SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

Molecular and Statistical BioPhysics Sector

PhD in Physics and Chemistry of Biological Systems

Characterizing Structure and Free Energy Landscape of Proteins by NMR-guided Metadynamics

Ph. D. Thesis of

Daniele Granata

Supervisors:

Alessandro Laio

Michele Vendruscolo

21st October 2013

To whom have the courage to
have doubts searching their way

To Alessio and Giorgia

Contents

Motivation

In the last two decades, a series of experimental and theoretical advances has made it possible to obtain a detailed understanding of the molecular mechanisms underlying the folding process of proteins [1, 2, 3, 4, 5, 6]. With the increasing power of computer technology [7, 8, 9, 10], as well as with the improvements in force fields [11, 12], atomistic simulations are also becoming increasingly important because they can generate highly detailed descriptions of the motions of proteins [13, 14, 15]. A supercomputer specifically designed to integrate the Newton's equations of motion of proteins, Anton [9], has been recently able to break the millisecond time barrier. This achievement has allowed the direct calculation of repeated folding events for several fast-folding proteins [16] and to characterize the molecular mechanisms underlying protein dynamics and function [17]. However these exceptional resources are available only to few research groups in the world and moreover the observation of few event of a specific process is usually not enough to provide a statistically significant picture of the phenomenon.

In parallel, it has also been realized that by bringing together experimental measurements and computational methods it is possible to expand the range of problems that can be addressed [4, 18, 19, 20, 21]. For example, by incorporating structural informations relative to transition states (Φ values) as structural restraints in molecular dynamics simulations it is possible to obtain structural models of these transiently populated states [22, 23], as well as of native [24] and non-native intermediates [25] explored during the folding process. By applying this strategy to structural parameters measured by nuclear magnetic resonance (NMR) spectroscopy, one can determine the atomic-level structures and characterize the dynamics of proteins [26, 27, 28, 29]. In these approaches the experimental information is exploited to create an additional term in the force field that

penalizes the deviations from the measured values, thus restraining the sampling of the conformational space to regions close to those observed experimentally [22].

In this thesis we propose an alternative strategy to exploit experimental information in molecular dynamics simulations. In this approach the measured parameters are not used as structural restraints in the simulations, but rather to build collective variables within metadynamics calculations. In metadynamics [30, 31], the conformational sampling is enhanced by constructing a time-dependent potential that discourages the explorations of regions already visited in terms of specific functions of the atomic coordinates called collective variables. In this work we show that NMR chemical shifts can be used as collective variables to guide the sampling of conformational space in molecular dynamics simulations.

Since the method that we discuss here enables the conformational sampling to be enhanced without modifying the force field through the introduction of structural restraints, it allows estimating reliably the statistical weights corresponding to the force field used in the molecular dynamics simulations. In the present implementation we used the bias exchange metadynamics method [32], an enhanced sampling technique that allows reconstructing the free energy as a simultaneous function of several variables.

By using this approach, we have been able to compute the free energy landscape of two different proteins by explicit solvent molecular dynamics simulations. In the application to a well-structured globular protein, the third immunoglobulin-binding domain of streptococcal protein G (GB3), our calculation predicts the native fold as the lowest free energy minimum, identifying also the presence of an on-pathway compact intermediate with non-native topological elements. In addition, we provide a detailed atomistic picture of the structure at the folding barrier, which shares with the native state only a fraction of the secondary structure elements.

The further application to the case of the 40-residue form of Amyloid beta, allows us another remarkable achievement: the quantitative description of the free energy landscape for an intrinsically disordered protein. This kind of proteins are indeed characterized by the absence of a well-defined three-dimensional structure under native conditions [33, 34, 35] and are therefore hard to investigate experimentally. We found that the free energy landscape of this peptide has ap-

proximately inverted features with respect to normal globular proteins. Indeed, the global minimum consists of highly disordered structures while higher free energy regions correspond to partially folded conformations. These structures are kinetically committed to the disordered state, but they are transiently explored even at room temperature. This makes our findings particularly relevant since this protein is involved in the Alzheimer's disease because it is prone to aggregate in oligomers determined by the interaction of the monomer in extended β-strand organization, toxic for the cells. Our structural and energetic characterization allows defining a library of possible metastable states which are involved in the aggregation process.

These results have been obtained using relatively limited computational resources. The total simulation time required to reconstruct the thermodynamics of GB3 for example (around 2.7 μs) is about three orders of magnitude less than the typical timescale of folding of similar proteins [36], simulated also by Anton in [16]. We thus anticipate that the technique introduced in this thesis will allow the determination of the free energy landscapes of wide range of proteins for which NMR chemical shifts are available. Finally, since chemical shifts are the only external information used to guide the folding of the proteins, our methods can be also successfully applied to the challenging purpose of NMR structure determination, as we have demonstrated in a blind prediction test on the last CASD-NMR target [37].

Chapter 1

Introduction

The great majority of the most relevant biological processes acts at the molecular level. On this stage proteins has probably the most significant role, participating in virtually every process within cells.

Many proteins are enzymes that catalyze biochemical reactions and are vital for the metabolism. Proteins have structural or mechanical functions acting as real cellular machinery, such as actin and myosin in muscle that are in charge of motion and locomotion of cells and organisms. Transmembrane proteins regulate the osmotic balance and the transportation of chemicals, nutrients and information, acting as receptors and signaling transducers across the cellular membrane. Others proteins are important for transporting materials, immune response, and several other functions.

Characterizing the tertiary structure of a protein and the energy of the interconversion between the different states can provide important clues about how the protein performs its function and information about the biological mechanisms in which it is involved. In particular, the knowledge of the interactions between proteins involved in deadly pathologies will help designing drugs and curing diseases, such as cancer, neurodegenerative diseases (Alzheimer, Huntington, Parkinson etc), virus infections. This explains why describing the structure and the functionality of a protein is one of the most important challenges in biophysics.

In this chapter we intend to provide the reader a background of the concepts

and terminology that will be presented and discussed throughout the thesis. After a brief introduction to the biochemistry of proteins, in Section 1.1 we will focus on the concept of folding and free energy landscape, analyzing the differences between well-structured and intrinsically disordered proteins. In Section 1.2 we will discuss the main techniques to characterize protein structure from experimental and computational point of view, with a particular attention on structure prediction methods. Finally, in Section 1.3 we will present some basic concepts related to nuclear magnetic resonance spectroscopy, with a special focus on chemical shifts and structure determination approaches.

1.1 Proteins: Life Building Blocks

A protein is a polymer chain of small subunits, called amino acids, which is defined by the nucleotide sequence of a gene. The genetic information, contained in the DNA, is first transcribed into messenger-RNA (mRNA) which is then translated into a linear chain of amino acids inside the ribosome. Finally the protein acquires its specific three dimensional structure which determines, in almost all cases, its functionality.

Each segment of three nucleotide basis of the mRNA specifies one out of 20 natural amino acids, each with specific features. However all amino acids possess common structural features (Fig. 1.1). They are composed of a central α carbon (C_α) to which an amino group (NH_3^+), a carboxyl group (COO^-), a hydrogen atom (H_α) and a variable side chain residue (R) are attached. In a protein, the amino acids are linked together by peptide bonds between the carboxyl and the amino groups of adjacent residues (lower part of Fig. 1.1). The serial assembly of consecutive peptide bonds then constitutes the entire polypeptide chain. Apart from the side chain ones, all these atoms constitutes the backbone of the protein and their relative arrangements is determinant for the local structure of the chain.

The side chains, instead, are what actually distinguishes the physical-chemical characteristics of the standard amino acids and, consequently, of the entire protein. They are usually classified based on their electrostatic properties, which can favor the formation of hydrogen bonds, salts bridges and in general their propensity to be exposed to the aqueous solvent, as for the charged (positive for Arginine,

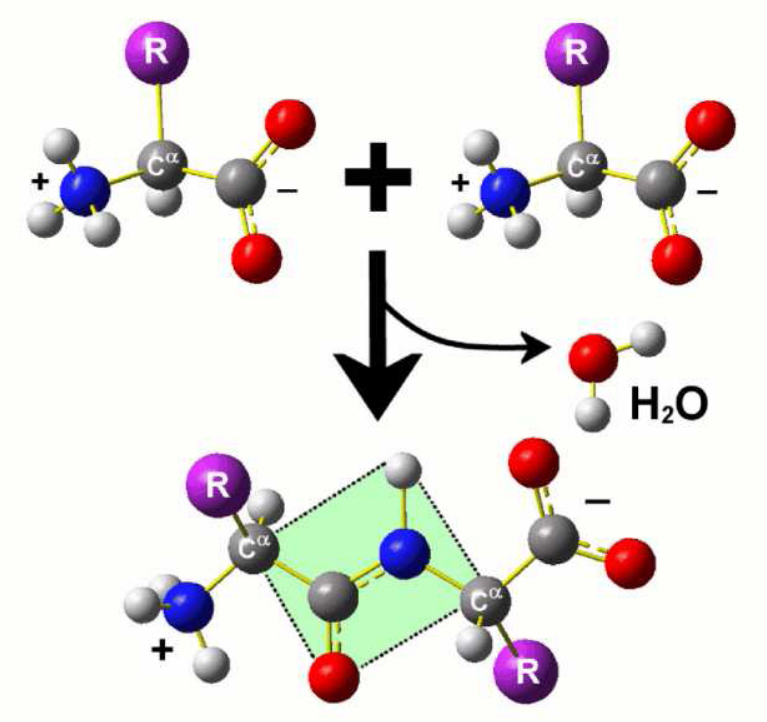

Figure 1.1: Representation of an amino acid (upper part) and of the peptide bond formation (lower part) between two consecutive amino acids

Histidine, Lysine and negative for Aspartic and Glutamic Acids) and polar ones (Serine, Threonine, Asparagine, Glutamine), or their propensity to be buried in the core of the protein, as for hydrophobic residues (Alanine, Leucine, Isoleucine, Methionine, Phenylalanine, Tryptophan, Tyrosine and Valine). Glycine, Proline and Cysteine has instead special structural properties.

1.1.1 Sequence-Structure-Function Paradigm and Protein Folding

Since the beginning of protein science [38, 39, 40, 41, 42] it has been clear that the amino acid sequence encodes all the information about the protein, determining univocally its structural characteristics and functions. As soon as the growing unfolded polypeptide chain exits the ribosome where it is synthesized, it starts to acquire a well defined three dimensional structure based on the sequence of the

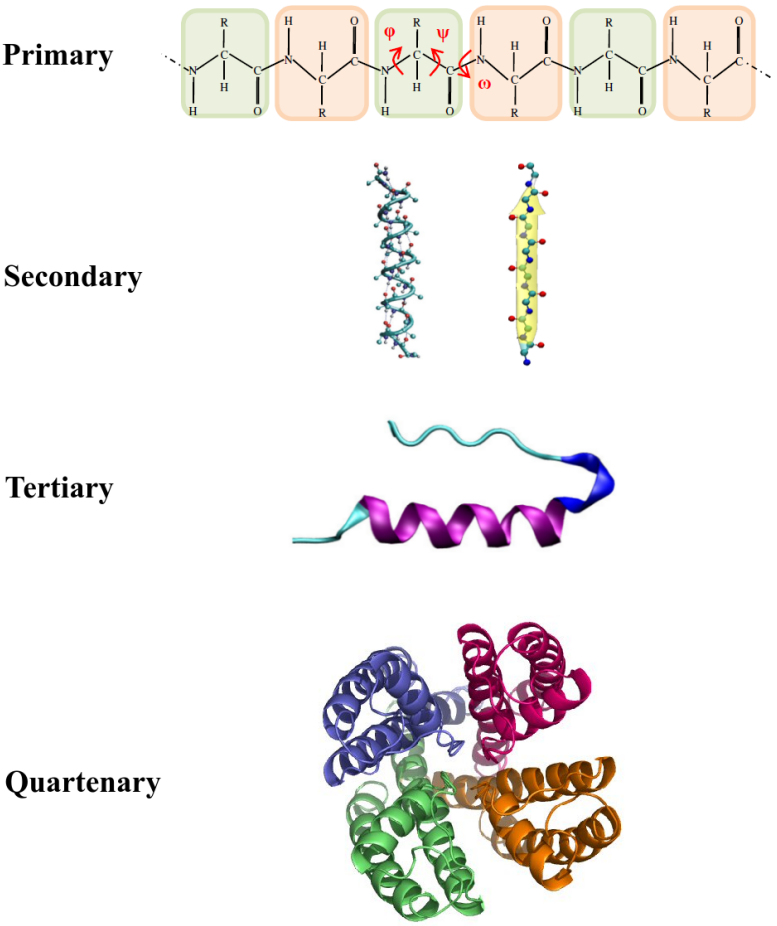

Figure 1.2: Schematic representation of the structural hierarchy in proteins. The primary structure can determine the formation of secondary structure elements which are then assembly to form the characteristic tertiary arrangement of a globular protein. The further interaction of multiple single domains can the generate the functional quaternary assembly of a complex

residues[42], through a folding mechanism. The acquisition of a stable structure, also called native state, is usually a fundamental requirement for the protein biological function [39].

This underlying hierarchy from sequence to function, is well represented by the four distinct aspects of a protein structure, reported in Fig. 1.2. The first level of protein structure is simply the amino acids sequence of a protein chain, which is called primary structure. This sequence of amino acids is specific for each protein. Amino acids interact with each other locally leading to secondary structure

elements such as α-helix or β-strands. The α-helix is formed by hydrogen bonds between the backbone atoms of consecutive residues in a protein. Also β-strands are determined by hydrogen bonding of the backbone atoms, but even for amino acids distant in sequence: they are usual present in two arrangements, parallel or anti-parallel. The assembly and the interaction of these structural motifs (α-helix and β-sheets), together with the formation of a hydrophobic core, produce the overall globular three dimensional shape of a protein, known as the tertiary structure. The further arrangement of distinct folded domains in a multi-subunit complex is called quaternary structure.

Protein Folding

This representation gives also a general view of the process by which a polypeptide chain folds from a random coil into a functional tertiary structure, called protein folding. At this first stage the polypeptide lacks any developed tridimensional structure. Then, based on their physicochemical properties, the amino acids start to interact with each other to produce the final stable conformation.

The early studies on the denaturation of proteins [43, 39, 40] and the relative observation that a denaturated protein could acquired back its native state by changing the solution conditions, led to the Anfinsen's statement of a thermodynamics hypothesis for protein folding. The reversibility of the process indeed implies that the native (functional) and the denaturated protein (non functional) can be treated as two separate thermodynamic states. Moreover this leads to the interpretation that the protein's function is determined by a well defined structure of the native state, which is the lowest in the Gibbs free energy for the whole system [42].

The folding mechanism occurs therefore under the combined action of enthalpic and entropic contributions from the protein and the solvent. The native conformation is determined by favorable interatomic interactions which lead to the formation of hydrogen bonds, as in secondary structure elements, and the neutralization salt bridges within the protein's core. At the same time its stabilization is enforced by the sequestration of hydrophobic amino acid side chains in the interior of the folded protein: this allows the exposure of charged and polar

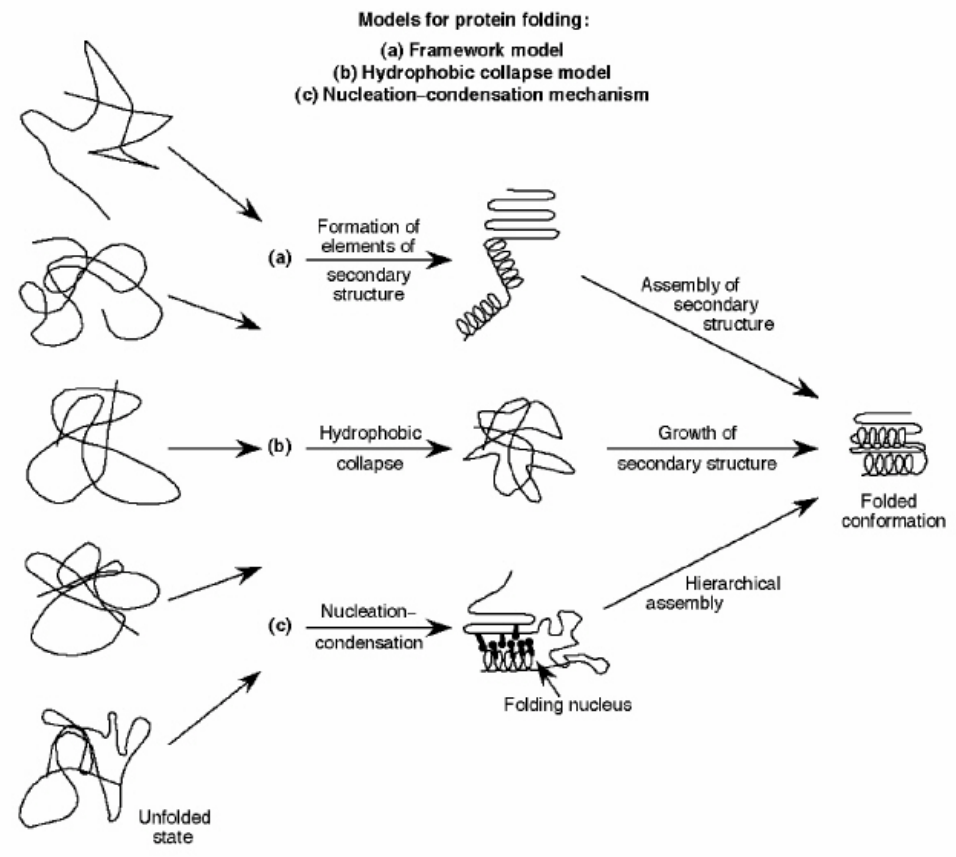

Figure 1.3: Three different structural models for the folding of a protein. Starting from an ensemble of unfolded conformations, the final native folds can occur: (a) though assembly of already formed secondary structure elements; (b) through the hydrophobic collapse in a central protein core and subsequent growth of the secondary structure; (c) formation of an initial folding nucleus, consisting of specific contacts inside the protein, followed by the others folded structural elements. Adapted from [44].

side chains on the solvent-accessible protein surface, and the water solvent to maximize its entropy, lowering the total free energy.

