied, the 10–residues ideal α–helix does not show any enhancement of folding rate, because its native state cannot be considered compact if compared to the average unfolded state. For the other structures, when decreasing R, folding is accelerated until a certain size R_I^{kin} is reached, then folding rates start to decrease. Table 6.3 reports R_I^{kin} values and the maximum extent of folding rates enhancement. For the β–hairpins, the drastic difference between R_I^{kin} and R_I^{eq} is likely due to the fact that for very small confining cages, even if the native state is not compromised, the structure is squeezed so much that chain reconfigurations towards the folded state become difficult. The same reason should explain the small differences between R_I^{kin} and R_I^{eq} of other structures.

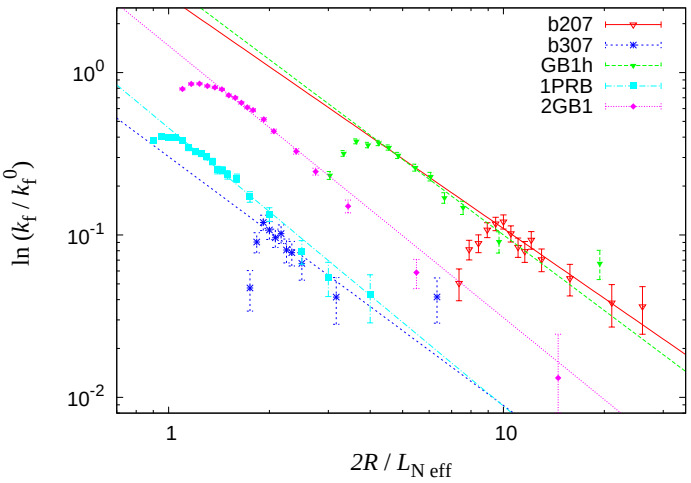


Figure 6.9: Shift in folding rates at $T = 0.9\,T_f^0$ as a function of confining cage radius R. Fits to equation 6.6 in ranges reported in table 6.3.

Table 6.3 also reports the γ values obtained through a fit to equation 6.6 while figure 6.9 shows the folding rates behavior together with fit lines. Exponents γ for enhancement of folding rate, lie in between the upper and lower values of 1.71 and 1.42 and are comparable with their equilibrium counterparts, especially if one considers the very low confinement regime for the two hairpins.

Theoretical values of γ ($\gamma = 5/3$ for a chain with excluded volume confined into a slit or a cylinder and $\gamma = 2$ for a gaussian chain into a slit, a cylinder or a sphere) are not directly comparable to the results of WSME model, which differs from these theories both for the geometry (our chain is neither self-avoiding nor gaussian) and for the presence of specific interactions, which are neglected by these theories. Nevertheless, the obtained results, both from thermodynamics and kinetics, for γ, are in the same range as the theoretical ones.

Finally it is worth to stress again that, for a 3–helix bundle and for protein G, WSME results are consistent with those obtained through a more realistic model by Best and Mittal [147] for confinement of the same proteins into a slit: γ values are consistent and also the maximum enhancements of folding temperatures and folding rates are in good accordance. The two model also agree in the fact that protein G is more affected by confinement but there is no accordance on the confinement radius at which the 3–helix bundle reaches its maximum folding temperature and its maximum folding rate.

Conclusions

In this Thesis a generalization of the WSME model, suitable to handle the mechanical unfolding phenomenon, has been used to study into details the mechanical unfolding of two real proteins, namely the 10th type III module of fibronectin ($FnIII_{10}$) and the wild type Green Fluorescent Protein (GFP). The same version of the WSME has been further modified in order to deal with confinement of proteins between two inert walls.

The WSME model, introduced first by Wako and Saitô and then reconsidered by Muñoz and Eaton, is an Ising–like model, where a binary variable, able to distinguish between a native and an unfolded conformation, is associated to each amino acid of a protein. Then, in its generalization treated in this Thesis, another variable specifies the direction of native stretches, i.e. of each sequence of consecutive native residues delimited by two non–native residue. We did the simplest choice of assuming only two possible direction, parallel or anti–parallel with respect to an external force, thus making this second one a spin–like variable. The loss of entropy, due to fixing peptide units in their native conformation, is taken into account by considering that, if a given residue is native, the total number of configurations is twice smaller than in the case in which that residue is non–native. The free energy of the model is designed by considering only native interactions and associating an energetic contribution to pairs of native residues belonging to native strings. A protein length dependent potential includes the coupling to the external force while the configurations space of the residue and spin variables account for the above entropic considerations.

Exploiting the fact that equilibrium thermodynamics of the model is exactly solvable, we have obtained equilibrium properties, such as the free energy landscape, of $FnIII_{10}$ and GFP when a constant force is applied to their ends. For both proteins, at the critical unfolding force the free energy landscape as a function of the end–to–end length showed few local minima, besides those associated with the native and the fully unfolded states. These local minima reflect the existence of intermediate states in the mechanical unfolding pathways that must be confirmed by considering the nonequilibrium unfolding kinetics. To this purpose, we run Monte Carlo (MC) simulations mimicking the pulling both at constant force and at constant velocity protocol. Thanks to the simplicity of the model we could probe force and speed ranges close to *in vivo* and experimental conditions, which was not possible in most previous simulations with more detailed models because of their high computational costs.

$FnIII_{10}$ is made by the packing of two antiparallel β–sheets. The β–strands are usually denoted with letters from A (the strand closest to the N–terminal) to G (the C–terminal one). The two sheets are made of strands ABE and DCFG. At high

enough constant force we observed two–state transitions only. At smaller forces and at all pulling speeds considered we observed several intermediates, denoted by A, G, AG and GF, based on the strands which are unfolded in each intermediate. Possible unfolding pathways are summarized in figure 5.4 for the constant force protocol and in figure 5.9 for the constant pulling speed protocol. Interestingly, the unfolding pathways depend on the applied force or on the pulling speed, which was already observed in ref. [104]. Such pathways become more complex at low forces and speeds, due to the increase in fluctuations. Previous simulations and experiments showed some discrepancies in the unfolding pathways, and our work is not going to resolve such discrepancies, but some general trends are confirmed. In particular the most frequently observed intermediate in our trajectories was AG, which was also observed in all previous simulations [109, 111, 112, 113, 104]. In addition, constant pulling speed trajectories always visit an intermediate with the C–terminal β–strand detached, which was also observed in most previous simulations [109, 112, 113, 104] and in AFM experiments [67]. On the other hand, we have never observed intermediate AB, which has been reported in many simulations [111, 113, 104] and experiments [67]. We have instead observed, at low enough forces and speeds, intermediates A and GF, which were previously reported only by Gao *et al.* [113] (A only) and Mitternacht *et al.* [104] (both A and GF). These intermediates have end–to–end lengths close to G and AG, respectively, and cannot be distinguished in the usual one–dimensional free energy landscape using the end–to–end length as a reaction coordinate. It is worth noting that in our trajectories we observe fluctuations between intermediates with similar lengths, that is between A and G or between AG and GF. Fluctuations between AG and GF, in particular, are observed in most trajectories at the lowest forces and pulling speeds we have considered, and therefore one could speculate that they have some biological significance.

From a more quantitative point of view, given the extreme simplicity of our model, it is remarkable that many quantities we can compute agree well with the results from AFM experiments or previous simulations with similar parameters. Our estimate for the native state unfolding length is $x_u = 0.34 \pm 0.01$ nm, to be compared with $x_u = 0.38$ nm from AFM results [110] and with $x_u = 0.4$ nm from the simulations by Mitternacht *et al.* [104]. The average rupture force we obtained for the native state is in the range 80 to 100 pN, to be compared with results from 75 to 100 pN reported by AFM studies [110, 67], and from 88 to 114 pN in the simulations by Mitternacht *et al.* [104]. Finally, our intermediate G has an average rupture force between 40 and 50 pN, to be compared with 50 pN found in experiments [67], though it must be mentioned that in such work the intermediate might be an average between the G and AB intermediates.

GFP is a protein constituted by 11 β–strands arranged in a barrel structure, with a short α–helix at the beginning of the polypeptide chain and other short α–helices along the barrel axis. Also in this case, WSME model managed to mimic unfolding pathways of GFP pulled at constant velocity and intermediates consistent with experiments [66, 126]. We found that unfolding proceeds through two different pathways: in the major one, at first, the N–terminal β–strand unravels with a length of the corresponding intermediate of about $10 \div 12.5$ nm, then also the second and third N–terminal β–strands unravel and the molecule corresponding intermediate length is around 20 nm. In the minor unfolding pathway the first strand to break

is the C–terminal one, later followed by the rupture of at least another β–strand (usually the 10–th one) before complete unfolding. Actually the N–terminal α–helix is the first secondary structure element to unravel in both pathways but this event is typically associated with very small signals, almost masked by fluctuations.

WSME model also describes correctly the most important qualitative aspects of the direction–dependent mechanical unfolding of the GFP, namely the orders of magnitude and ranking of the unfolding forces corresponding to different pulling directions [62] but, from a more quantitative point of view, our energy barriers and unfolding forces are systematically larger than those observed in experiments [62]. Furthermore, we have exploited the dependence of the unfolding force on the pulling direction to investigate a force sensor based on a GFP polyprotein where each module is linked with a different geometry to the nearest neighbouring modules, so as to experience the force along different direction, yielding a device whose luminescence depends (in a discrete way) on the force. It is worth noting that such a device may be used in *in vivo* experiments, to measure forces at molecular level, e.g. inside cells, in a non–invasive way.