Although the physics and the chemistry of these interactions are well understood, the folding pathway can be very complicated, because the extent of the different contributions can determine way to reach the folding states. Three are the most common models for a folding process, as schematized in Fig. 1.3 [45, 46, 47]: a) first the formation of the secondary structure occurs, followed

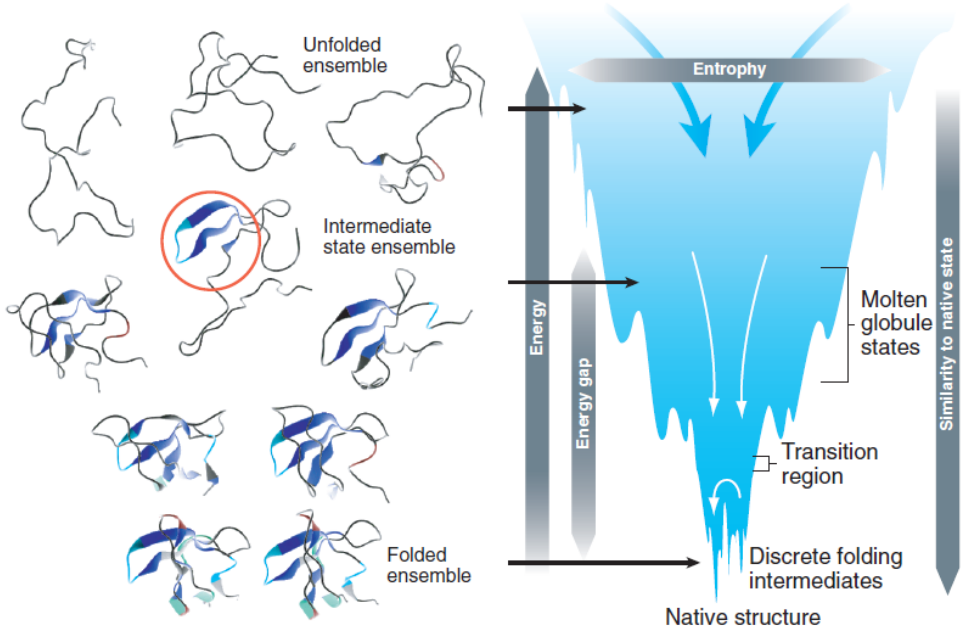

Figure 1.4: Schematic representation of the folding free energy of a protein. Folding occurs through the progressive organization of ensembles of structures on a funnel-shaped free energy landscape (right). Conformational entropy loss during folding is compensated by the free energy gained as more native interactions are formed. Kinetics is determined by the local roughness of the landscape, relative to thermal energy. Adapted from [49]

by their assembly in the final structure; b) the hydrophobic collapse of the hydropathic residues takes place before, then the secondary structure forms; c) the fold is determined by the initial condensation of a folding nucleus formed by few important contacts, which constitutes the seed for the native structure formation. The amino acid sequence is the determinant of the choice of the followed pathway, which can obviously result by the combination of these models, especially in large protein where different region can adopt different folding strategies [48].

The Free Energy Landscape

The heterogeneity of the folding mechanism reflects the high complexity of the free energy landscape of proteins. This theoretical concept was introduced [52, 1, 53] to give a qualitatively description of protein folding, based on the "principle of minimal frustration" [54]. According to this principle, nature selected amino acid

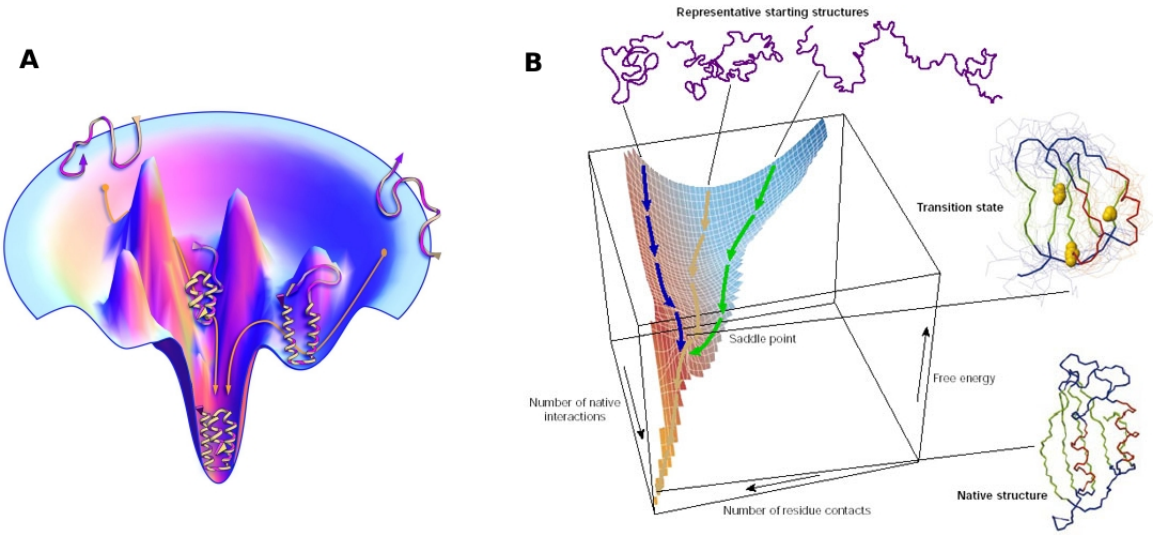

Figure 1.5: Qualitative description of two typical free energy landscape for globular protein. (a) The folding can be quite complicated, exhibiting high kinetic barriers and folding intermediates, following different folding pathway. (b) The folding can occurs in a simple two-state system, in which the unfolded ensemble and the native state are separated by only one high free energy barrier (saddle point or transition state), following in practice a unique pathway. Adapted respectively from [50] and [51].

sequences so that the folded state of the protein is very stable and the formation of favorable interactions along the folding pathway makes the acquisition of the relative structure a fast process. As shown in Fig. 1.4, the protein conformation is initially unfolded and dominated by entropy, but at high Gibbs free energy; the progressive collapse with a consequent loss in entropy is overcome by the enthalpic contribution from electrostatics with the formation of more compact structures (molten globule states), still very different and separated by high energetic barriers from the native state. The following formation of the secondary structures, native contacts and of the hydrophobic core gives the final gain in internal enthalpy and solvent entropy which determines a deep energy gap between the native state and the other conformations.

This perspective depicts the landscape as a global energy funnel [55] determined by these favorable contributions and directed towards the free energy minimum, which is highly stable and reachable through a large number of pathways

instead of a single mechanism, which makes the folding a rapid process, usually on the timescale of milliseconds or even microseconds. This overall picture, indeed, allows overcoming the paradox posed by Levinthal [56], who observed that if a protein were folded by random sampling of all possible conformations ($3^{2(N-1)}$ only considering three possible orientations for the ψ and ϕ backbone angles), the folding process would take a time longer than the age of the universe.

However, even though nature has reduced the level of frustration through sequence selection, depending on the protein, the landscape can be highly complex. It can be a simple two state system, with the unfolded state separated from the native one by a single free energy barrier (see Fig. 1.5b), which is populated by structures with the fundamental folding nucleus already formed (transition state ensemble). Or it can have several local energy minima (metastable states) with intermediates which can act as kinetic traps for the folding process, as described by Fig. 1.5a. In addiction, external factors can further complicate this picture: temperature, pH, molecular crowding and the interaction with other molecules can alter the free energy landscape, determining the changes in conformation which are proper of the protein function, modifying the equilibrium between the native state and an intermediate, or the denaturation of the protein, with the associated loss of its functionality, or even the degeneration in misfolded state which can then interact and aggregate with other proteins forming amorphous compounds, that can be toxic and lethal for the cells.

1.1.2 Some Exceptions: Intrinsically Disordered Proteins

For decades the sequence-structure-function paradigm, together with its theoretical justification, have constituted a solid dogma. The successes of X-ray crystallography which provided the final physical representation of a protein [58, 59, 60], literally crystalized the idea that every protein possesses a native state with a well defined three dimensional structure, required for its biological function. But soon some ambiguous observations started to appear: loops, known to be important for function, were missing in high-resolution structures [61, 62]. Nuclear magnetic resonance spectra also revealed that some proteins with known biological function did not have a stable and defined structure in solution [63]. These findings

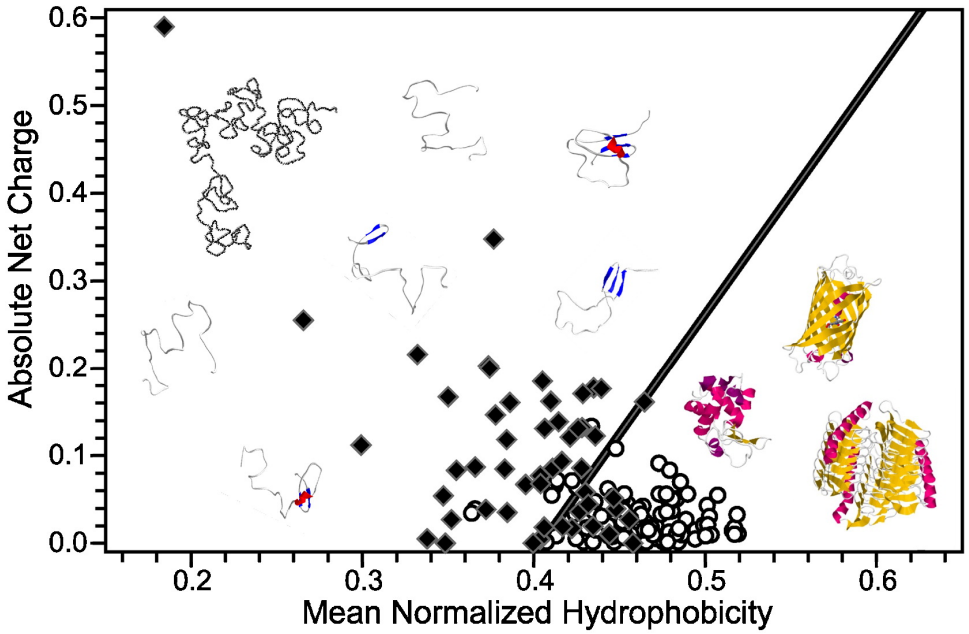

Figure 1.6: Difference in sequence composition between well-structured and intrinsically disordered proteins. The two-dimensional scatter plot shows a clear separation between this two kind of proteins as a function of the mean hydrophobicity and of the net charge. The folded protein clusters in the low-right corner (empty dots), characterized by a high hydrophobicity and a low net charge, the disordered proteins (black dots) are instead characterized by low hydrophobicity and a higher net charge. Adapted from [57]

were initially attributed to the dynamical properties of proteins (in particular to loops, linkers or terminal regions), or even associated with misfolding-related diseases, since several unstructured proteins, such as α-synuclein, β-amyloid, amylin or p53, are involved in several human "disorders" (respectively Parkinson's and Alzheimer's disease, diabetes, cancer) [64].

Instead, it is now well established that many proteins lack a specific tertiary structure under functional conditions [33, 34, 35]. They exist as an ensemble of flexible and mobile conformations. This flexibility can be localized in particular regions of the protein (these are referred to as intrinsically disordered regions or partially folded proteins), or can regard the entire protein length (natively unstructured, natively unfolded, intrinsically disordered are the most common nomenclature).

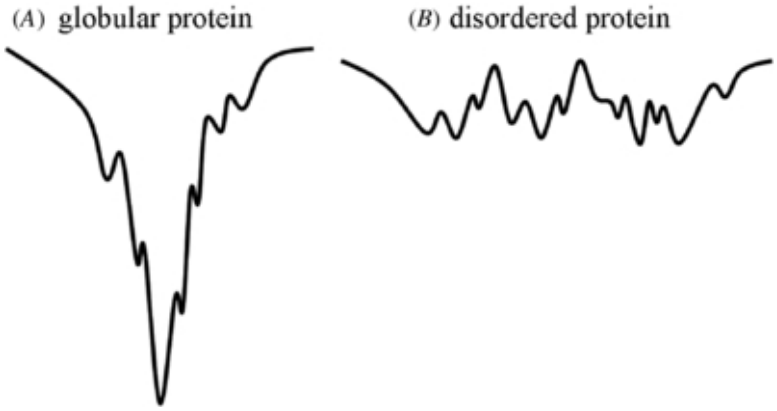

Figure 1.7: Schematic representation of the energy landscape of a globular protein (a) and a disordered protein (b). The energy of the system is sketched against a single coordinate of the conformational space. Adapted from [67])

From a structural point of view these disordered proteins or regions, are not simple random coils. Experimental evidences showed that they often populates extended conformations but with local transient secondary structure elements [65, 66], with short helical, β-bridges and in particular polyproline II helix.

Comparing the amino acid sequence, well-structured and intrinsically disordered proteins show important differences [68, 57]. Unstructured proteins are characterized by a low content of bulky hydrophobic amino acids and a high proportion of polar and charged amino acids. Thus disordered sequences cannot bury a critical hydrophobic core to determine the crucial solvent entropic gain like for stable globular proteins. The graph in Fig. 1.6 show how these two categories of proteins clusters in two well separated groups, simply based on the mean hydrophobic and net charged content of their primary structure, observation which has been the basis of many disorder predictors [68, 69, 70, 71].

A possible schematic representation of the free energy landscape of an intrinsically disordered protein is reported in Fig. 1.7 [67]. While globular proteins are usually characterized by a deep energy minima which sequesters the system in a specific structures, the landscape of an intrinsically disordered protein can be interpreted as very rugged, with several local minima separated by low energy gaps, comparable to the thermal fluctuations of the external bath. This allows disordered proteins to adopt an ensemble of rapidly interconverting and

easy-to-access conformations, instead of a single stable one. In this thesis we will provide a complete and quantitative view of this naif picture, able to capture more profound properties of intrinsic disordered proteins (see Chapter 6).

As discussed above for structured proteins, the interaction with ligands and external factors usually define the functionality of a protein, modifying its free energy landscape to favor conformational changes. This is even more striking for intrinsically disordered proteins: the interaction with their natural targets allows the selection of a particular structured conformation [72, 73], lowering its relative free energy. The coupled folding and binding allows the complementary burial of a large surface area involving a small number of amino acids: to achieve the same surface burial with only folded proteins would require a much larger protein [74], with a higher metabolic cost for the cell. In many cases they also shows the ability to bind in different conformations to different partners [75], which can be explained by their barrieless energy landscape. This is often done using also different regions of the sequence [76], which means that they can link two targets, acting as a sort of hub to favor their localization and interaction.

1.1.3 Revising the Protein Paradigm

The ability of disordered proteins to bind, and thus to exert their function, demonstrates that stability and acquisition of a well-defined three dimensional structure is not a required condition for all proteins. However the Anfinsen's thermodynamics hypothesis, with the protein in equilibrium between functional and non functional states, is still valid, but in a statistical physics sense. While for globular proteins the native, functional states is both a structural and thermodynamic state, for disordered protein is only a thermodynamic state, composed by an ensemble of heterogeneous structural conformations. A particular three dimensional structure should be thought, in a more loose way, as a single realization of the system in equilibrium with all the other possible configurations. The profound diversity between structured and unstructured proteins in sequence, and consequently in energetics and structural characterization, indeed reflects the functional purpose and role of these two kinds of proteins.

The standard view of the paradigm applies perfectly to enzymes. In catalytic

reactions, enzymes usually works through the stabilization of their transition state [77]: in a catalysis the protein binds more tightly to the transition state than to the native one, in order to lower the activation barrier and to accelerate the reaction rate. Transition states are usually derived from the native conformation by very slight movements of atoms. This implies an accurate prior positioning of the key amino acids and consequently a well-ordered protein structure is prerequisite for activity.

The peculiar properties of disordered proteins are instead perfectly suitable for signaling, regulation or control. Due to the absence of high barriers in their free energy landscape, when disordered regions bind to signaling partners, the energetic cost required to determine a transition from a disordered to an ordered conformation is quite low and easily compensated by specific contact interactions on a large surface area. Also the ability of these protein to readily bind to multiple partners by changing shape to associate with different targets[75, 76] is a characteristic feature of these proteins and of their energetic landscape, because it allows them to regulate the association of two other targets. The combination of high complementarity with low affinity and sometimes also with promiscuous binding properties are fundamental for signaling and regulation and would be very difficult to evolve between two ordered structures.

1.2 Protein Structural Characterization

The combination of experimental techniques and computational approaches employing the theoretical concepts described so far, has been able to clarify different aspects of protein folding.

This task has been successfully achieved for proteins which possess a well-defined structure like enzymes, but is still a great challenge in the case of highly-dynamic proteins, such as partially-folded and intrinsically disordered ones. Moreover the high-throughput of new proteins due to the sequencing of entire genomes[78] makes the application of standard experimental techniques unaffordable, because too slow and expensive. This motivated the development of theoretical and computational methods and bioinformatics tools capable of modeling or extracting structural information only from the sequence by comparison of known structures,

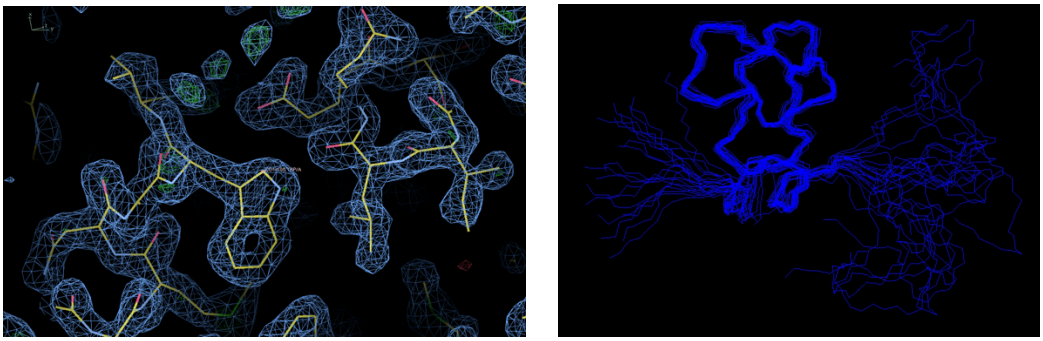

Figure 1.8: Examples of protein structures solved by X-ray crystallography (on the left) and NMR spectroscopy (on the right) . Adapted respectively from [80] and [81].

or even to fold protein *ab initio*, based only on physical principals.