To introduce confinement of a protein into a slit, we resorted to the fact that the WSME model for mechanical unfolding under a constant force allows an exact solution that involves a recursive scheme that builds the partition function adding a residue at each step. This scheme can be expanded in powers of $e^{\beta f L}$, where f and β are respectively the external force and the inverse temperature, and L is the end–to–end length of the protein. In such a way it is possible to obtain a recursive scheme whose final results is the partition function as a function of the end–to–end length and which does not depend on the force f. Suitably amending the expanded recursive scheme it is possible to introduce the confinement, still maintaining the possibility to obtain an exact analytical solution. Making use of such an exact solution and resorting to MC simulations, we have respectively studied the thermal stability and the kinetic properties upon confinement of three ideal and three real protein structures.

From a theoretical point of view, confinement have been studied by modelling the native state as a rigid sphere and the non–native state as a polymer chain [142,143]. The effect of confinement is to forbid the most expanded configurations of the polymer chain, thus involving an entropy–based increase in the free energy of the unfolded state, which has consequences on both thermodynamics and kinetics of folding. A direct implication of such reduced stability of the non–native state is that melting temperatures (T_f) and folding rates (k_f) of proteins should follow the scaling law $\Delta T_\mathrm{f} \sim \Delta \ln k_\mathrm{f} \sim R^{-\gamma}$, where R is the typical size of confinement. Notwithstanding the simplicity of the model and its unidimensionality, the obtained results follow the general trend of previous experimental simulations [145,146,147] confirming the above scaling law when the size of confinement is not too low. In fact, changing the distance $2R$ between the walls, we found two confinement regimes: provided the native state is compact, starting from large R and decreasing R, confinement first enhances the stability of the folded state until a given value of R; then a further decrease of R leads to a decrease of folding temperature and folding rate. We found that in the low confinement regime both unfolding temperatures and logarithm of folding rates scale as $R^{-\gamma}$ where γ values lie in between 1.42 and 2.35 according to the considered protein. The γ values obtained for unfolding temperatures from

exact solutions at equilibrium are consistent with those found for folding rates enhancement by MC simulations. Finally, for a 3–helix bundle and for protein G, which are two of the real proteins we have considered, our results are consistent with those obtained through a more realistic model by Best and Mittal [147] for confinement of the same proteins into a slit: γ values are consistent and also the maximum enhancement extents of folding temperatures and folding rates are in good accordance. The two models also agree in the fact that protein G is more affected by confinement with respect to the 3–helix bundle.

To conclude this Thesis, we note that, as it generally happens for other Gō–type models, WSME model, notwithstanding its simplicity, is a powerful tool to investigate protein folding. In particular, for mechanical pulling of proteins and their confinement, we showed that it is able to capture the basic physics of folding and to reproduce, to a great extent, experimental results and achievements of simulations based on more detailed models. Major limitations of the model are connected to its unidimensionality but these limitations do not generally affect the validity of the results. However, one can, in principle, think to clear this hurdle by enlarging the number of the possible directions of a native stretch. When considering mechanical unfolding, this solution introduces some problems because it is necessary to introduce additional and not trivial constraints in the configurations space in order to keep the two residues to which the force is applied along the force axis. To investigate confinement, being the force formally set equal to zero, such a complication does not occur. One can, for example, go through the choice of allowing six native stretch directions, two for each cartesian axis. In this way MC simulation would be easy to perform and the model is, in principle, still exactly solvable but, *de facto*, computationally unmanageable because of a too high RAM requirement.

Another possibility to cure unidimensionality would be to substitute, for each residue k, the single binary variable with a pair of degree of freedom associated with the dihedral angles (ϕ_k, ψ_k) and to consider the residue as native if both ϕ_k and ψ_k assume their native value. Values of the pair (ϕ_k, ψ_k) can be limited to the native angles plus the typical angles in Ramachandran plot, i.e. to one native and about three non–native configurations. In this way the possibility to exploit the exact solution of the original model is lost but three–dimensionality of the chain is gained allowing configurations more similar to the real ones.

Bibliography

[1] T.E. Creighton. *Proteins: structures and molecular properties*. Freeman, New York., 1983.

[2] Selkoe D.J. Folding proteins in fatal ways. *Nature*, 426:900, 2003.

[3] F. Chiti and C.M. Dobson. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.*, 75:333, 2006.

[4] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44, 1968.

[5] C. Levinthal. How to fold graciously. *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.*, 1969.

[6] A.R. Fersht. Nucleation mechanisms in protein folding. *Curr. Op. Struct. Biol.*, 7:3, 1997.

[7] A.R. Fersht. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Op. Struct. Biol.*, 5:79, 1995.

[8] T.E. Creighton. *Protein Folding*. Freeman, New York, 1992.

[9] A.V. Finkelstein and O.V. Galzitskaya. Physics of protein folding. *Physics of Life Reviews*, 1:23, 2004.

[10] G. Binnig, C.F. Quate, and Ch. Gerber. Atomic Force Microscope. *Phys. Rev. Lett.*, 56:930, 1986.

[11] Simmons R.M., Finer J.T., Chu S., and Spudich J.A. Quantitative measurements of force and displacement using an optical trap. *Biophys. J.*, 70:1813, 1996.

[12] S.B. Fowler, R.B. Best, J.L. Toca Herrera, T.J. Rutherford, A. Steward, E. Paci, M. Karplus, and J. Clarke. Mechanical unfolding of a Titin Ig Domain: Structure of unfolding intermediate revealed by combining AFM, Molecular Dynamics Simulations, NMR and Protein Engineering. *J. Mol. Biol.*, 322:841, 2002.

[13] R.L. Baldwin. How does protein folding get started? *Trends Biochem Sci.*, 14:291, 1989.

[14] M. Karplus and D.L. Weaver. Protein–folding dynamics. *Nature*, 260:404, 1976.

[15] S.C. Harrison and R. Durbin. Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. USA*, 82:4028, 1985.

[16] M. Schaefer, C. Bartels, and M. Karplus. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.*, 284:835, 1998.

[17] Y. Duan and P.A. Kollman. Pathways to a Protein Folding Intermediate Observed in a 1–Microsecond Simulation in Aqueous Solution. *Science*, 282:740, 1998.

[18] Shea J.-E. and Brooks III C.L. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Re. Phys. Chem.*, 52:499, 2001.

[19] N. Gō and H. Taketomi. Reversible unfolding of Individual Titin Immunoglobulin domains by AFM. *Proc. Natl. Acad. Sci. U.S.A.*, 75:559, 1978.

[20] K.A. Dill, S. Bromberg, K. Yue, K.M. Ftebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein folding – A perspective from simple exact models. *Protein Sci.*, 4:561, 1995.

[21] A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of Protein Folding: A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.*, 235:1614, 1994.

[22] Veitshans T., Klimov D., and Thirumalai D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence–dependent properties. *Fold. Design*, 2:1, 1997.

[23] A. Kolinski and J. Skolnick. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. R.G. Landes Company, 1996.

[24] H. Wako and N. Saitô. Statistical Mechanical Theory of the Protein Conformation. I. General Considerations and the Application to Homopolymers. *J. Phys. Soc. Jpn*, 44:1931, 1978.

[25] H. Wako and N. Saitô. Statistical Mechanical Theory of the Protein Conformation. II. Folding Pathway for Protein. *J. Phys. Soc. Jpn*, 44:1939, 1978.

[26] V. Muñoz, P.A. Thompson, J. Hofrichter, and W.A. Eaton. Folding dynamics and mechanism of β–hairpin formation. *Nature*, 390:196, 1997.

[27] V. Muñoz, E.R. Henry, J. Hofrichter, and W.A. Eaton. A statistical mechanical model for β–hairpin kinetics. *Proc. Natl. Acad. Sci. USA*, 95:5872, 1998.

[28] V. Muñoz and W.A. Eaton. A simple model for calculating the kinetics of protein folding from three–dimensional structures. *Proc. Natl. Acad. Sci. USA*, 96:11311, 1999.

[29] P. Bruscolini and A. Pelizzola. Exact solution of the Muñoz–Eaton model for protein folding. *Phys. Rev. Lett.*, 88:258101, 2002.

[30] A. Imparato, A. Pelizzola, and M. Zamparo. Ising–like model for protein mechanical unfolding. *Phys. Rev. Lett.*, 98:148102, 2007.

[31] R. Ravindra, S. Zhao, H. Gies, and R. Winter. Protein encapsulation in mesoporous silicate: The effects of confinement on protein stability, hydration, and volumetric properties. *J. Am. Chem. Soc.*, 126:12224, 2004.

[32] D. Bolis, A.S. Politou, G. Kelly, A. Pastore, and P.A. Temussi. Protein stability in nanocages: A novel approach for influencing protein stability by molecular confinement. *J. Mol. Biol.*, 336:203, 2004.

[33] M.S. Cheung, D.K. Klimov, and D. Thirumalai. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci. USA*, 102:4753, 2005.

[34] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland, New York, 2002.

[35] Sumner J.B. The isolation and crystallization of the enzyme urease. Preliminary paper. *J. Biol. Chem.*, 69:435, 1926.

[36] A. Radzicka and R. Wolfenden. A proficient enzyme. *Science*, 267:90, 1995.

[37] http://www.pdb.org/pdb/home/home.do.

[38] B.K. Ho, A Thomas., and R. Brasseur. Revisiting the Ramachandran plot: Hard–sphere repulsion, electrostatics, and h–bonding in the α–helix. *Protein Sci.*, 12:2508, 2003.

[39] B.K. Ho and R. Brasseur. The Ramachandran plots of glycine and pre–proline. *BMC Structural Biology*, 5:14, 2005.

[40] Sanger F. The arrangement of amino acids in proteins. *Adv. Protein Chem.*, 7:1, 1952.

[41] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223, 1973.