1.2.1 Experimental Techniques

If we look at the Protein Data Bank (PDB)[79], the most important repository for the 3-D structural data of proteins and nucleic acids, almost all solved macromolecular structure are obtained by X-ray crystallography (90%) and NMR spectroscopy (10%). At the moment these two methods are only capable of resolving structure at atomic resolution. Nevertheless they present severe limitations.

X-ray Crystallography

In this method a macromolecule structure is determined from the diffraction pattern of a X-ray beam on crystals which contain the protein[82]. By measuring the angles and intensities of these diffracted beams, an electron density map of the crystal can be reconstructed together with the protein structure, fitting the atoms of the polypeptide chain into the map. This technique provides only a static representation of the protein, with little information about its dynamics apart from the presence of flexible regions which are missing in the diffraction pattern. It is also important to stress that experimental conditions required for X-ray crystallography can be critical: indeed several proteins are difficult or impossible to crystalize, such as membrane and disordered proteins. This also means that since X-ray structures are the most abundant in the PDB database, some protein

families are overrepresented with respect to others. This has produced a bias in our understanding of protein science, as we have already mentioned in Section 1.1.2. Even for easy target as globular proteins the unphysical tight packing of the crystal can alter significantly the final structure.

NMR Spectroscopy

NMR experiments [83, 84, 85] provides structural information in the form of distance and angular restrains which allow to infer and predict the local arrangements of the protein atoms. All the restrains are then combined to produce the final possible conformations, usually by solving a distance geometry problem. This method allows to study proteins in solution and therefore closer to physiological conditions. Moreover it can provide also important information about the dynamics of the system, allowing also the characterization of folding intermediates or transiently populated states, as for intrinsically disordered proteins. However also this technique is affected by limitations: in particular as the size of the system increases, the extraction of spatial restrains becomes harder, and accurate structural prediction is limited to proteins of few hundreds of residues. Since some aspects of this technique are central for this thesis, they are discussed in more details in Section 1.3

Other techniques

Also electron microscopy (including cryo-electron microscopy and electron crystallography) can produce lower-resolution structural information about very large protein complexes, including assembled viruses [86] and two-dimensional crystals of membrane proteins[87]. Finally, other experimental methods such as UV circular dichroism , small-angle X-ray scattering (SAXS), mass spectrometry, hydrogen exchange, fluorescence resonance energy transfer (FRET) and atomic force microscopy (AFM) can provide important quantitative structural information for secondary structure population, compactness, residues exposure and presence of metastable states along the protein folding.

1.2.2 Theoretical and Computational Approaches

Following the idea that the amino acid sequence determines the structural properties of the protein, many theoreticians started to deal with the challenging problem of predicting the tertiary structure using only the notion of the primary one [88]. The development of bioinformatics, in particular of sequence alignment methods [89, 90, 91], and of effective potential for native folding discrimination [92, 93, 94, 95, 96] gave important contributions in this direction. The performance of current methods is assessed every two years in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction), where research groups try to predict the structure from sequences of soon-to-be solved proteins [97].

Two are the usual strategies to address this problem: comparative (or template-based) modeling and *de novo* (or *ab initio*) structure prediction.

Comparative modelling

This approach is based on two converging observations. Although the number of actual proteins is huge [79], they are characterized by a limited set of tertiary folds [98, 99, 100]. The second observation follows directly from the protein paradigm, that is similar sequence should determine similar structures. These considerations originate two type of methods. Homology modeling tries to built the atomistic description of a target sequence out of the experimental three-dimensional structure of one or more related homologous proteins [101, 102]. These are identified through sequence alignments and since the protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy also on a very distantly related template. The main bottleneck is actually represented by the accuracy of sequence alignment [103].
Threading algorithms [104, 105, 106] instead scan the amino acid sequence of the target against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. Protein threading treats the template in an alignment as a structure, and both sequence and structure information extracted from the alignment are used for prediction. Protein threading can be still effective

in case of no significant homology.

De novo structure prediction

These methods attempts building three-dimensional protein models "from scratch", based on general principles that govern protein folding energetics and/or statistical tendencies of conformational features that native structures acquire, without the use of explicit templates from previously solved structures. A general paradigm for de novo prediction usually involves an efficient sampling algorithm in the conformational space, guided by the minimization of physics-based or knowledge-based scoring functions [94, 95] such that a large set of candidate ("decoy") structures are generated. Native-like conformations are then selected from these decoys using scoring functions as well as conformer clustering. The process can involve coarse graining and successive refinement stages [107]. These procedures normally require vast computational resources, and have thus only been carried out for tiny proteins and few research groups. To predict protein structure de novo for larger proteins will require better algorithms and larger computational resources, using powerful supercomputers (such as Blue Gene [7], MDGRAPE [8] or Anton [9]) or distributed computing strategies (such as Folding@Home [10, 108] and Rosetta@Home [109]).
Also molecular dynamics simulations can be considered and applied as a particular case of *ab initio* predictions, as spectacularly shown by the recent result obtained by the D.E. Shaw's group [16]. This specific topic will be discussed in the introduction to the next chapter.

1.3 NMR of Proteins

Nuclear magnetic resonance spectroscopy is an experimental technique that exploits the magnetic properties of certain atomic nuclei. It determines the physical and chemical properties of atoms or the molecules in which they are contained, and can provide detailed information about the structure, dynamics, reaction state, and chemical environment of molecules[83, 110, 85]. For these reasons NMR spectroscopy has a fundamental role in structural biology, being able to

give atomic resolution description of biomolecule in solution, as we mentioned previously in Section 1.2.1.

1.3.1 Chemical Shifts

Chemical shifts are the most readily and accurately measured NMR observables. When placed in a magnetic field, atoms which possess a nuclear spin (such as ^{1}H,^{1}3C, ^{1}5N) absorb electromagnetic radiation at a frequency characteristic of the isotope and proportional to the strength of the external field. These resonance frequencies are then registered on one or multi-dimensional spectra specific for the nucleus type. The chemical shift quantifies the difference of the frequency for a specific atom with respect to the known one of a reference compound: in this way it provides information about the chemical environment of the corresponding atom, because it reveals the local modification of the magnetic field induced by the presence of neighboring atoms, due to shielding effects of their electrons. The assignment of the peaks of the spectra to the corresponding atoms of the primary structure allows then the reconstruction of the local arrangement of a molecule.

These properties are very useful in protein science. The high sensitivity to the surrounding environment allows the chemical shift to indicate if the atom is involved in a hydrogen bond, or the presence of a close aromatic ring. The use of multi-dimensional NMR experiments [110], such as J-coupling, nuclear Overhauser effect (NOE) and correlation spectroscopy, permits in particular to analyze the local dihedral conformation of the chemically bound atoms and to identify the residues involved in secondary structure elements. In addiction, exploiting the NOE spectroscopy (NOESY), it is possible to detect also specific contacts between two atoms, providing upper limits on their distance and consequently important information about the tertiary structure of the protein.

Since the acquisition time of the NMR signals is usually on the millisecond time scale, the signal registered by the chemical shifts and other NMR observables is an ensemble average of all the conformations explored by the protein during that time. This is particularly important because it allows NMR spectroscopy to provide quantitative information of the dynamics and on the folding of the protein, revealing the presence of intermediates, domain fluctuations and

transient population of different secondary structure elements. This is why this experimental technique has been finding also a huge application in the study of intrinsically disordered proteins.

1.3.2 Structure Determination Protocol

The typical protocol for protein structure determination by nuclear magnetic resonance (NMR) spectroscopy involves a number of sequential steps [85, 111]:

- Resonance assignment. The chemical shifts observed in multidimensional NMR spectra are assigned specifically to their corresponding atoms in the protein sequence.
- Spatial restraints extraction. Thousands of nuclear Overhauser effects (NOEs) are identified in multidimensional NOE spectroscopy spectra, assigned two specific pair of interacting atom and converted into interatomic distance restraints. Additional conformational restraints come, for example, from measurements of residual dipolar couplings and scalar couplings, which provide information about the orientation of the backbone dihedral angles and about the secondary structure of the macromolecule.
- Structure generation. Software programs are then used to solve a geometry constraint satisfaction problem and to generate a set of protein conformations (called bundle of conformers, or models) which satisfy the experimental and spacial restraints. This step is usually performed and reiterated several times together with the previous one, in order to maximize the agreement with the conformational restraints and to minimize the number of violations.
- Structure refinement and validation. The bundle of conformers is usually energetically refined through restrained molecular dynamics simulations or through the minimization of pseudoenergy function. Finally the quality of the models is assessed through standard validation tools [112, 113, 114].

Many attempts has been made to partially and fully automatize the different stages of this protocol combining different strategies [115, 116, 117, 118] and com-

putational resources [119, 120], trying also to simplify the protocol and skipping the NOESY assignment step [121, 122]. Recently, a community-wide experiment, CASD-NMR [37] (see Section 1.3.3), conceptually similar to CASP competition [97], has been introduced to benchmark the different methodologies [111].

1.3.3 CASD-NMR

CASD-NMR [37] is a community-wide experiment involving developers of software tools and protocols for the automated calculation of protein structures from NMR data, with the scope of evaluating and favoring the improvement of the different methodological strategies in the field [111]. CASD-NMR collects and makes available to the participants NMR data sets (chemical shifts assignment, residual dipolar coupling, and NOE peak lists) for which the corresponding protein structure, obtained with the traditional manually curated procedures, is not publicly available at the time of release. These data can be used by the participants for protein structure determination using fully automated methods as if they would directly deposit them into the PDB. The results are then analyzed through various validation tools and compared with the original released PDB structure.

Using the method presented in this thesis and explained in the following chapters, we have participated to the last submitted data set (HR2876C) to benchmark our methodology on a blind test. The results are presented in Section 4.3.2.

Chapter 2

Theoretical and Methodological Background

During the last decades, molecular dynamics (MD) simulations[123, 13] have been optimized properly in order to obtain an accurate and reasonable picture of biological systems in agreement with experimental results. But, as was mentioned previously, the major problem with MD is that it is a daunting task to study phenomena, like protein folding, protein aggregation, ion channeling, etc, that happen in time scales on average much larger than the commonly accessible time scales in simulations, in the range of microseconds. Coarse-grained approaches has been developed, schematizing groups of atoms in effective beads and reducing the number of degree of freedom to be integrated. The pay off of these approaches, which allow to access longer time scales, is the loss of a detailed atomistic description exploiting generally empirical more than physics-based interactions (often also difficult to be parametrized) and which can bring to less reliable results. Especially in biomolecular simulation the main computational cost is represented by water molecules, usually in a ratio 1:20 with solute atoms, but the attempts to treat water implicitly[124] in the force fields weren't able to reproduce properly the properties of the system, as we also verified in our study (see section 4.3.1).

Recently high-performance computing strategies like purpose built machines [9, 7]or distributed computing [10] were able to reach time scales of millisecond

in full-atom simulation in explicit solvent [16, 14]. However these exceptional resources are available only to few research groups in the world and even in this case the observation of few event of a specific process is usually not enough to provide a statistically significant sampling of the phenomenon.

Powerful alternatives to these brute force approaches are provided by several methodologies aimed at accelerating rare events and conformational transitions using commonly available resources. Most of these approches can provide also the statistical weights of the sampled configuration, allowing the reconstruction of the thermodynamics of the system consistent with the one obtained by a plain MD simulation.

In this chapter we will discuss the general basis behind our methodology, in particular molecular dynamics (Section 2.1) and metadynamics (Section 2.3). Finally in Section 2.4 we will discuss the Bias-Exchange metadynamics, an extension for problems with a huge number of degrees of freedoms, underlining especially how it is possible to reconstruct the free energy of a system employing this technique, which the fundamental enhanced sampling approach that we employed in this thesis.

2.1 Molecular Dynamics

Molecular dynamics is a computational technique aimed at simulating the time-evolution of atoms and molecules, under the action of the forces generated by a potential that provides approximations of the physics and chemistry of the system under investigation[123]. Because molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically. MD simulation circumvents this problem by using numerical methods. The forces over a particle in MD are approximated with a sum of analytical terms describing the chemical bonds, the van der Waals, the electrostatics interactions, etc., globally called the force field. The force field parameters are fitted on the potential energy surface evaluated with quantum approaches in small molecules representing parts of the system or derived from experimental data. During the last two decades, a lot of effort has been dedicated for the force field parameter optimization[125, 126, 11, 127]. AMBER [126] and CHARMM

[125] are most commonly used force fields. Even though these force field have been quite successful in reproducing a lot of experimental systems, we note that they still have important limitations [11, 128].

In a normal MD simulation, given a certain force field $V(\mathbf{R}^N)$ and the positions and momenta of a particle i at time t as $\mathbf{R}_i(t)$, $\mathbf{P}_i(t)$, respectively, the accelerations over the particles are computed using $\ddot{\mathbf{R}}_i(t) = -\nabla_i V/M_I$, and then the equations of motion are numerically integrated on time step (Δt) to find the final positions $\mathbf{R}_i(t + \Delta t)$ and final momenta $\mathbf{P}_i(t + \Delta t)$. The most commonly used integrator is the Velocity Verlet [129]:

$$\mathbf{R}_i(t + \Delta t) = \mathbf{R}_i(t) + \frac{\mathbf{P}_i(t)}{M_i}\Delta t + \frac{1}{2}\ddot{\mathbf{R}}_i(t)\Delta t^2 \tag{2.1}$$

$$\mathbf{P}_i(t + \Delta t) = \mathbf{P}_i(t) + M_i\frac{\ddot{\mathbf{R}}_i(t) + \ddot{\mathbf{R}}_i(t)}{2}\Delta t \tag{2.2}$$

This procedure is repeated iteratively and the system is evolved in time. The choice of the time step Δt is quite crucial and depends on the kind of simulated system. If it is too large the integration won't be accurate resulting in a unrealistic evolution and numerical instabilities in the simulation. If it is too small the computational time and resources need to observe meaningful events along the simulation will not be affordable (see also next Section). For atomistic simulations the appropriate Δt is usually of the order of 1-2 fs. Most of the MD simulations are done within the canonical ensemble, where the number of particles (N), volume (V) and temperature (T) are conserved. In the NVT ensemble, the energy of endothermic and exothermic processes is exchanged with a thermostat. A variety of thermostat methods are available to add and remove energy from a system in a more or less realistic way. Popular techniques to control temperature include velocity rescaling[130], the Nosé-Hoover thermostat [131, 132], and the Berendsen thermostat [133]. In general the result of a MD simulation will depend crucially on the force field, system size, thermostat choice and parameters, time step and integrator.

2.2 Rare Events and Computing Free Energy

Several methods have been developed to accelerate the rare events in MD or Monte-Carlo simulation [31] and important successes have been achieved in various fields, ranging from solid state physics to quantum chemistry [134, 135, 136, 137]. Most of these techniques are unfortunately only partially useful for biophysical applications, because of the enormous amount of degrees of freedom involved. In the simulation of a normal size protein with explicit water, one has to treat explicitly approximately 10^4 atoms, which correspond to a 10^4 dimensional configuration space, while the more interesting processes depends usually on concerted and very slow collective motions.

A methodology that seems to offer a general route for studying such complex problems is rthe eplica exchange molecular dynamics (REMD, also known as parallel tempering, or multicanonical ensemble method [134, 138]). Several replicas of the same system are run at different temperatures (usually between 270 and 500 K) to enhance the conformational sampling, then from time to time exchanges between the replicas are attempted according a Monte-Carlo scheme: if the move is accepted the two configuration are swapped and the two simulations continue the new configurations. At the end the thermodynamics of the system is reconstructed from the conformations sampled at the desired temperature. Depending on the system, also this method can be extremely computationally demanding because it requires a large number of replicas to guarantee an adequate acceptance rate, and the simulations have to be run for very long time to accumulate enough statistics. Moreover the force field parameters are usually optimized at room temperature and so artifacts in the sampling can arise from simulating it at high temperature.

Other approaches try to address the problem reducing the dimensionality of the system into a few reaction coordinates or collective variables (CVs), that are assumed to provide a coarse-grained and (possibly) comprehensive description of the system, and then to explore the free energy surface as a function of these variables.

2.2.1 Dimensional Reduction

Considering a system of particles of coordinates x coupled to a thermostat bath of temperature T, it evolves under the influence of a potential $V(x)$, following the canonical equilibrium distribution:

$$P(x) = \frac{1}{Z} e^{-\beta V(x)} \tag{2.3}$$

where $\beta = 1/k_{\mathrm{B}}T$ and $Z = \int dx e^{-\beta V(x)}$ is the partition function of the system. For system characterized by a huge number of degrees of freedom, as for biological ones, like proteins, the $P(x)$ has an incredibly large dimensionality. In order to describe the system in more simple terms, what is done is to consider the reduced probability distributions in terms of some reaction coordinates or collective variables $s(x)$. Namely, instead of monitoring the full trajectory $x(t)$ of the system, a reduced trajectory $s(t) = s(x(t))$ is analyzed. The probability distribution $P(s)$ can be written as:

$$P(s) = \frac{1}{Z} \int dx e^{-\beta V(x)} \delta(s - s(x)) \tag{2.4}$$

or for an infinitely long trajectory it can be evaluated by the histogram of s:

$$P(s) = \lim_{t \to \infty} \frac{1}{t} \int dt e^{-\beta V(x)} \delta(s - s(t)) \tag{2.5}$$

In real application $P(s)$ is estimated as

$$P(s) \simeq \frac{1}{n \Delta s} \sum_{t=1}^{n} \chi_s(s(t)) \tag{2.6}$$

where the characteristic function $\chi_s(x) = 1$ if $x \in [s, s + \Delta s]$ and zero otherwise. If the system is ergodic and the dynamics allows an equilibrium distribution at an inverse temperature β, the knowledge of $P(s)$ allows to define the free energy of the system in terms of the reduced coordinate s:

$$F(s) = -\frac{1}{\beta} \ln P(s) \tag{2.7}$$

allowing the description of the equilibrium properties of the system as a function

of a relevant and smartly chosen set of variables. For instance, if one is interested in the binding of two small molecules the distance between the corresponding centers of mass can be a good reaction coordinates to describe and to study the free energy profile of the process.

2.2.2 Computing the Free Energy

Different methods for computing free energy profiles have been developed. For example, umbrella sampling [139] is a commonly used method to enhance the sampling of rare events, in which the normal dynamics of the system is biased by a suitably chosen bias potential $V_B(s(x))$ that depends on x only via $s(x)$ and works as an extra term added to the potential $V(x)$. For an appropriate choice of the bias potential the system will be sampled more efficiently than in the normal case. The biased probability distribution is

$$P_B(x) = \frac{1}{Z_B} e^{-\beta(V(x)+V_B(s(x))} \tag{2.8}$$

where Z_B is the canonical partition function for the potential $V(x)+V_B(x)$. Measuring a probability distribution in the presence of a bias $V_B(s(x))$ will provide a measure for the unbiased free energy and for the unbiased probability distribution. As in eq. 2.4 we can evaluate $P_B(s)$:

$$\begin{aligned} P_B(s) &= \frac{1}{Z_B} \int dx e^{-\beta(V(x)+V_B(s(x)))} \delta(s - s(x)) = & (2.9) \\ &= \frac{Z}{Z_B} e^{-\beta V_B(s)} \frac{1}{Z} \int dx e^{-\beta V(x)} \delta(s - s(x)) = & (2.10) \\ &= \frac{Z}{Z_B} e^{-\beta V_B(s)} P(s) & (2.11) \end{aligned}$$

and consequently $P(s)$

$$P(s) = \frac{Z_B}{Z} e^{\beta V_B(s)} P_B(s) \tag{2.12}$$

So using eq. 2.7 we have:

$$F(s) = -\frac{1}{\beta} \ln P_B(s) - V_B(s) + f_B \tag{2.13}$$

where $f_B = \frac{1}{\beta} \ln \frac{Z}{Z_B}$ is a constant which doesn't depend on s. Finally $P_B(s)$ can be estimated as in eq. 2.6 on the biased trajectory.

It can be shown[139] that the optimal efficiency is obtained when the biased potential is $V_B(s(x)) = -F(s)$, because the resulting free energy landscape experienced by the system will be flat, able to diffuse in a barrierless conformational space. Unfortunately in real systems $F(s)$ is not known, so the main problem that arises is how to construct $V_B(s(x))$ without a detailed knowledge of the system. In order to solve this problem, an efficient strategy to apply is the weighted histogram method (WHAM) [140, 141], in which several histograms, constructed with different umbrellas $V_{Bi}(s(x))$, are combined in order to reconstruct a single estimate of $F(s)$. The principal limitation of these and similar methods is that the computational cost scales exponentially with the number of reaction coordinates used. This is the case also for history-dependent search methods, such as local elevation [137], Wang-Landau sampling [142]and metadynamics [30], which allow a free energy reconstruction only as a function of a few variables.

2.3 Metadynamics

Metadynamics is a computational technique aimed at enhancing the sampling of the conformational space of complex molecular systems [30]. Conceptually it is a generalization of umbrella sampling. In fact the enhancement is obtained through a bias that acts on a small number of parameters, referred to as collective variables (CVs), $s(x)$, which provide a coarse-grained description of the system, and are explicit and differentiable functions of the Cartesian coordinates x. In the case of metadynamics the bias takes the form of a history-dependent potential constructed as a sum of Gaussian distributions centered along the trajectory of the CVs [31]

$$V_{\mathrm{G}}(s(x), t) = w \sum_{t'=\tau_{\mathrm{G}}, 2\tau_{\mathrm{G}}, \ldots} \exp\left(-\frac{(s(x) - s(x(t')))^2}{2\sigma_s^2}\right) \tag{2.14}$$

where the sum is for $t' < t$. Three parameters enter into the definition of V_{G}:
(i) the height w of the Gaussian distributions,
(ii) the width σ_s of the Gaussian distributions, and

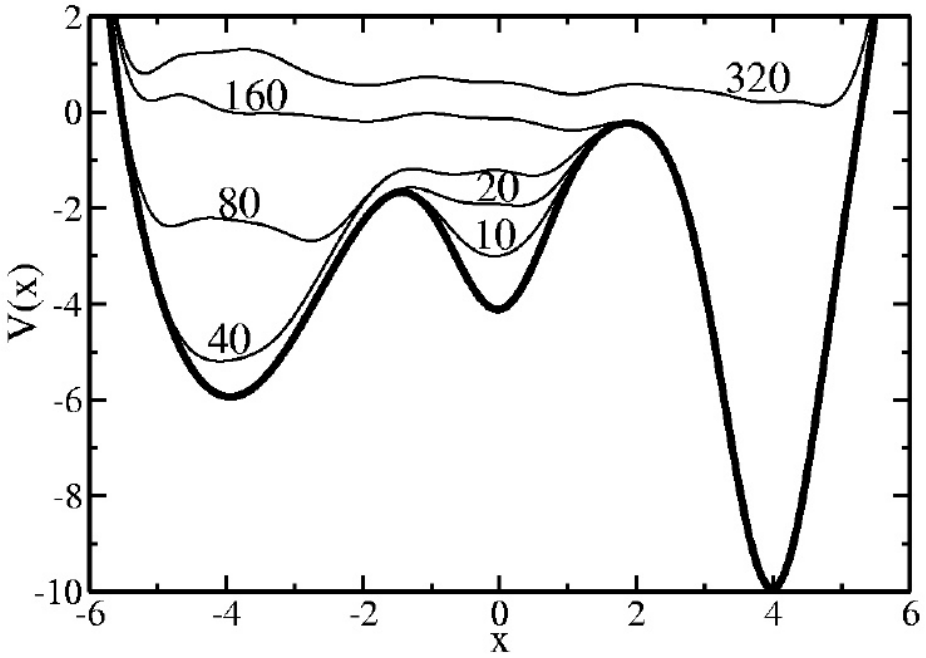

Figure 2.1: One dimensional example of the free energy profile filled by metadynamics bias potential. The different profiles correspond to the bias potential at $t = N\tau_{\mathrm{G}}$, where N is the number on top of each profile and represents the amount of Gaussians deposited until time t

(iii) the frequency τ_{G}^{-1} at which the Gaussian distributions are deposited.

These three parameters determine the accuracy and efficiency of the free energy reconstruction. If the Gaussians are large, the free energy surface will be explored at a fast pace, but the reconstructed profile will be affected by large errors. Instead if the Gaussian are small or are deposited infrequently, the reconstruction will be accurate, but it will take a longer time. Typically the width σ_s is chosen to be of the order of the standard deviation of the CV in a preliminary unbiased simulation in which the system explores a local minimum in the free energy surface [31]. The bias potential, in time, fills the minima in the free energy surface, allowing the system to efficiently explore the space defined by the CVs. Fig. 2.1 reports an example of how metadynamics acts with an one-dimensional bias potential.

2.3.1 Free energy estimate and convergence

For any choice of the parameter w, the bias potential $V_{\mathrm{G}}(s,t)$ can also represents an unbiased estimator of the free energy as a function of the reaction coordinate s. It has been demonstrated [143] that after a transient time t_{eq}, which corresponds to the time needed to fill all the relevant free energy minima of the system (in Fig. 2.1 $t_{eq} = 320\tau_{\mathrm{G}}$), $V_{\mathrm{G}}(s,t)$ reaches a stationary state in which it grows evenly fluctuating around an average. So if $V_{\mathrm{G}}(s,t)$ after t_{eq} is an unbiased estimator of $-F(s)$, at finite t the time average of $V_{\mathrm{G}}(s,t)$ defined as

$$\overline{V_{\mathrm{G}}}(s) = \frac{1}{t - t_{eq}} \int_{t_{eq}}^{t_{sim}} dt' V_{\mathrm{G}}(s,t'), \tag{2.15}$$

show deviations from $-F(s)$ which becomes smaller and smaller as t increases. This time average represents the best estimate of the free energy that can be obtained with metadyanmics.

2.3.2 Criticalities

To obtain an accurate description and free energy surface by this approach, the choice of CVs (see also 2.4.1) is fundamental. If an important variable is missing, the free energy estimate will be characterized by large fluctuations and errors. Moreover, as long as the CVs are uncorrelated, the time required to reconstruct a free energy surface for a given accuracy scales exponentially with the number of CVs, like in ordinary umbrella sampling. Therefore, the performance of the algorithm rapidly deteriorates as the dimensionality of the CV space increases. This makes impractical to obtain an accurate calculation of the free energy when the dimensionality is high. Unfortunately, this is often the case for complex reactions such as protein folding, in which it is very difficult to select a priori a limited number of variables which describes appropriately the process.

2.4 Bias-Exchange Metadynamics

The bias-exchange metadynamics (BE-META) method allows to overcome the difficulties discussed above [32]. The method is based a combination of replica

exchange [138] and metadynamics. Multiple metadynamics simulations of the system at the same temperature are performed. Each replica is biased with a time-dependent potential acting on a different collective variable. Exchanges between the bias potentials in the different variables are periodically allowed according to a replica exchange scheme. Moves are accepted with a probability:

$$P_{ab} = min(1, e^{\beta\left(V_{\mathrm{G}}^{a}(x^{a},t)+V_{\mathrm{G}}^{b}(x^{b},t)-V_{\mathrm{G}}^{a}(x^{b},t)-V_{\mathrm{G}}^{b}(x^{a},t)\right)}) \ . \tag{2.16}$$

If the exchange move is accepted, the trajectory that was previously biased in the direction of the first variable, continues its evolution biased by the second and vice versa. A 2-dimensional example is shown in Fig. 2.2. Two simulations are performed on two replicas of the system at the same temperature, respectively biased by x- and y-variable. From time to time, the two replicas are allowed to exchange configurations, accepting the exchange according to Eq 2.16. As a result, the metadynamics potential almost exactly compensates the free energy, both as a function of x and y.

By this approach, a relatively large number of different variables can be biased, and a high-dimensional space can be explored after a sufficient number of exchanges. Indeed the computational cost of adding an extra variable increases just linearly, not exponentially as in a normal metadynamics.

The result of the simulation is not a free-energy hyper-surface in several dimensions, but several (less informative) low-dimensional projections of the free energy surface along each of the CVs. The high-dimensional hyper-surface can still be reconstructed [144] using the method summarized in Section 2.4.3.

2.4.1 Choice of the Collective Variables in the BE

Similarly to other methods that reconstruct the free energy as a function of a set of generalized coordinates, in the BE-META the choice of the CVs plays an essential role in determining the convergence and efficiency of the free-energy calculation. If the chosen set of CVs does not distinguish different metastable states of the system, the simulation will be affected by hysteresis as not all of the important regions of the conformational space will be explored. To choose an appropriate set, one needs to exploit some basic knowledge on the topological,

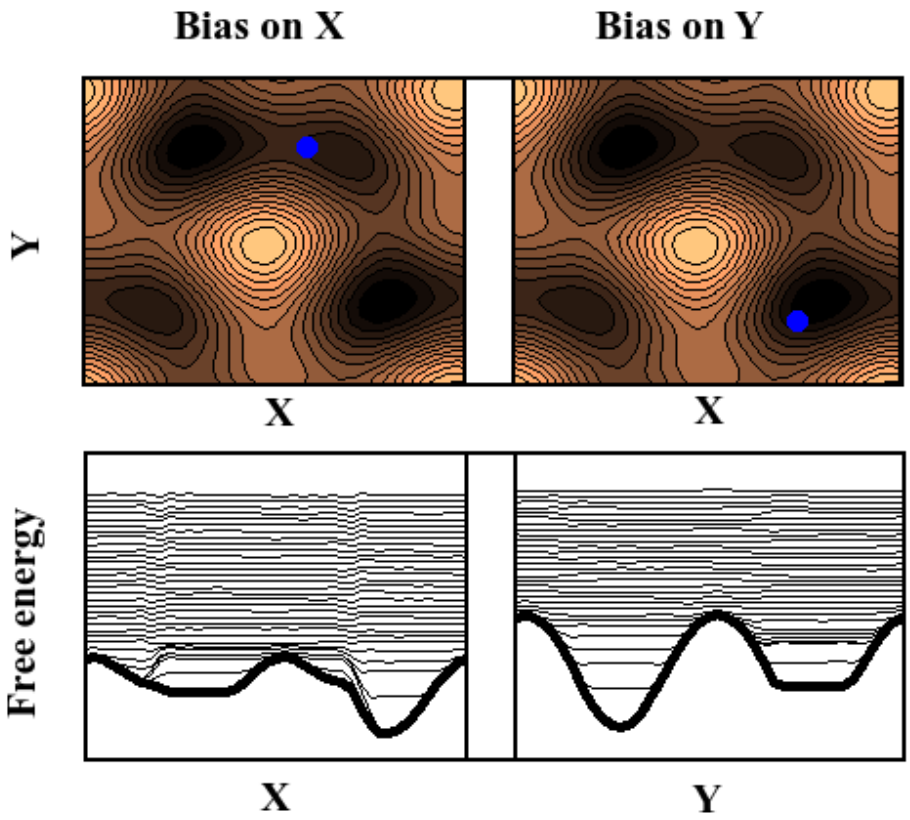

Figure 2.2: An example of a bias-exchange metadynamics reconstruction performed in a system with the potential depicted in the top panel. The simulation is performed with two replicas, biasing with metadynamics x and y respectively. $F(x)+V_{\mathrm{G}}(x,t)$ and $F(y)+V_{\mathrm{G}}(y,t)$, represented respectively in the bottom panels, are approximately flat at all times indicating that the CV jumps introduced by the exchange moves can efficiently lead to a good level of convergence in free energy calculation.

chemical, and physical properties of the system. Although there is no *a priori* recipe for finding the correct set of CVs, in the BE-META method the number of variables can be relatively large, making the selection less critical.

2.4.2 Choice of the BE Parameters

As mentioned in Section 2.3, the choice of the parameters w and σ_s influences the accuracy and efficiency of the free energy reconstruction. Artifacts tend to

arise when the free energy landscape is highly inhomogeneous, being characterized by the simultaneous presence of very shallow and very narrow free energy basins [145]. The parameters of the Gaussian distributions should be chosen in such a way that the maximum force introduced by a single Gaussian distribution (w/σ_s) is smaller than the typical derivative of the free energy [145]. To choose these parameters, we follow a scheme previously proposed in [31] and [146]. In particular, the width σ_s has to be chosen of the order of the standard deviation of the CV, performing several preliminary unbiased simulations starting from different folded and unfolded configurations, in which the system explored a local minimum in the free energy surface. Moreover the force introduced by a single Gaussian distribution should be smaller than the typical derivative of the free energy. In the cases studied in this thesis we verified that the choice of the parameters was correct (see Table 3.1 and Fig. 3.3 in Section 3.3). All the values of the parameters used in this work are reported in Section 3.3.

2.4.3 Free Energy Reconstruction

The BE-META method allows the free energy of the system to be reconstructed once the bias potentials reach convergence [32]. The convergence of the bias potential $V_{\mathrm{G}}(s,t)$ is monitored as in standard metadynamics [31, 143] (see Section 2.3.1) and is evaluated independently over the profile reconstructed by each replica. A lack of stationary fluctuations in the biased profile, typically due to the neglect of some important slow degree of freedom, is a signal that allows to improve the simulation setup. If an important degree of freedom is not explicitly biased, V_{G} does not converge to a stationary shape, and taking a time average of the potential of the form in Eq 2.15 is not meaningful [147]. In this case, one should analyze the trajectory and find the "hidden" variable responsible for the large fluctuations of V_{G}, and add it to the set of CVs.

In order to estimate the relative probability of the different states, the low-dimensional free energy surfaces (FES) obtained from the BE-META calculations are exploited to estimate, by a weighted-histogram procedure, the free energy of a finite number of structures representative of all the configurations explored by the system. The CV space is subdivided so that all the frames of the BE-

META trajectories are grouped in sets (microstates) whose members are close to each other in the CV space [144]. Since the scope of the overall procedure is to construct a model to describe the thermodynamic and kinetic properties of the system, it is important that the microstates are defined in such a way that they satisfy three properties: i) the microstates should cover densely all the configuration space explored in BE-META, including the barrier regions; ii) the distance in CV space between nearest neighbor microstates centers should not be too large; iii) the population of each microstate in the BE-META trajectory has to be significant, otherwise its free energy estimate will be unreliable. A set of microstates that satisfy these properties is defined dividing the CV space in small hypercubes forming a regular grid. The size of the hypercube is defined by its side in each direction: $ds = (ds_1, ds_2, ..., ds_n)$ where n is the number of collective variables used in the analysis. This procedure determines directly how far the cluster centers are in CV space. Each frame of the BE-META trajectory is assigned to the hypercube to which it belongs and the set of frames contained in a hypercube defines a cluster.

The free energy F_α of each microstate α is estimated by a weighted-histogram analysis approach (WHAM) [139, 144], conceptually similar to what has been described in eq. 2.11 and 2.13 . In the WHAM approach, the effect of the bias is removed, thus resulting in the free energy of a finite number of clusters that are representative of all the configurations explored by the system. The free energy of a cluster α is given as

$$F_\alpha = -T \log \frac{\sum_i n_\alpha^i}{\sum_j e^{\frac{1}{T}\left(f^j - V_\alpha^j\right)}} \tag{2.17}$$

where n_α^i is the number of times the microstate α is observed in the trajectory i and V_α^i is the bias potential acting on microstate α in the trajectory i. As described by eq. 2.15 V_α^i is estimated as the time average of the history-dependent potential acting on the trajectory i, evaluated in s_α, the center of cluster α

$$V_\alpha^i = \overline{V_{\mathrm{G}}^i}\left(s_\alpha\right) = \frac{1}{t_{sim} - t_{eq}} \int_{t_{eq}}^{t_{sim}} dt' V_{\mathrm{G}}^i\left(s_\alpha, t'\right). \tag{2.18}$$

where t_{sim} is the total simulation time and t_{eq} is the time after which the bias potentials converge. The normalization constants f^j appearing in Eq. 2.17 are determined self-consistently like in the standard WHAM method [144]. Corrections taking into account the variation of the bias over different structures assigned to the same cluster α have also been described previously [144].

Once the free energy of the different microstates has been obtained, the ensemble average of any observable that is function of the atomic coordinates of the system can be calculated as

$$\langle O \rangle = \frac{\sum_\alpha O_\alpha e^{-F_\alpha/T}}{\sum_\alpha e^{-F_\alpha/T}} \tag{2.19}$$

where the sums run over all the bins, O_α is the average value of the observable in the bin α, that is usually estimated as an arithmetic average on all the configurations belonging to the bin.