[42] Y-C. Tang, H-C. Chang, A. Roeben, D. Wischnewski, N. Wischnewski, M.J. Kerner, F.U. Hartl, and M. Hayer-Hartl. Structural features of the GroEL–GroES Nano–Cage required for rapid folding of encapsulated protein. *Cell*, 125:903, 2006.

[43] Editorial:. So much more to know. *Science*, 309:78, 2005.

[44] H. Frauenfelder, F. Parak, and R.D. Young. Conformational Substates in Proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451, 1988.

[45] A. Ansari, J. Berendzen, S.F. Bowne, H Frauenfelder., I.E. Iben, T.B. Sauke, E. Shyamsunder, and R.D. Young. Protein states and proteinquakes. *Proc. Natl. Acad. Sci.*, 82:5000, 1985.

[46] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.*, 35:539, 2010.

[47] Falke J.J. A moving story. *Science*, 295:1480, 2002.

[48] K.A. Dill and D. Shortle. Denatured states of proteins. *Annu. Rev. Biochem.*, 60:795, 1991.

[49] O.B. Ptitsyn. Molten Globule and Protein Folding. *Adv. Protein Chem.*, 47:83, 1995.

[50] V.S. Pande and S. Rokhsar. Is the molten globule a third phase of proteins? *Proc. Natl. Acad. Sci.*, 95:1490, 1998.

[51] Southall N.T., Dill K.A., and Haymet A.D.J. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106:521, 2002.

[52] K. Sneppen and G. Zocchi. *Physics in Molecular Biology*. Cambridge University Press, 2005.

[53] P.L. Privalov. Stability of proteins: Proteins which do not present a single cooperative system. *Adv. Protein Chem.*, 35:1, 1982.

[54] Zwanzig R., Szabo A., and Bagchi B. Levinthal's paradox. *Proc. Natl. Acad. Sci USA*, 89:20, 1992.

[55] S.E. Jackson. How do small single–domain proteins fold? *Fold. Design*, 3:R81, 1998.

[56] Bryngelson J.D. and Wolynes P.G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci*, 84:7524, 1987.

[57] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284, 1940.

[58] K.W. Plaxco, K.T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277:985, 1998.

[59] A. Matouschek, J.T. Kellis Jr., L. Serrano, and A.R. Fersht. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, 340:122, 1989.

[60] Sánchez I.E. and Kiefhaber T. Origin of unusual ϕ–values in Protein Folding: Evidence against specific nucleation sites. *J. Mol. Biol.*, 334:1077, 2003.

[61] H. Dietz, M. Bertz, M. Schlierf, F. Berkemeier, T. Bornschlögl, J.P. Junker, and M. Rief. Cysteine engineering of polyproteins for single–molecule force spectroscopy. *Nat. Protocols*, 1:80, 2006.

[62] H. Dietz, F. Berkemeier, M. Bertz, and M. Rief. Anisotropic deformation response of single protein molecules. *Proc. Natl. Acad. Sci.*, 103:12724, 2006.

[63] G.E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721, 1999.

[64] C. Jarzynski. Nonequilibrium Equality for Free Energy differences. *Phys. Rev. Lett.*, 78:2690, 1997.

[65] M. Rief, M. Gautel, F. Oesterhelt, J.M. Fernandez, and H.E. Gaub. Reversible unfolding of Individual Titin Immunoglobulin domains by AFM. *Science*, 276:1109, 1997.

[66] H. Dietz and M. Rief. Exploring the energy landscape of GFP by single–molecule mechanical experiments. *Proc. Natl. Acad. Sci.*, 101:16192, 2004.

[67] L. Li, H.H-L. Huang, C.L. Badilla, and J.M. Fernandez. Mechanical unfolding intermediates observed by Single–molecule Force Spectroscopy in a Fibronectin type III module. *J. Mol. Biol.*, 345:817, 2005.

[68] S. Kumar and M.S. Li. Biomolecules under mechanical force. *Physics Reports*, 486:1, 2010.

[69] J. Brujić, Z.R.I. Hermans, K.A. Walther, and J.M. Fernandez. Single–molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin. *Nature Physics*, 2:282, 2006.

[70] C.Cecconi, E.A. Shank, S. Marqusee, and C. Bustamante. DNA molecular handles for single–molecule protein–folding studies by optical tweezers. *Methods Mol. Biol.*, 749:255, 2011.

[71] G.I. Bell. Models for the specific adhesion of cells to cells. *Science*, 200:618, 1978.

[72] E. Evans and K. Ritchie. Dynamic strength of molecular adhesion bonds. *Biophys. J.*, 72:1541, 1997.

[73] E. Evans and K. Ritchie. Strength of a Weak Bond Connecting Flexible Polymer Chains. *Biophys. J.*, 76:2439, 1999.

[74] M. Carrion-Vazquez, A.F. Oberhauser, S.B. Fowler, P.E. Marszalek, S.E. Broedel, J. Clarke, and J.M. Fernandez. Mechanical and chemical unfolding of a single protein: A comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 96:3694, 1999.

[75] G. Hummer and A. Szabo. Kinetics from Nonequilibrium Single–Molecule Pulling Experiments. *Biophys. J.*, 85:5, 2003.

[76] O.K. Dudko, A.E. Filippov, J. Klafter, and M. Urbakh. Beyond the conventional description of dynamic force spectroscopy of adhesion bonds. *Proc. Natl. Acad. Sci. USA*, 100:11378, 2003.

[77] O.K. Dudko, G. Hummer, and A. Szabo. Intrinsic Rates and Activation Free Energies from Single–Molecule Pulling Experiments. *Phys. Rev. Lett.*, 96:108101, 2006.

[78] M. Raible, M. Evstigneev, P. Reimann, F.W. Bartels, and R. Ros. Theoretical analysis of dynamic force spectroscopy experiments on ligand–receptor complexes. *J. Biotechnol.*, 112:13, 2004.

[79] A. Irbäck and S. Mohanty. Folding Thermodynamics of Peptides. *Biophys. J.*, 88:1560, 2005.

[80] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235:318, 1987.

[81] A. Pelizzola. Exactness of the cluster variation method and factorization of the equilibrium probability for the Wako–Saitô–Muñoz–Eaton model of protein folding. *J. Stat. Mech.: Theory Exp.*, page P11010, 2005.

[82] M. Zamparo and A. Pelizzola. Kinetics of the Wako–Saitô–Muñoz–Eaton model of protein folding. *Phys. Rev. Lett.*, 97:068106, 2006.

[83] M. Zamparo and A. Pelizzola. Rigorous results on the local equilibrium kinetics of a protein folding model. *J. Stat. Mech.: Theory Exp.*, page P12009, 2006.

[84] P. Bruscolini, A. Pelizzola, and M. Zamparo. Downhill versus two–state protein folding in a statistical mechanical model. *J. Chem. Phys.*, 126:215103, 2007.

[85] P. Bruscolini, A. Pelizzola, and M. Zamparo. Rate determining factors in protein model structures. *Phys. Rev. Lett.*, 99:038103, 2007.

[86] A. Imparato, A. Pelizzola, and M. Zamparo. Protein mechanical unfolding: A model with binary variables. *J. Chem. Phys.*, 127:145105, 2007.

[87] A. Imparato and A. Pelizzola. Mechanical unfolding and refolding pathways of ubiquitin. *Phys. Rev. Lett.*, 100:158104, 2008.

[88] A. Imparato, A. Pelizzola, and M. Zamparo. Equilibrium properties and force–driven unfolding pathways of RNA molecules. *Phys. Rev. Lett.*, 103:188102, 2009.

[89] M. Caraglio, A. Imparato, and A. Pelizzola. Pathways of mechanical unfolding of FnIII_{10}: Low force intermediates. *J. Chem. Phys.*, 133:065101, 2010.

[90] M. Caraglio, A. Imparato, and A. Pelizzola. Direction dependent mechanical unfolding and green fluorescent protein as a force sensor. *Phys. Rev. E*, 84:021918, 2011.

[91] M. Faccin, P. Bruscolini, and A. Pelizzola. Analysis of the equilibrium and kinetics of the ankyrin repeat protein myotrophin. *J. Chem. Phys.*, 134:075102, 2011.

[92] K. Itoh and M. Sasai. Dynamical transition and proteinquake in photoactive yellow protein. *Proc. Natl. Acad. Sci.*, 101:14736, 2004.

[93] M. Zamparo, A. Trovato, and A. Maritan. Simplified exactly solvable model for β–amyloid aggregation. *Phys. Rev. Lett.*, 105:108102, 2010.

[94] V.I. Tokar and H. Dreyssé. Exact solution of a one–dimensional model of strained epitaxy on a periodically modulated substrate. *Phys. Rev. E*, 71:031604, 2005.

[95] M. Zamparo. An exactly solvable model for a β–hairpin with random interactions. *J. Stat. Mech.: Theory Exp.*, page P10013, 2008.

[96] E. Evans. Probing the relation between force–lifetime–and chemistry in single molecular bonds. *Annu. Rev. Biophys. Biomol. Struct.*, 30:105, 2001.

[97] M. Zamparo. Ph.D. thesis, Politecnico di Torino, 2009.

[98] F. Rossi. *Theory of Semiconductor Quantum Devices*. Springer Heidelberg Dordrecht London New York, 2011.

[99] J.Liphardt, B. Onoa, S.B. Smith, I. Tinoco Jr., and C. Bustamante. Reversible Unfolding of Single RNA Molecules by Mechanical Force. *Science*, 292:733, 2001.

[100] B. Onoa, S. Dumont, J. Liphardt, S.B. Smith, I. Tinoco Jr., and C. Bustamante. Identifying Kinetic Barriers to Mechanical Unfolding of the T. thermophila Ribozyme. *Science*, 299:1892, 2003.