An important issue is how many and which CVs should be used in the clustering procedure. It is not necessary to use all the CVs that have been explicitly biased in one replica, as some of these CVs might prove to be a posteriori less relevant for the process, or strongly correlated with other variables. The variables used for the cluster analysis must provide an accurate and effective description of the system. An accurate description entails a set of clusters where each member contains consistently similar structures, and thus with very similar free energy. If the variables are too few, a cluster will contain structures that are very different from each other. On the other hand, performing the analysis in a very high-dimensional CV space will lead to poor statistics. Finally, it is possible to estimate the error of the free energy reconstruction looking at the root-mean-square deviations of the two profile of V_{G} in the first and second half of the simulation after the t_{eq}, as shown by the errorbars in Fig. 3.3.

All the analysis has been done by METAGUI[148], a graphic user-interface for VMD [149] for analyzing meta- and molecular dynamics simulations, visualizing the structures assigned to each microstate for different choices of the CVs.

Chapter 3

NMR-guided Metadynamics: Methods and Details

In this thesis we introduced an approach for integrating experimental data (in particular NMR chemical shifts) in MD simulations to drive the sampling of a biomolecular system, combining the experimental information with a powerful enhanced sampling technique, bias exchange-metadynamics, which allow reconstructing also the free-energy landscape. In this way, we were able to address a very challenging problem as protein folding, usually extremely resources-demanding, exploiting ordinary computational power, accelerating the sampling by 2-3 orders of magnitudes in the case of the GB3 protein[16, 150] (see Chapter 4 and 5). Moreover this enables a high-quality structural description of the different states of the protein, in particular for the folded state corresponding to the chemical shifts, designing NMR-guided metadynamics as a perspective method also for chemical shifts-based structure determination.

In this chapter we will focus on the implementation of this method, describing the sets of collective variable suitable for a protein folding study, highlighting in particular the novel collective variable introduced here, based on the chemical shift predictor Camshift (see Section 3.1.1). We will describe also a slightly different implementation of the general approach, Restrained metadynamics (Section 3.2), which can be very useful and computationally cheaper if one is interested only in characterizing the structure corresponding to the given experimental data,

rather than in the free energy reconstruction.

3.1 NMR-guided Metadynamics

This method is based on the introduction, within a BE-metadynamics framework, of a new collective variable (Camshift CV) which is able to quantify the overlap with known experimental chemical shifts of the structures sampled during the simulation, This approach is immediately generalizable to any other experimental observables (for instance residual dipolar couplings, cryo-electron microscopy map, fluorescence resonance energy transfer or small-angle X-ray scattering), provided that it is possible to predict or back-calculate the corresponding observable as a differentiable function of the atomic coordinates at each step of the simulation. Differently from other methods which modify directly the force field adding harmonic-like terms to favor the sampling of structure consistent with the experimental data [151, 27], the free energy resulting from our approach represents properly the statistical weights of the underlying force field. This is very important also from the search-algorithm point of view, because it does not introduce any frustration in the sampling which could prevent the system to reach the right fold if large conformational rearrangement are required, allowing the system to explore configurations also with no overlap with the chemical shifts. Since the chosen implementation chosen (see Section 3.1.1) relies on the general agreement with the data, the technique can work also in the case of sparse, incomplete, or even partially incorrect experimental data (as in a wrong chemical shift assignment).

3.1.1 Camshift Collective Variable

To predict the NMR chemical shifts corresponding to a given structure we used the Camshift method [152], which is based on an approximation of the chemical shifts as polynomial functions of interatomic distances. Unlike other previously developed methods for the semiempirical calculation of protein backbone chemical shifts [153, 154, 155], the functions used in Camshift are differentiable, thus allowing the forces to be computed and the CV to be defined as a penalty function

based on the differences between the experimentally measured and the calculated backbone chemical shifts ($^1H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$,^{13}C', ^1H_N, ^{15}N) [27, 29]. Since the chemical shifts are extremely sensitive to the details of the local configuration and environment of the atoms (see Section 1.3.1), the aim of this CV is to reproduce the local rearrangement of the protein compatible with the experimental data, expecially when it approaches low values, which corresponds to a better overlap between the predicted and the experimental chemical shifts. Even if the calculation of the chemical shifts is restricted to the backbone atoms, some contributions depends also on the orientation of the side-chains [152]. Therefore the forces applied by metadynamics to all the atoms involved in the calculation of a CV help also the slow transition of side-chain dihedral angles in finding the correct arrangement that is crucial to avoid a bad steric hindrance and to reach the correct fold. In the following chapters we will discuss also how this last variable is essential in the folding process to access the route towards the native basin of a protein and consequently to reach convergence in the free energy calculations.

Implementation of the Camshift CV

The implementation of Camshift as a collective variable requires the structure-based calculations of the chemical shifts. As described in [152] the chemical shift of a given atom is calculated as

$$\delta_{calc} = \delta_{coil} + \delta_{dihedrals} + \delta_{rings} + \delta_{backbone} + \delta_{side-chains} + \delta_{through-space} \tag{3.1}$$

where, δ_{coil} is a residue-dependent constant, and $\delta_{dihedrals}$ is calculated using the ϕ, ψ and χ_1 dihedral angles as

$$\delta_{dihedrals} = p_1 \cos(3(\theta + p_4)) + p_2 \cos(\theta + p_5) + p_3 \tag{3.2}$$

where p_i are empirical coefficients. The δ_{rings} term, which takes into account the ring current contributions, is defined using the classical point-dipole method [156]. The $\delta_{backbone}$, $\delta_{side-chains}$ and $\delta_{through-space}$ terms are defined as

$$\delta_X = \sum_{j,k} \alpha_{jk} d_{jk}^{\beta_{jk}} \tag{3.3}$$

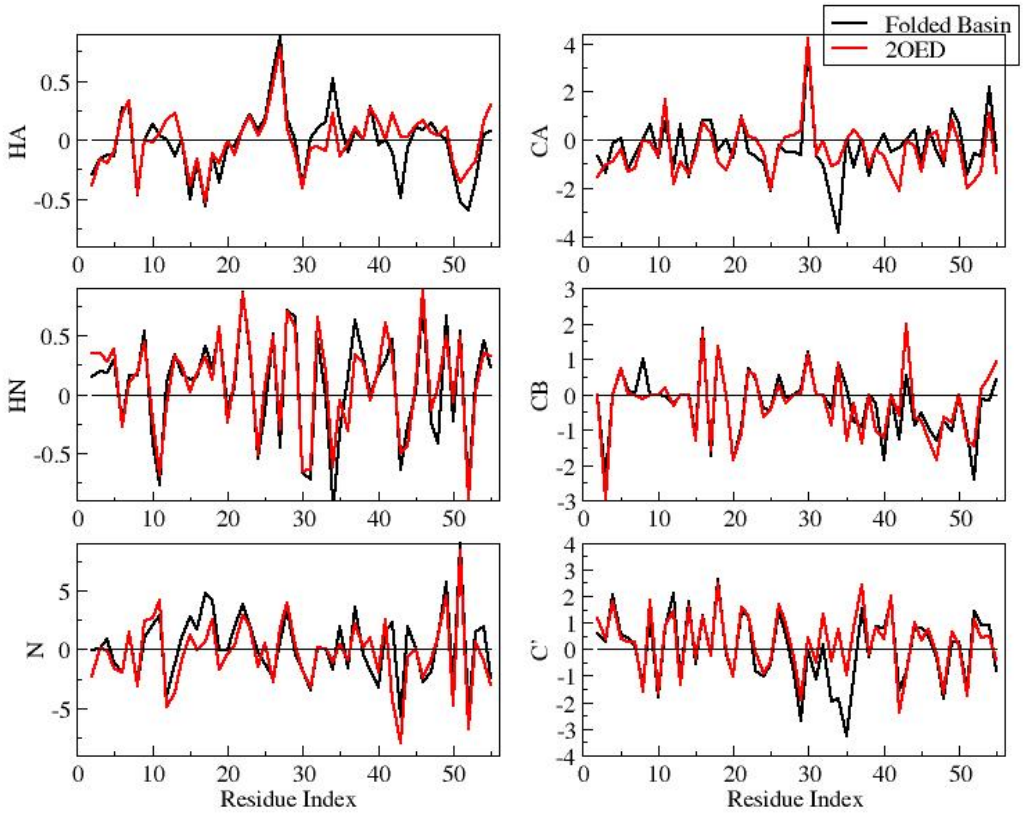

Figure 3.1: Difference between the chemical shifts calculated with the Camshift method [152] and the corresponding experimental values [157] for the structures in the free energy minimum (black line) of Fig. 5.1 (see Chapter 5) and for the experimental structure (PDB code 2OED, red line) reference. The values on the y axis are in ppm.

where j, k defines a pair of atoms at distance d; α and β are empirical coefficients. For $\delta_{backbone}$ the atoms are selected from the neighbouring residues along the chain, for $\delta_{side-chains}$ the atoms are those of the same residue, while for $\delta_{through-space}$ the atoms are selected among those within a radius of 0.5 nm and do not belong to the current and neighbouring residues. In Fig. 3.1 we report for a structure explored during the simulation and the experimental reference (PDB code 2OED), the difference between the chemical shifts calculated by the Camshift method and the corresponding experimental values [157] for the different atom types.

Since all these terms are defined as differentiable functions of the atomic

coordinates, it is possible to compute their derivatives and the corresponding forces in molecular dynamics simulations [27, 29]. The collective variable is then defined as

$$\text{Camshift} = \sum_{i=1}^{N} \sum_{j} E_{ij} \tag{3.4}$$

where i runs over the residues of the protein and j runs over the different atom types (H_α, H_N, N, C_α, C_β and C'). E_{ij} has the functional form [27, 29]:

$$E_{ij} = \begin{cases} 0 & \text{if} \quad |\delta^{ij}_{calc} - \delta^{ij}_{exp}| \leq n\epsilon_j \\ \left(\frac{|\delta^{ij}_{calc} - \delta^{ij}_{exp}| - n\epsilon_j}{\beta_j}\right)^2 & \text{if} \quad n\epsilon_j < |\delta^{ij}_{calc} - \delta^{ij}_{exp}| \leq x_0 \\ \left(\frac{x_0 - n\epsilon_j}{\beta_j}\right)^2 + \gamma \tanh\left(\frac{2(x_0 - n\epsilon_j)(|\delta^{ij}_{calc} - \delta^{ij}_{exp}| - x_0)}{\gamma\beta_j^2}\right) & \text{for } x_0 < |\delta^{ij}_{calc} - \delta^{ij}_{exp}| \end{cases} \tag{3.5}$$

where δ_{exp} and δ_{calc} are the experimental and calculated chemical shifts, respectively. The function E_{ij} has a flat bottom (Fig. 3.2) so that the chemical shifts calculated to within a given accuracy of the experimental value do not produce a penalty. The width of the flat region of the potential is determined by the term $n\epsilon_j$, where n is a tolerance parameter and ϵ_j is the accuracy of the Camshift predictions used for the chemical shifts of type j [152]. The penalty is harmonic until the deviation reaches a cutoff value x_0, at which point the penalty grows according to a hyperbolic tangent function defined to maintain a continuous derivative at the point x_0. The magnitude of the penalty is scaled for each chemical shift type j by the variable β_j, which is a function of the variability of that chemical shift in folded proteins reported in the Biological Magnetic Resonance Bank (BMRB) database [158]. The scaling factor β_j is used to obtain relative contributions of comparable magnitude of each chemical shift type to the CV value. The parameter γ determines how large the penalty can grow for deviations beyond x_0. In this investigation the simulation was run with $n = 1$ for all chemical shifts. The harmonic truncation point x_0 was set to 4.0 ppm for H_α and H_N, and 20.0 ppm for N, C_α, C_β and C′. The penalty truncation factor γ was set to 20 for all chemical shifts. These values of x_0 and γ result in an essentially harmonic penalty for most chemical shifts, with penalties only reaching the hyperbolic tangent region of the penalty function in the case of very large outliers [27, 29].

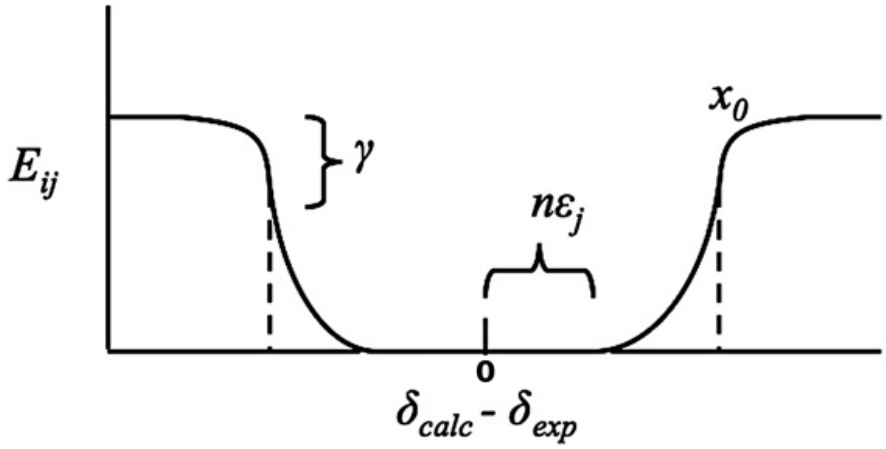

Figure 3.2: Graphical representation of the functional form of E_{ij} used to calculate the Camshift collective variable (adapted from [27])

The collective variable and the forces, which were derived analytically, have been implemented explicitly into a modified version of PLUMED [159] that will be made public in a future release.

3.1.2 Definition of the Collective Variables

In order to explore exhaustively the conformational space of a protein, together with Camshift we used several other CVs:

- **AlphaRMSD**, **ParaBetaRMSD**, **AntiBetaRMSD**. These CVs count how many fragments of 6 residues (6 in a row for α-helices and 3+3 for β-sheets) belong to an α-helix and β-sheet, by computing their RMSD with respect to an ideal α-helix and β-sheet conformation [160]:

$$S = \sum_{\alpha} n \left[\mathrm{RMSD} \left(\{\mathbf{R}_i\}_{i \in \Omega_\alpha}, \{\mathbf{R}^0\} \right) \right] \tag{3.6}$$

$$n\,(\mathrm{RMSD}) = \frac{1 - (\mathrm{RMSD}/0.1)^n}{1 - (\mathrm{RMSD}/0.1)^m} \tag{3.7}$$

  where n is a function switching smoothly between 0 and 1, the RMSD is measured in nm, and $\{\mathbf{R}_i\}_{i \in \Omega_\alpha}$ are the atomic coordinates of a set Ω_α of six residues of the protein, while $\{\mathbf{R}^0\}$ are the corresponding atomic positions

of ideal α-helical and β-sheet conformations; m, n are exponents that allow to tune the smoothness of the function.

- **Coordination Number**. This CV, which is used to quantify the number of contacts between the side chain heavy atoms of hydrophobic residues, is defined as

$$C_N = \sum_{i,j} C_{ij}$$

with

$$C_{ij} = \frac{1 - (\frac{r_{ij}}{r_0})^n}{1 - (\frac{r_{ij}}{r_0})^m} \tag{3.8}$$

where r_{ij} is the distance between atoms or groups i and j, r_0 is the distance value to consider two atoms in contact and m, n are exponents that allow to tune the smoothness of the function.

- **AlphaBeta Similarity**. We considered two CVs of this type, corresponding to the χ_1 and χ_2 side chain dihedral angles, respectively, for hydrophobic and polar amino acids. These CVs are designed to enhance the side chain packing searching, which is crucial for protein folding. The CVs are defined as

$$AB_{Sim} = \sum_i \frac{1}{2}[1 + cos(\chi_i - \chi_i^{ref})] \tag{3.9}$$

where we have chosen χ_i^{ref} as the mean value of the corresponding dihedral angle from a library of folded proteins extracted from the PDB.

- **Camshift CV**. This CV evaluates the overlap between experimental and calculated chemical shifts of the sampled structure. Its implementation is described in the previous Section 3.1.1.

The first three CVs act at the secondary structure level, by quantifying, respectively, the fraction of α-helical, anti-parallel and parallel β-sheet content of the protein. This allows a comprehensive sampling of the secondary structure for all the backbone conformations of the protein. The following three CVs act at the tertiary structure level through the exploration of the number of hydrophobic contacts and of the packing of the side chain dihedral angles χ_1 and χ_2 for hy-

drophobic and polar side chains. Combined together these CVs allows to rapidly open and close the core of the protein, exploring new internal sidechain arrangements of the protein, which are one of the slowest motion in protein dynamics. The coordination number, since it is extended to the hydrophobic sidechain atoms, quantifies also the degree of compactness of a structure, as we will discuss in 6.

All these CVs are implemented in PLUMED [159] for Gromacs [161]. They are publicly available and, apart from Camshift, have been used in several previous protein folding studies by metadynamics [144, 160].

3.2 Restrained Metadynamics

In this thesis we introduce also a variant of the general approach aimed at achieving a rapid and efficient search if one is only interested to find the structure corresponding to the experimental chemical shifts. It consists in a normal Bias-Exchange procedure using ordinary variables, adding a restraining potential acting on all the replicas. Since we are interested to bias the sampling towards structures consistent with the experimental chemical shifts, the restraining potential is the Camshift collective variable itself, applied to all the replicas. This determines an oriented bias also on the other biasing potentials which speeds up the conformational search towards the right fold. The approach is similar to the one used in [27], where a restrained molecular dynamics simulation was applied in a temperature replica exchange framework. In Section 4.3 we will show that exploiting the sampling power of the Bias Exchange strategy with traditional variables in combination with a coupling to the restraining potential do not frustrate the search and allows reaching better results in terms of accuracy due the direct use of Camshift bias in the sampling, and exploiting only 7 seven replicas instead of 21 of the method proposed in [27]. Also in comparison with the standard BE-META approach (see Section 4.1) we gain on order of magnitude in simulation time required to obtain the same structural accuracy, and even exploiting implicit solvent simulation.

The approach has been implemented introducing an umbrella (see Section 2.2.2) on the Camshift CV which acts simultaneously with the metadynamics bias

potentials on each replica of the Bias-Exchange simulation. Since the functional form of the collective variable (eq. 3.1.1 and Fig. 3.2) is already similar to a positively defined harmonic penalty, we used just a linear coupling to avoid too strong forces which can lead to frustration and numerical instabilities. Thus the bias acting on all the replicas is

$$V_B(s(x)) = m(s(x) - s_0) \tag{3.10}$$

where $s(x) = s(x(t))$ is the value of the Camshift CV for a specific configuration at time t, and s_0 is the reference value, that we set to zero in order to push the system towards the best agreement with the experimental data.