[101] A. Imparato, F. Sbrana, and M. Vassalli. Reconstructing the free–energy landscape of a polyprotein by single–molecule experiments. *Europhys. Lett.*, 82:58006, 2008.

[102] A. Imparato, S. Luccioli, and A. Torcini. Reconstructing the free energy landscape of a mechanically unfolded model protein. *Phys. Rev. Lett.*, 99:168101, 2007.

[103] B.T. Andrews, S. Gosavi, J.M. Finke, J.N. Onuchic, and P.A. Jennings. The dual–basin landscape in GFP folding. *Proc. Natl. Acad. Sci. USA*, 105:12283, 2008.

[104] S. Mitternacht, S. Luccioli, A. Torcini, A. Imparato, and A. Irbäck. Changing the Mechanical Unfolding Pathway of $FnIII_{10}$ by Tuning the Pulling Strength. *Biophys. J.*, 96:429, 2009.

[105] B. Geiger, A. Bershadsky, R. Pankov, and K.M. Yamada. Transmembrane crosstalk between the extracellular matrix and the cytoskeleton. *Nat. Rev. Mol. Cell. Biol.*, 2:793, 2001.

[106] Pankov R. and Yamada K.M. Fibronectin at a glance. *J. Cell. Sci.*, 115:3861, 2002.

[107] Williams C.M., Engler A.J., Slone R.D., Galante L.L., and Schwarzbauer J.E. Fibronectin Expression Modulates Mammary Epithelial Cell Proliferation during Acinar Differentiation. *Cancer Research*, 68:3185, 2008.

[108] E. Ruoslahti and M.D. Pierschbacher. New perspectives in cell adhesion: RGD and integrins. *Science*, 238:491, 1995.

[109] Krammer A., Lu H., Isralewitz B., Schulten K., and Vogel V. Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch. *Proc. Natl. Acad. Sci. USA*, 96:1351, 1999.

[110] A.F. Oberhauser, C. Badilla-Fernandez, M. Carrion-Vasquez, and J.M. Fernandez. The mechanical hierarchies of Fibronectin observed with Single-molecule AFM. *J. Mol. Biol.*, 319:433, 2002.

[111] E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.*, 288:441, 1999.

[112] Klimov D.K. and Thirumalai D. Native topology determines force–induced unfolding pathways in globular proteins. *Proc. Natl. Acad. Sci. USA*, 97:7254, 2000.

[113] Gao M., Craig D., Vogel V., and Schulten K. Identifying Unfolding Intermediates of FN–III_{10} by Steered Molecular Dynamics. *J. Mol. Biol.*, 323:939, 2002.

[114] H. Erickson. Reversible unfolding of fibronectin type iii and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin. *Proc. Natl. Acad. Sci. USA*, 91:10114, 1994.

[115] H. Li, W.A. Linke, A.F. Oberhauser, M. Carrion-Vazquez, J.G. Kerkvliet, H. Lu, P.E. Marszalek, and J.M. Fernandez. Reverse engineering of the giant muscle protein titin. *Nature*, 418:998, 2002.

[116] S.V. Litvinovich and K.C. Ingham. Interactions between Type III Domains in the 110 kDa Cell–binding Fragment of Fibronectin. *J. Mol. Biol.*, 248:611, 1995.

[117] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single–molecule pulling experiments. *Proc. Natl. Acad. Sci. USA*, 98:3658, 2001.

[118] P. Szymczak and M. Cieplak. Stretching of proteins in a force–clamp. *J. Phys.: Condens. Matter*, 18:L21, 2006.

[119] S. Luccioli, A. Imparato, S. Mitternacht, A. Irbäck, and A. Torcini. Unfolding times for proteins in a force clamp. *Phys. Rev. E*, 81:010902, 2010.

[120] R.Y. Tsien. The Green Fluorescent Protein. *Annu. Rev. Biochem.*, 67:509, 1998.

[121] M. Zimmer. Green Fluorescent Protein (GFP): applications, structure, and related photophysical behavior. *Chem. Rev.*, 102:759, 2002.

[122] M. Zimmer. GFP: from jellyfish to the nobel prize and beyond. *Chem. Soc. Rev.*, 38:2823, 2009.

[123] T.D. Craggs. Green fluorescent protein: structure, folding and chromophore maturation. *Chem. Soc. Rev.*, 38:2865, 2009.

[124] J. Dopf and T.M. Horiagon. Deletion mapping of the aequorea victoria green fluorescent protein. *Gene*, 173:39, 1996.

[125] X. Li, G. Zhang, N. Ngo, X. Zhao, S.R. Kain, and C.-C. Huang. Deletions of the Aequorea victoria Green Fluorescent Protein define the minimal domain required for fluorescence. *J. Biol. Chem.*, 272:28545, 1997.