Within this approach we renounce to the free energy reconstruction because, albeit theoretically feasible, sampling fluctuations in the exploration of structures with large values of Camshift would dominate the reconstruction because of their large weight $e^{\beta V_B(s)}$ (see eq. 2.12).

3.3 Simulations Details

In this section we provide the simulation details for the simulations described in Chapter 4, 5 and 6 for the third immunoglobulin-binding domain of streptococcal protein G (GB3), and the 40-residue form of Amyloid beta (Abeta 40), for which we were able to characterize the free energy landscape, and for the CASD-NMR targets where we applied the Restraint Metadynamics approach to determine the structure corresponding to the experimental chemical shifts. As force fields for our molecular dynamics simulations we have used AMBER99SB-ILDN [127] and CHARM22* [11] which are considered among the best force fields for protein folding simulation [128]. The choice of the force field has been made based on the state-of-the-art at the time we started the simulations. In particular we exploited AMBER99SB-ILDN for well-structured proteins (GB3 protein and CASD-NMR target) and CHARM22* for Abeta40, considering also that the latter force field was successfully employed in an extensive molecular dynamics simulation of an intrinsically disordered protein [162].

Collective Variable (s)	w/σ_s (kJ/mol)	$\overline{\lvert\partial F/\partial s\rvert}$ (kJ/mol)
Camshift (σ_s=1)	0.3	3.4
Camshift (σ_s=0.5)	0.6	3.4
Coord. Number (σ_s=10)	0.03	0.35
ParaBeta-RMSD (σ_s=0.1)	3.0	13.7
AntiBeta-RMSD (σ_s=0.2)	6.0	13.0

Table 3.1: Comparison between the maximum force introduced by a single Gaussian ($\frac{w}{\sigma_s}$) and the average of the derivative of the free energy with respect to the specific collective variable ($\overline{\lvert\frac{\partial F}{\partial s}\rvert}$). w is equal to 0.3 kJ/mol.

3.3.1 GB3 Protein

GB3 is a protein of 56 residues characterized by a fold with two β-hairpins, located respectively in the N- and C-terminus, bridged together in a parallel β-strand, and a central α-helix, which connects the two hairpins (see Fig. 4.1B in the next chapter). Starting from an unfolded state at 5.7Å from the reference structure of GB3 protein (PDB 2OED [163]), obtained by a simulated annealing procedure, we ran 380 ns of molecular dynamics simulation enhanced by a BE-META scheme for 7 different replicas ($\sim 2.7\mu$s in total), each one biased by a different history-dependent potential acting on one of the collective variables described in Section 3.1.2:

- **Camshift** Parameters: Gaussian width $\sigma = 1$.
- **AlphaRMSD**. Parameters: $m = 4$, $n = 2$, R_0=0.08, σ =0.2
- **ParaBetaRMSD** Parameters: $m = 12$, $n = 8$, R_0=0.08, σ =0.1
- **AntiBetaRMSD**. Parameters: $m = 12$, $n = 8$, R_0=0.08, σ =0.2.
- **Coordination Number**. Parameters: $m = 8$, $n = 4$, R_0=0.4, σ =10
- Two **AlphaBeta Similarities**. Parameters: σ =0.5 for both replicas.

The simulations were performed using the GROMACS 4.5.3 package [161] employing the AMBER99SB-ILDN force field [127] and the TIP3P water model

[164]; the protein was solvated by 6524 water molecules in a 212 nm^3 cubic periodic box. The particle-mesh Ewald method [165] was used for long-range electrostatic with a short-range cutoff of 1 nm. A cut-off was used for Lennard-Jones interactions at 1.2 nm. All bond lengths were constrained to their equilibrium length with the LINCS algorithm [166]. The time step for the molecular dynamics simulation was set at 2.0 fs and the Nose-Hoover thermostat [131, 132] with a relaxation time of 1ps was used. The atomic coordinates and the energy were saved every 1 ps. Concerning the metadynamics setup, one-dimensional Gaussian functions of height$w = 0.30$ kJ/mol were added every 4 ps, and exchanges of the bias potentials were attempted every 20 ps.

After 120 ns of simulation, in which very wide regions of the CVs were explored, we introduced loose upper boundaries to help the convergence of the bias potentials [147], to not waste time in meaningless region of the CVs space. At this time we also reduced to 0.5 the Gaussian width of the Camshift CV, and doubled the σ of the AlphaRMSD CV. A second BE-META simulation of 300 ns was run with the same setup on 6 replicas, excluding the Camshift CV, to benchmark the importance and the power of this CV to fold the protein. Finally a standard molecular dynamics simulation of 200 ns was performed to evaluate the stability of the intermediate state.

In Table 3.1 we checked that the choice of the metadynamics parameters were consistent with the explored free energy landscape. This is confirmed also by the corresponding shape of the free energy projections along the collective variables used in the analysis, showed in Fig. 3.3: the profiles are homogeneous and smooth, with the minima wider than Gaussian width.

The implicit solvent simulations described in Section 4.3.1, has been performed with the same setup, but by a stochastic dynamics with the Generalized Born Solvent Accessibility (GBSA, [167]) approximation based on the OBC algorithm. The starting configuration for the simulation was obtained with an equilibration procedure in implicit solvent at 400 K from the same configuration used in the explicit solvent simulation.

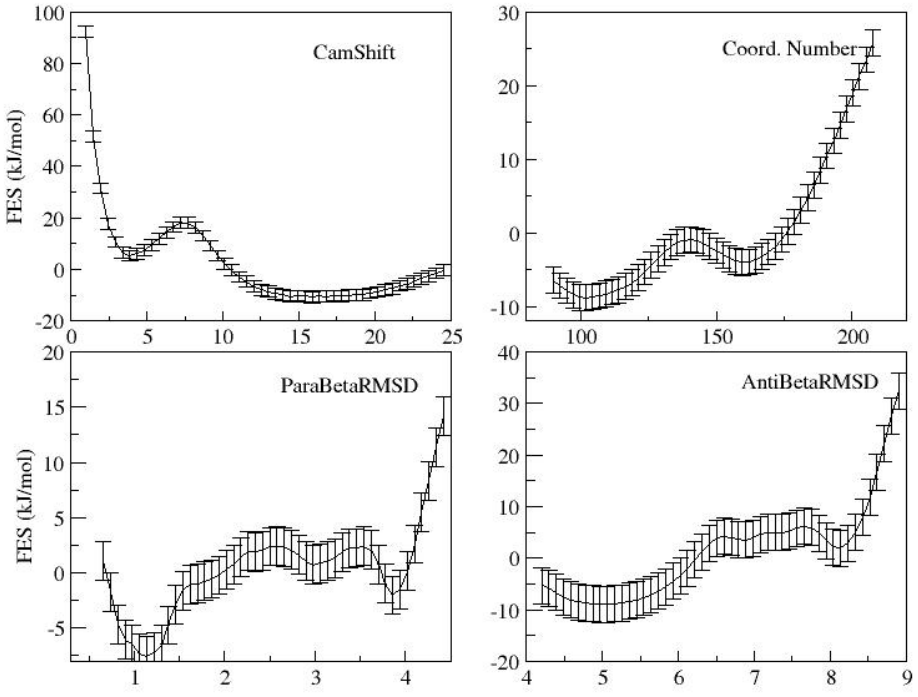

Figure 3.3: Convergence of the free energy surfaces as function of different collective variables for GB3. The values on the x axis are relative to the corresponding collective variable (CV) on the top right corner of each panel. The explicit forms of the CVs are reported in Section 3.1.2

3.3.2 Abeta40

Amyloid beta is a peptide of 36-43 amino acids that is processed from a bigger transmembrane glycoprotein, the amyloid precursor protein (APP). It usually originated by sequential cleavage of the APP, and the 40-residue form of Amyloid beta (we will refer to it as Aβ40 or Abeta40) is the most common isomorfs. These kinds of peptides are well known for their propensity to aggregate into plaques composed of a tangle of ordered structures obtained by the regular superposition of parallel and anti-parallel β-strands, called fibrils, protein fold shared by other peptides associated with protein misfolding disease. After a rapid growing these aggregates becomes usually toxic for the cells. In particular the amyloid plaques are the main component of the deposits found in the brains of patients affected

by Alzheimer's disease.

To characterize the free energy landscape of Abeta40, we used a setup similar to the previous one, performing BE-metadynamics simulation of 310 ns on 8 replicas ($\sim 2.5\mu$s in total), the same as the previous case, just adding an extra replica biased by Camshift CV:

- **Camshift** Parameters: Gaussian width $\sigma = 1$.
- **AlphaRMSD**. Parameters: $m = 12$, $n = 6$, R_0=0.1, σ =0.2
- **ParaBetaRMSD** Parameters: $m = 12$, $n = 6$, R_0=0.1, σ =0.1
- **AntiBetaRMSD**. Parameters: $m = 12$, $n = 6$, R_0=0.1, σ =0.2.
- **Coordination Number**. Parameters: $m = 8$, $n = 4$, R_0=0.4, σ =10
- Two**AlphaBeta Similarities**. Parameters: σ =0.3 for hydrophobic residues and σ =0.075 for polar residues.

The GROMACS 4.5.3 package [161] was used, employing the Charmm22* force field [11] and the TIP3P water model [164]; the protein was solvated by 6461 water molecules and 3 sodium ions in a 203 nm^3 dodecahedron periodic box. The particle-mesh Ewald method [165] was used for long-range electrostatic with a short-range cutoff of 0.9 nm. A cut-off was used for Lennard-Jones interactions at 0.9 nm. All bond lengths were constrained to their equilibrium length with the LINCS algorithm [166]. The time step for the molecular dynamics simulation was set at 2.0 fs and the Nose-Hoover thermostat [131, 132] with a relaxation time of 1ps was used. The atomic coordinates and the energy were saved every 1 ps. Concerning the metadynamics setup, one-dimensional Gaussian functions of heigth $w = 0.30$ kJ/mol were added every 5 ps, and exchanges of the bias potentials were attempted every 20 ps. Also in this case, after 100 ns of simulation, in which very wide regions of the CVs were explored, we introduced loose upper boundaries to help the convergence of the bias potentials [147]. The experimental chemical shifts [168] were taken from the BMRB databank [158], code 17795.

3.3.3 CASD-NMR Target

The simulations for the CASD-NMR target HR2876C, an unsolved 97-residue long protein (PDB reference 2M5O, released after our model submission), described in 4.3.2, has been performed by a stochastic dynamics employing the Generalized Born Solvent Accessibility (GBSA, [167]) approximation based on the OBC algorithm, with the same setup used for the GB3 simulations (AMBER99SB-ILDN has been employed). The starting configuration for the simulation was obtained by homology modeling from the amino acid sequence, using the HHpred server [91] and Modeller [101, 102].

Chapter 4

Structure Determination using Chemical Shifts

As we have discussed in Section 1.2.2, one of the field where computational approaches to biological issues played and still play a major role is in the prediction and determination of the three dimensional structure of a protein. In particular, these approaches have been really useful in simplifying and speeding up the experimental protocols for structure determination in X-ray crystalography and NMR spectroscopy.

In this chapter we present the first results of the application to plain structure prediction of the methodology introduced in the previous chapter. Starting from the simple knowledge of the backbone chemical shifts ($^1H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$,^{13}C', ^1H_N, ^{15}N), we show how it is possible to combine this experimental information with an enhanced sampling technique based on molecular dynamics simulations to determine the protein structure in NMR spectroscopy with the same atomistic accuracy of an experimental structure. The application of this method could avoid lots of time and resource demanding experiments for NOESY assignment and structural restraints extraction, necessary in the standard structure generation procedure, improving the prediction power of this experimental technique, especially for larger proteins. Moreover the molecular dynamics framework provides also a "natural" way for the acquisition of the native conformation exploiting its folding mechanism (see also chapter 5), and guarantees by itself a high quality of

the stereochemical parameters, avoiding directly bad steric clashes and dihedral angles violations by the force field.

In Section 4.1 we apply our approach to the case of a GB3 protein, showing in particular how the use of the new collective variables based on the chemical shifts (Camshift CV, see Section 3.1.1) is crucial to properly fold the protein. In the attempt to validate the sampled model of the protein we were able also to establish a new way to assess and rank the quality of different NMR models (Section 4.2), demonstrating that we would been able to identify correctly the structure which is more consistent with the experimental data even if we didn't know the reference PDB structure in advance. Finally we discuss a slight modification of the general approach (NMR Restraint metadynamics) which is more computationally efficient for structure determination and refinement, showing the results of a successful application to a blind test on a CASD-NMR target.

4.1 Folding of GB3 using Chemical Shifts as Collective Variables

We performed molecular dynamics simulations of GB3 at 330 K in explicit solvent with the Amber99SB-Ildn force field [127]. In order to enhance conformational sampling, we used the NMR-metadynamics scheme described in Section 3.1 and [150], based on Bias-Exchange approach, with 7 replicas. We started the simulations from a structure at 5.7 Å from the reference structure (PDB code 2OED [163]), and run them for a total of 380x7 ns ($\sim 2.7\mu$s in total). For each replica we used a different metadynamics history-dependent potential acting on a different collective variable (see Section 2.4 and 3.1.2). Three CVs act at the secondary structure level, by quantifying, respectively, the fraction of αhelical, anti-parallel and parallel β-sheet content of the protein. Other three CVs act a the tertiary structure level through the exploration of the number of hydrophobic contacts and of the packing of the side chain dihedral angles χ_1 and χ_2 for hydrophobic and polar side chains. The seventh CV, Camshift (see Section 3.1.1), measures the difference between the experimental and calculated chemical shifts. The detailed setup is described in Section 3.3.1

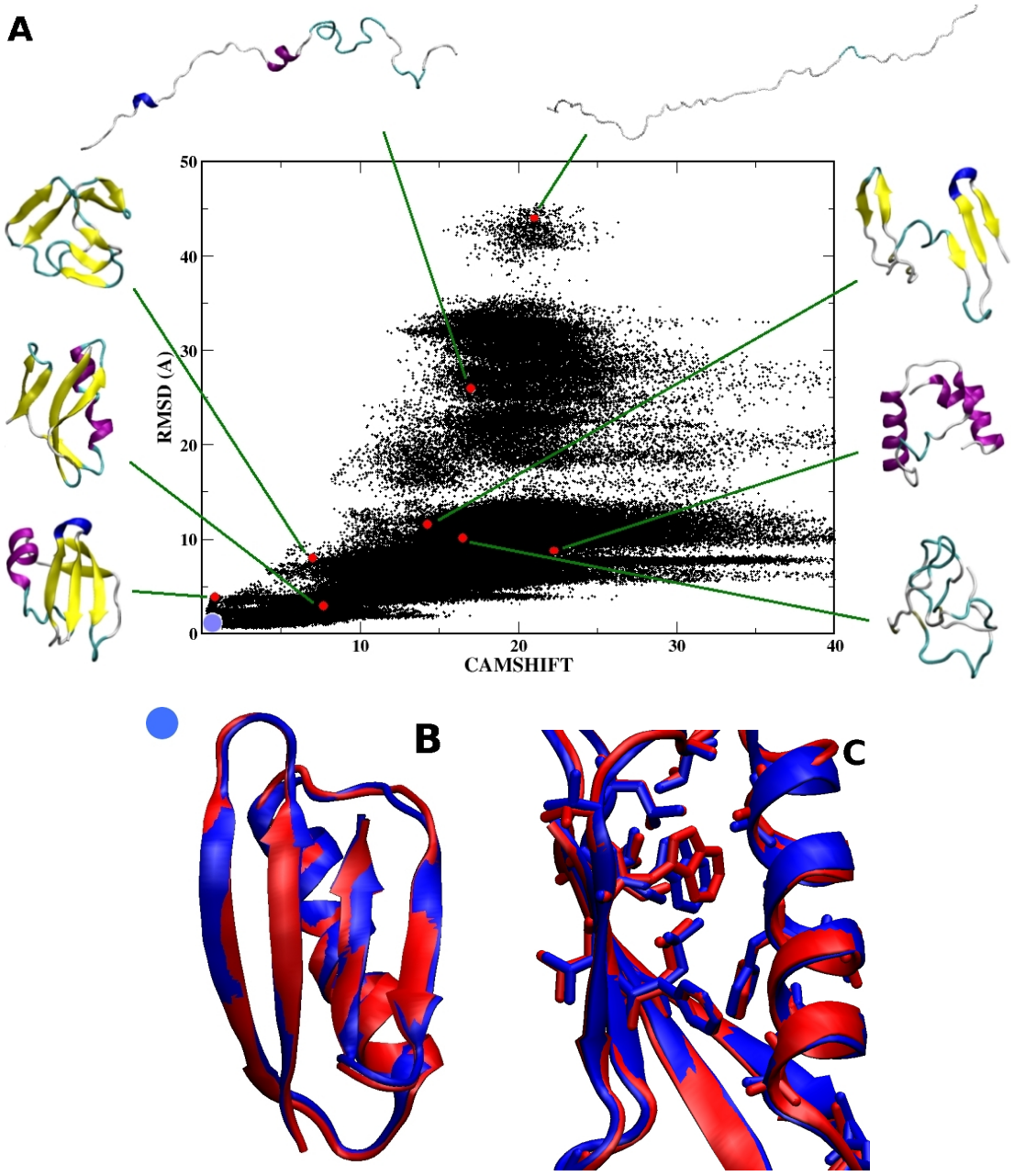

Figure 4.1: A) Representation of the conformational sampling achieved by the approach introduced in this work. The conformations visited are shown as function of the Camshift collective variable and of the backbone RMSD from the reference structure (PDB 2OED). B) Structure with the lowest RMSD (0.5 Å) from the reference structure. C) Illustration of the accuracy of the side-chain packing of the structure in (B).

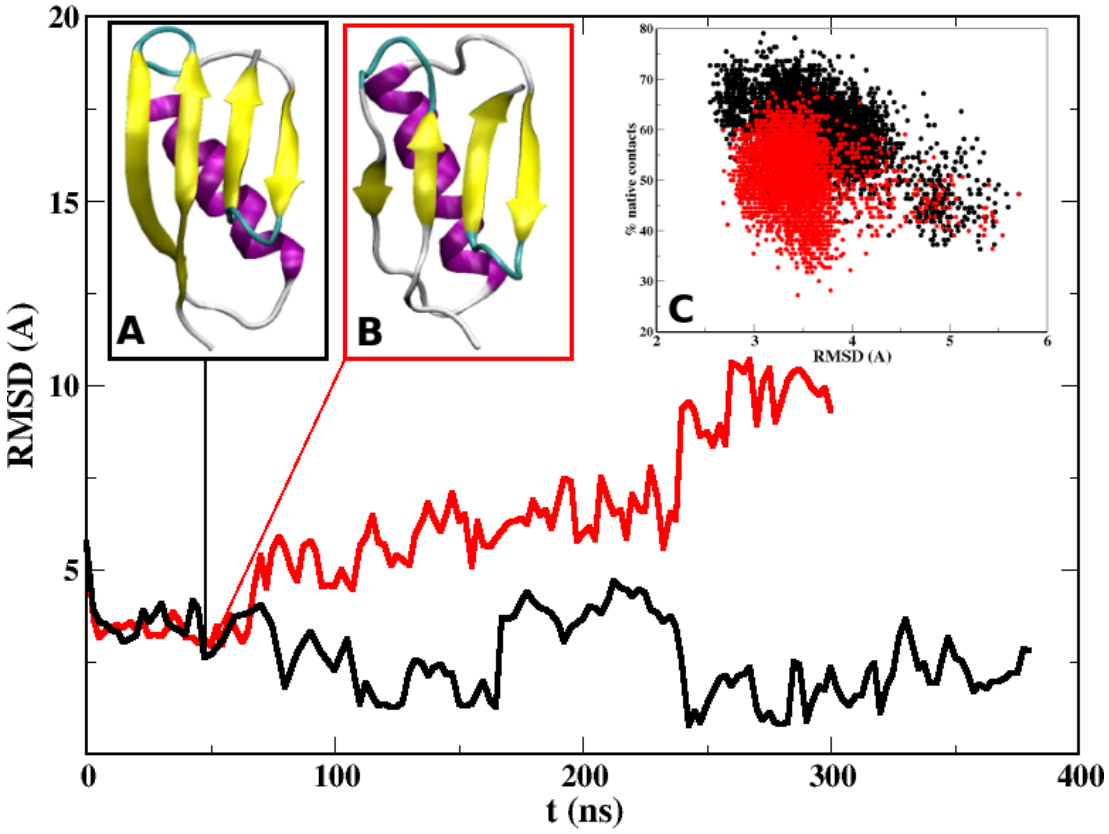

Figure 4.2: Time series of the trajectories that achieve the lowest RMSD value for the simulations with (black line) and without (red line) the Camshift CV. Insets: A and B) Lowest RMSD structures, and C) Percentage of native contacts in each conformation in the first 50 ns in the two simulations.

Our method allows to explore a wide range of structures, ranging from extended to compact. Representative examples are shown in Fig. 4.1A. Native-like conformations are visited multiple times, reaching a backbone RMSD of 0.5Å from the reference structure (PDB code 2OED). In these native-like structures the internal packing of hydrophobic side-chains is practically identical to that observed in the reference structure (Fig. 4.1C). In the calculations that we performed, this level of accuracy could be reached only by using a Bias-Exchange scheme in which the Camshift CV is included in the collective variables set. To demonstrate this point, we performed another simulation with the same setup, using the six CVs discussed above that describe the secondary and tertiary structures, but not the Camshift CV. The difference between the two simulations is

substantial. In the simulation without the Camshift CV the best configuration has a RMSD of 2.7Å (Fig. 4.2, inset B). After 50 ns the RMSD starts increasing constantly (red line) and the folded state is not explored at all. By contrast, the simulation with the Camshift CV visits several times the folded state, with several unfolding-refolding events. During the first 50 ns, not only the latter simulation performed better, reaching a RMSD of 2.5Å, but it also forms the correct secondary and tertiary contacts, in particular the ones involved in the formation of the first β-hairpin (Fig. 4.2, inset A), which is critical for the folding of this protein[169, 170]. The fraction of native contacts is also systematically higher in the simulation employing Camshift CV (Fig. 4.2, inset C). These results indicate that the folding events observed later in the simulation are due to the systematic bias induced by the Camshift CV variable towards the correct local topology in the native state. This is indeed the key of the success of the methodology: the rapid formation of native contacts induced by the chemical shifts bias allows to flatten the rugged free energy landscape and consequently to overcome barriers and kinetic traps, pushing the folding process downward the funnel. This picture is also confirmed by the correlation between Camshift CV and RMSD in Fig. 4.1A and 4.3 and by the results observed in Section 4.3.

4.2 A New Scoring Scheme for NMR Models

The correlation between Camshift CV and RMSD reveals that the new CV can be used also as a scoring function to assess and rank the quality of different NMR models. This would have allowed us to recognize the right fold for the GB3 protein even if the reference structure were unknown, as in a blind structure prediction starting only from the chemical shifts assignment.

The thermal average of Camshift CV allows ranking correctly a structure.

In Figure 4.3-A we plot the RMSD from the PDB structure against the value of Camshift CV for an ensemble of structures generated by bias exchange metadynamics. Clearly, all the conformations at low RMSD are also characterized by a

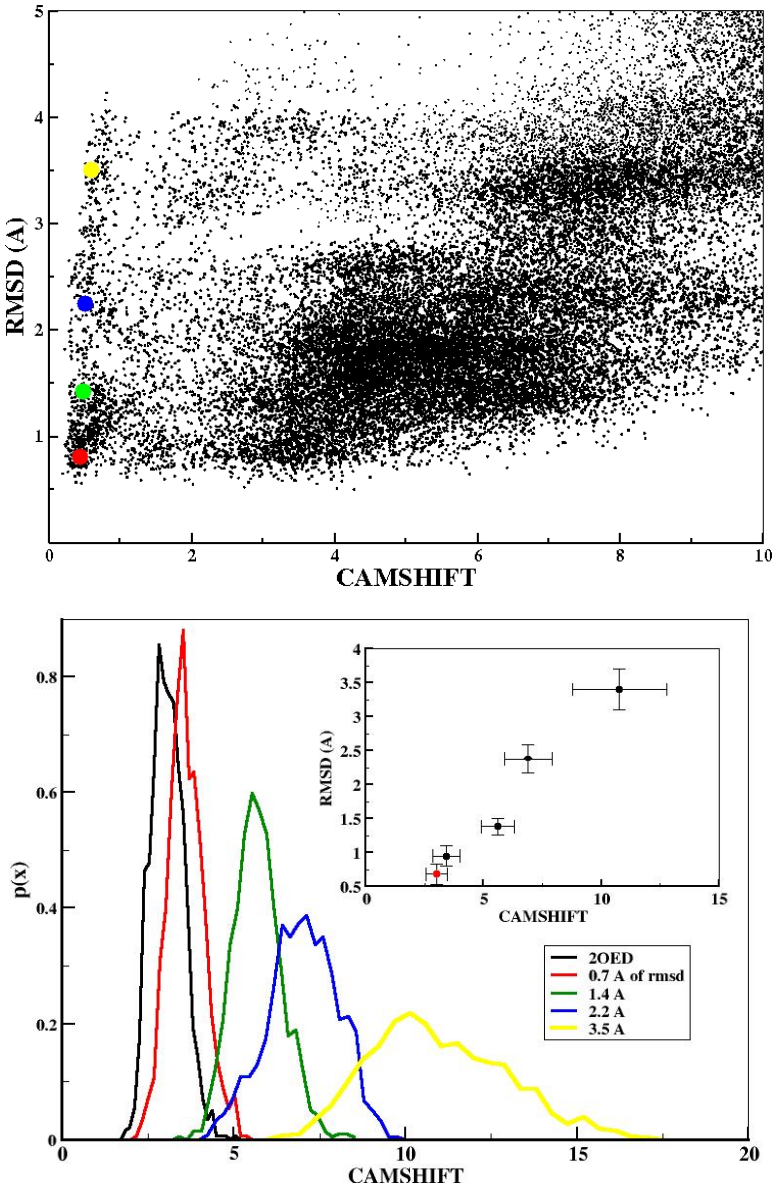

Figure 4.3: Above the detail of Fig.4.1 for low value of Camshift and selected structure for right fold discrimination. Below the probability distribution of Camshift score for free MD of 10 ns starting from the PDB reference (black) and the structures respectively at 0.7 (red), 1.4 (green), 2.2 (blue) and 3.5 (yellow) Å of RMSD

relatively low value of Camshift CV. However, we also observe some structures with a low value of similarity, but with a RMSD of 3 Å or more from the native state. Visual inspection reveals that these structure are topologically and locally similar to the native fold, but with differences involving, for instance, the packing of the hydrophobic core.

Thus, simply selecting from the simulation the configurations corresponding to a low Camshift value is not a sufficient criterion for localizing univocally the folded state if the reference structure is unknown. However, the structure corresponding to the folded state can be easily found by computing the probability distribution of Camshift CV in relatively short finite temperature MD runs started from the candidate structures. In order to demonstrate this point ee considered four different configurations with Camshift value smaller than 1 with different RMSDs from the reference (respectively at 0.7, 1.4, 2.2 and 3.5 Å, marked with different colors in Fig. 4.3-A) and we run a MD simulation of 10 ns starting from each of these structures. The probability distributions of Camshift in these four runs is shown in Fig. 4.3-B: clearly, the closer a structure is to the folded state, the lower on average its Camshift value is. This is also pointed out in the inset of Fig. 4.3-B, where the average of Camshift CV is plotted as a function of the average RMSD, which shown a striking linear correlation between the two variables. Remarkably, the probability distribution of the run started from the configuration at 0.7 Å (red line) is indistinguishable from the one of the run starting from the PDB coordinates (black line), indicating once again that the structures obtained by our procedure are statistically indistinguishable from the folded state, with an accuracy beyond the experimental resolution. It is worth to be noticed from the inset that also the variance of both Camshift and RMSD increase with the distance from the reference, as shown by the error bars: the further a structure is from the native state, the less stable it will be and viceversa.

The results shown in Fig. 4.3 clearly indicate that the value of Camshift in a single configuration can be affected by relatively large fluctuations, similar to those observed in the finite temperature value of any observable that depends on the position of several atoms, but it demonstrates also that structures can be unambiguously ranked by computing a thermal average of Camshift.

In our opinion this is a general result which is valid for all structure-based

scoring functions, as we have already shown in [96]. On one side this cures artifacts due to fluctuations of the atomic coordinates, quite common especially in the scoring of a large number of structures and since these functions usually rely on thresholds and cut-offs definitions; and more importantly the averaging procedure represents an ensemble average in the neighborhood of a local energy minimum which has a straightforward statistical physics meaning and takes also in consideration that proteins and their structures are dynamical and not static entities.

4.3 Structure Determination by Restrained Metadynamics

In order to establish a method only for protein structure determination purpose which is more competitive in terms of efficiency with the state-of-the-art prediction techniques from NMR data [111], we applied to GB3 the same setup described above but using implicit solvent approximation. This reduces drastically the degrees of freedom of the system (from $\sim$ 20000 to 862 atoms) but as we will see without success in the right fold determination. Then we show that the Restraint Metadynamics approach explained in Section 3.2, allows reaching this goal for GB3. The same technique has been applied successfully in a real blind test within CASD-NMR competition.

4.3.1 Implicit Solvent

We used the setup of the previous simulation for GB3 protein and described in Section 3.3 with the same 7 replicas, but with an implicit solvent approximation, Generalized Born Solvent Accessibility (GBSA, [167]), to reduce the computational cost. The simulation has been performed at a higher temperature T = 400 K to enhance the conformational transitions since usually folded proteins tends to exhibit a higher melting temperature in implicit solvent simulation [171], overstabilizing folded states [172]. As it is shown by the black line of Fig. 4.4, in 120 ns of simulation the standard Bias-Exchange approach involving Camshift CV is not able to fold properly the system, reaching in the best case structures at 2.6Å

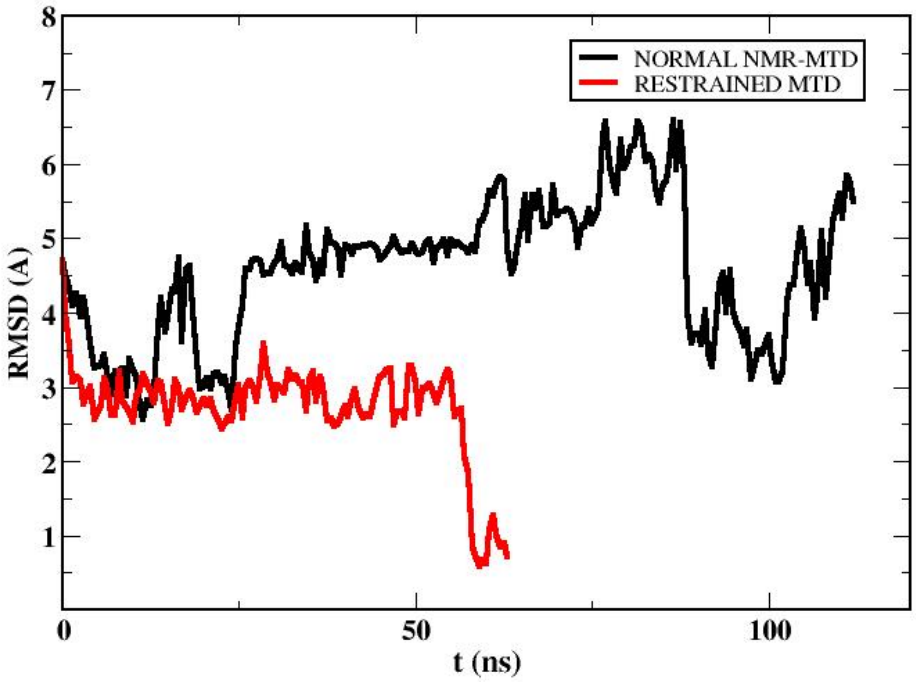

Figure 4.4: Time series of the trajectories that achieve the lowest RMSD value from the reference state (PDB 2OED) for the implicit solvent simulations of GB3 protein with standard NMR-metadynamics approach (black line) and with Restrained metadynamics (red line)

of backbone RMSD from the reference PDB just in the first part of the simulation and then drifting away from the folding basin, at variance with the previous simulation (Fig. 4.2). Even if the force field used is the same (Amber99SB-Ildn), the implicit solvent approximation has a negative impact on the dynamical and folding properties of the system, which is not able to form the right native contacts and, consequently, to access the folding basin. However adding a soft restraining potential on the Camshift CV on all the replicas, the system folds with the same accuracy of the explicit solvent simulation and in a shorter time (red line in Fig. 4.4), respectively at 60 and 240 ns for the same structural similarity. Despite of the implicit solvent approximation, the great increase in the efficiency of the

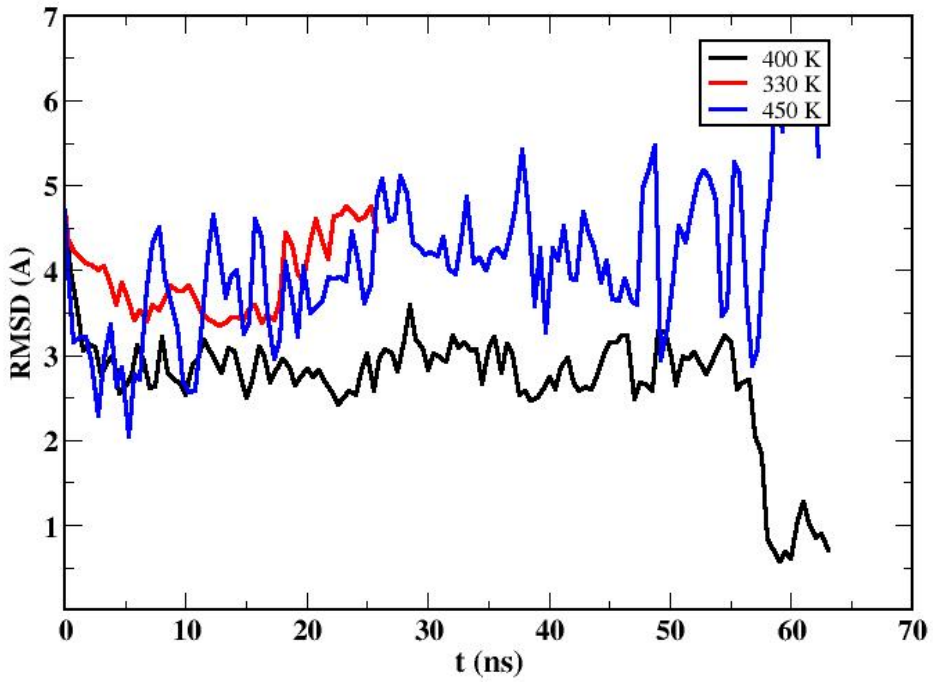

Figure 4.5: Effect of the temperature on Restraint metadynamics simulations in implicit solvent for GB3 protein. In black the time series of lowest RMSD reached by the simulation at 400 K, in red the one at 330 K and in blue the one at 450 K.

algorithm is due to the simultaneous action of the Camshift restraint on all the replicas: this determines a sort of "oriented" bias on all collective variables acting on the various walkers towards a general better agreement with experimental data, accelerating the formation of the local native rearrangements.