[126] M. Mickler, R.I. Dima, H. Dietz, C. Hyeon, D. Thirumalai, and M. Rief. Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations. *Proc. Natl. Acad. Sci. USA*, 104:20268, 2007.

[127] R. Perez-Jimenez, S. Garcia-Manyes, S.R.K. Ainavarapu, and J.M. Fernandez. Mechanical Unfolding Pathways of the Enhanced Yellow Fluorescent Protein Revealed by Single Molecule Force Spectroscopy. *J. Biol. Chem.*, 281:40010, 2006.

[128] A. Nagy, A. Mlnsi-Csizmadia, B. Somogyi, and D. Lorinczy. Thermal stability of chemically denatured green fluorescent protein (GFP): A preliminary study. *Thermochimica Acta*, 410:161, 2004.

[129] M. Schlierf, Z.T. Yew, M. Rief, and E. Paci. Complex Unfolding Kinetics of Single–Domain Proteins in the Presence of Force. *Biophys. J.*, 99:1620, 2010.

[130] R.J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26:597, 2001.

[131] A.P. Minton. Influence of excluded volume upon macromolecular structure and associations in 'crowded' media. *Curr. Opin. Biotechnol.*, 8:65, 1997.

[132] A.P. Minton. Implication of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.*, 10:34, 2000.

[133] R.J. Ellis. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.*, 11:114, 2001.

[134] S.B. Zimmerman and S.O. Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli. *J. Mol. Biol.*, 222:599, 1991.

[135] M.J. Kerner, D.J. Naylor, Y. Ishihama T. Maier, H.-C. Chang, A.P. Stines, C. Georgopoulos, D. Frishman, M. Hayer-Hartl, M. Mann, and F. Ulrich Hartl. Proteome–wide Analysis of Chaperonin–Dependent Protein Folding in Escherichia coli. *Cell*, 122:208, 2005.

[136] A. Brinker, G. Pfeifer, M.J. Kerner, D.J. Naylor, F.U. Hartl, and M. Hayer-Hartl. Dual function of protein confinement in chaperonin-assisted protein folding. *Cell*, 107:223, 2001.

[137] S.E. Radford. GroEL: More than just a folding cage. *Cell*, 125:831, 2006.

[138] A. Gershenson and L.M. Gierasch. Protein folding in the cell: challenges and progress. *Curr. Opin. Struct. Biol.*, 21:32, 2001.

[139] D. Eggers and J.S. Valentine. Molecular confinement influences protein structure and enhances thermal protein stability. *Protein Sci.*, 10:250, 2001.

[140] B. Campanini, S. Bologna, F. Cannone, G. Chirico, A. Mozzarelli, and S. Bettati. Unfolding of green fluorescent protein mut2 in wet nanoporous silica gels. *Protein Sci.*, 14:1125, 2005.

[141] Hong J. and Gierasch L.M. Macromolecular crowding remodels the energy landscape of a protein by favoring a more compact unfolded state. *J. Am. Chem. Soc.*, 132:10445, 2010.

[142] HX. Zhou. Protein folding and binding in confined spaces and in crowded solutions. *J. Mol. Recognit.*, 17:368, 2004.

[143] HX. Zhou. Protein folding in confined and crowded environments. *Arch. Biochem. Biophys.*, 469:76, 2008.

[144] D. Thirumalai, D.K. Klimov, and G.H. Lorimer. Caging helps proteins fold. *Proc. Natl. Acad. Sci. USA*, 100:11195, 2003.

[145] D.K. Klimov, D. Newfield, and D. Thirumalai. Simulations of β–hairpin folding confined to spherical pores using distributed computing. *Proc. Natl. Acad. Sci. USA*, 99:8019, 2002.

[146] F. Takagi, N. Koga, and S. Takada. How protein thermodynamics and folding mechanisms are altered by chaperonin cage: Molecular simulations. *Proc. Natl. Acad. Sci. USA*, 100:11367, 2003.

[147] J. Mittal and R.B. Best. Thermodynamics and kinetics of protein folding under confinement. *Proc. Natl. Acad. Sci. USA*, 105:20233, 2008.

[148] M. Hayer-Hartl and A.P. Minton. A simple semiempirical model for the effect of molecular confinement upon the rate of protein folding. *Biochemistry*, 45:13356, 2006.

[149] P.G. de Gennes. *Scaling Conceps in Polymer Physics*. Cornell University Press, 1979.

[150] T. Sakaue and E. Raphaël. Polymer chains in confined spaces and flow-injection problems: Some remarks. *Macromolecules*, 39:2621, 2006.

[151] M.S. Cheung and D. Thirumalai. Effects of Crowding and Confinement on the Structures of the Transition State Ensemble in Proteins. *J. Phys. Chem. B*, 111:8250, 2007.

[152] D.L. Pincus and D. Thirumalai. Crowding Effects on the Mechanical Stability and Unfolding Pathways of Ubiquitin. *J. Phys. Chem. B*, 113:359, 2009.