We investigated also the role of the temperature in the sampling, performing other two simulations with the same setup, but at 330 K and 450 K. As expected the amplitude of the RMSD fluctuations, and consequently of the structural changes, in the three runs is proportional to the temperature. In the simulation at 330 K (red line in Fig. 4.5) the relative low temperature affects the sampling slowing down the dynamics and the conformational exploration, even

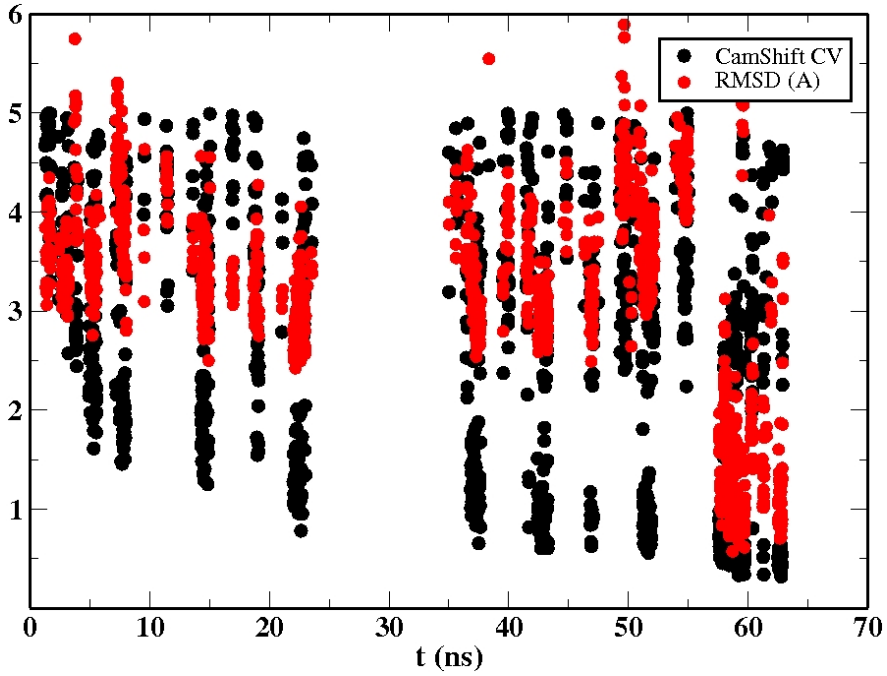

Figure 4.6: Values of the Camshift CV (black dots) and of the relative RMSD (red dots) from the reference structure for the configurations sampled in the replica bias by Camshift CV, during the Restrained metadynamics simulation at 400 K in implicit solvent.

in the first part of the run where all the performed simulations showed a net improvement in the structure similarity. The simulation at 450 K instead (blue line in Fig. 4.5) is able to visit immediately structures with a better agreement with the PDB reference, but in this time window the large fluctuations induced by the temperature does not allow the correct formation of the contacts to descend rapidly the funnel to the native fold.

The largest structural improvements usually happen in the first part of the simulation and the improvement usually corresponds to a new minima of the Camshift bias-potential, as shown in Fig 4.6. This suggests that selecting the structure with the lowest value of Camshift in the beginning of the simulation and starting a completely new simulation (without the previous deposited bias potentials) can fasten the procedure to reach the folded state. Analyzing the first 5 ns of the simulation at 450 K (blue line in Fig.4.7, we selected the structure with the minimum value of Camshift (which is at 3.1 Å of RMSD from the PDB

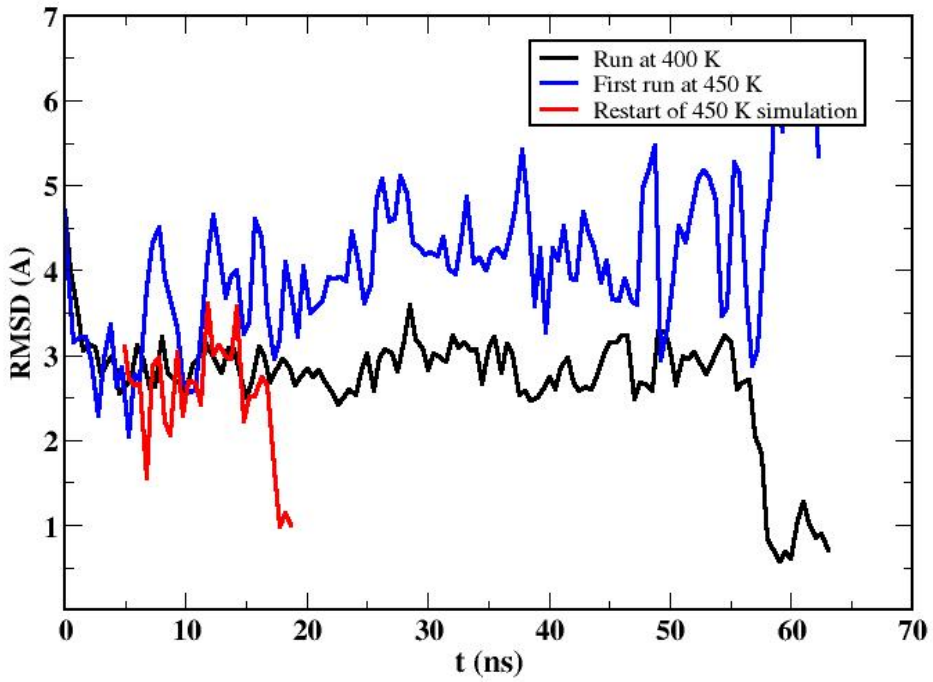

Figure 4.7: Effect of starting a new Restrained simulation from a structure at a low value of Camshift CV of a previous run. In black and in blue line the time series of the lowest from the previous simulation at respectively 400 and 450 K, and in red the same time series for a new simulation started from the configuration with the lowest Camshift value in the simulation at 450 K.

reference), and we started a new Restrained metadynamics simulation from this configuration. The red line in Fig.4.7 shows that in this way also the simulation at 450 K is able to obtain structures perfectly consistent with the experimental structure in less than 20 ns of simulation for 7 replicas in implicit solvent, and has been obtain with less than one day of computation on a total of 21 processors of a normal computer.

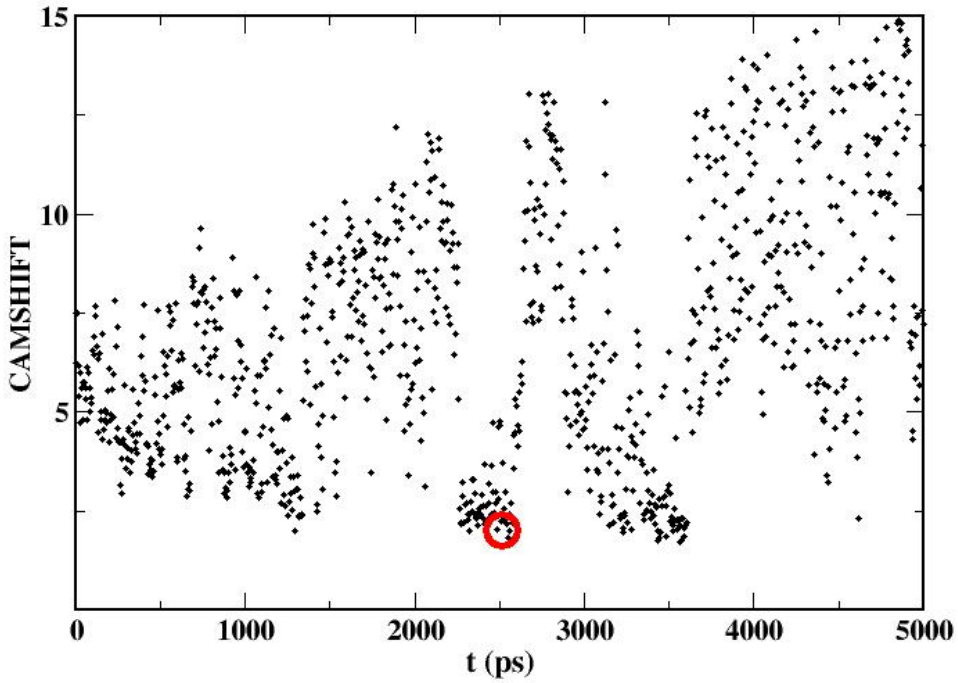

Figure 4.8: Time series of Camshift CV in the first Restrained metadynamics run for HR2876C CASD-NMR target. The red circle show the selection in the first Camshift minimum of the run of the structure, used then as starting configuration for the second run

4.3.2 A Blind Test: CASD-NMR Target

We tested our approach also in a blind test, participating at the structure determination of a CASD-NMR target (see Section. 1.3.3), HR2876C, an unsolved 97-residue long protein. After downloading the chemical shift assignment from [173], we generated the starting configuration for our simulation by homology modeling from the residues sequence, using the HHpred server [91] and Modeller [101, 102]. Then two Restrained metadynamics simulation were run for 5 ns each at 400 K in implicit solvent (see Section 3.3.3 for details). During the first run we selected a structure from the first minimum explored by the Camshift replica (red circle in Fig. 4.8), that has been used as starting configuration for the second

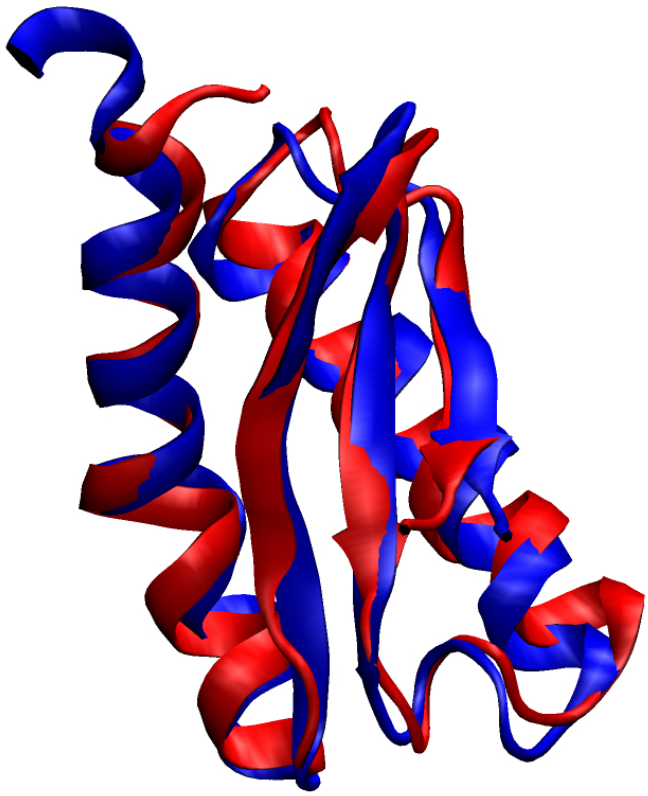

Figure 4.9: Comparison between the submitted structure (in blue) with the lowest backbone RMSD and the one of the model (in red) from the reference pdb (code 2M5O)

run. We finally submitted, as our solved structure, the centroids of the cluster analysis of the structures belonging to Camshift minima in the two runs (like the ones at 2.5 ns and 3.5 ns in Fig. 4.8): unfortunately the rescoring procedure through MD simulations proposed in Section 4.2 was not affordable due to lack of time and resources, so we ordered the structure based on the Camshift average on the elements of the clusters, from the lowest to the highest.

In [174] the external analysis of the RMSD in the range between ASP16 and ASP93 residue from the reference PDB (code 2M5O), released publicly after the submission, and the quality assessment for our models are reported. The starting

configuration obtained by the homology modeling was at 2.55 Å, the average backbone RMSD of our models is 1.31±0.23 Å (1.73±0.24 Å for all heavy atoms) with a best configuration at 0.84 Å (1.25 Å for all heavy atoms); the main defects are on the borders of the range selected for the RMSD calculation, as shown in Fig. 4.9. Moreover the similarity deteriorates with the model number, confirming that, also in this approximation, the Camshift averaging is a good indicator to rank the different models. Also the quality parameters (like distance, dihedral or Ramachandran violations) show good values, that could be further improved using the approach in explicit solvent simulation: indeed the Restrained NMR-metadynamics approach can be obviously applied also in explicit solvent, with likely better results in term of quality and structure similarity to the experimental coordinates due to a more accurate force field.

Chapter 5

Free Energy Landscape of a Globular Protein

As we have discussed in Section 1.1.1, the use of free energy landscapes rationalizes a wide range of aspects of protein behaviour by providing, in particular, a clear illustration of the different states accessible to these molecules, as well as of their populations and pathways of interconversion. However, the quantitative determination of the free energy landscapes of proteins by computational methods is very challenging as it requires an extensive sampling of their conformational spaces.

Together with the accurate structural characterization of the folded state that we discussed in the previous chapter (Section 4.1), the NMR-guided metadynamics that we introduce in this thesis allows also the description of the free energy landscape of a protein with relatively limited computational resources. As in this approach the chemical shifts are not used as structural restraints, the resulting free energy landscapes correspond to the force fields used in the simulations.

In this chapter we illustrate this approach for the third immunoglobulin-binding domain of protein G from streptococcal bacteria (GB3). Our calculations reveal the existence of a folding intermediate of GB3 with a partially non-native topology (Section 5.1.1). Furthermore, the availability of the free energy landscape enables the folding mechanism of GB3 to be elucidated by analyzing the conformational ensembles corresponding to the native, intermediate and unfolded

states, as well as the transition states between them (Section 5.1.2). Taken together these results show that by incorporating experimental data as collective variables in metadynamics simulations it is possible to enhance the sampling efficiency by two or more orders of magnitude with respect to standard molecular dynamics simulations[16], and thus to estimate free energy differences between the different states of a protein with a k_BT accuracy by generating trajectories of just a few microseconds.

5.1 Thermodynamics of GB3 Folding

Bias-Exchange Metadynamics allows the free energy of a system to be reconstructed once the bias potentials become stable [32]. In order to obtain such a characterization for GB3 protein, we have analyzed the simulation obtained by standard NMR-guided metadynamics approach and described in Section 4.1, following the scheme explained in Section 2.4.3). After selecting the CVs that are most effective to discriminate different states of the system, the CVs space is divided in hypercubes which represent microstates of the system. All the frames explored after the equilibration time t_{eq} are then assigned to the corresponding microstates according to their CV values: the relative conformations within each hypercube should be structurally consistent to define a proper microstate of the system, otherwise the hypercube size must be reduced, repeating the frames assignment. Then a free energy value is computed for the microstate as described by eq. 2.17, based on the values of the bias potentials and on the number of assigned frames in that specific microstates. In our study we have chosen the Camshift, Coordination Number, Anti- and Para-BetaRMSD CVs; the relative free energy profiles are reported in Fig. 3.3. The error on the free energy difference of the microstates corresponding to the three local free energy minima in Fig. 5.1 is approximately of 3 kJ/mol.

The molecular dynamics simulations reached convergence after approximately 240 ns, as at this point the bias potentials acting on all the replicas started to become stationary [144]. We then continued the simulations for further 140 ns in order to acquire enough statistics: so a total of approximately 2.7μs (380 ns on 7 replicas) has been spent to fully describe the free energy landscape and

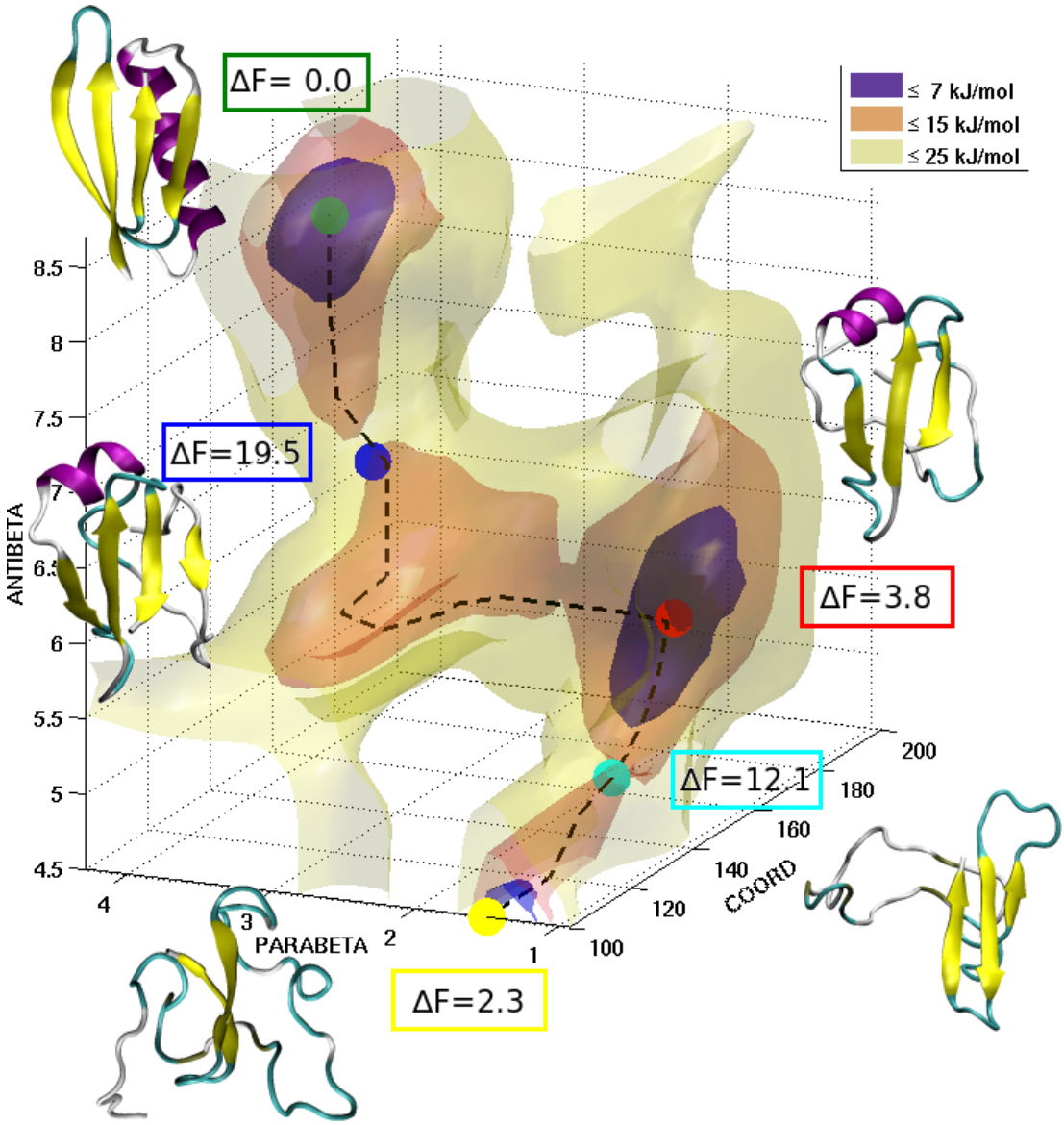

Figure 5.1: A) Three dimensional representation of the free energy landscape of GB3 as function of three collective variables (see main text). Along the folding pathway (black dashed line) the most relevant structures are reported with their relative free energy values: the native state N in green, the transition state TS2 in blue, the intermediate state I in red, the transition states TS1 in cyan and the unfolded ensemble U in yellow.

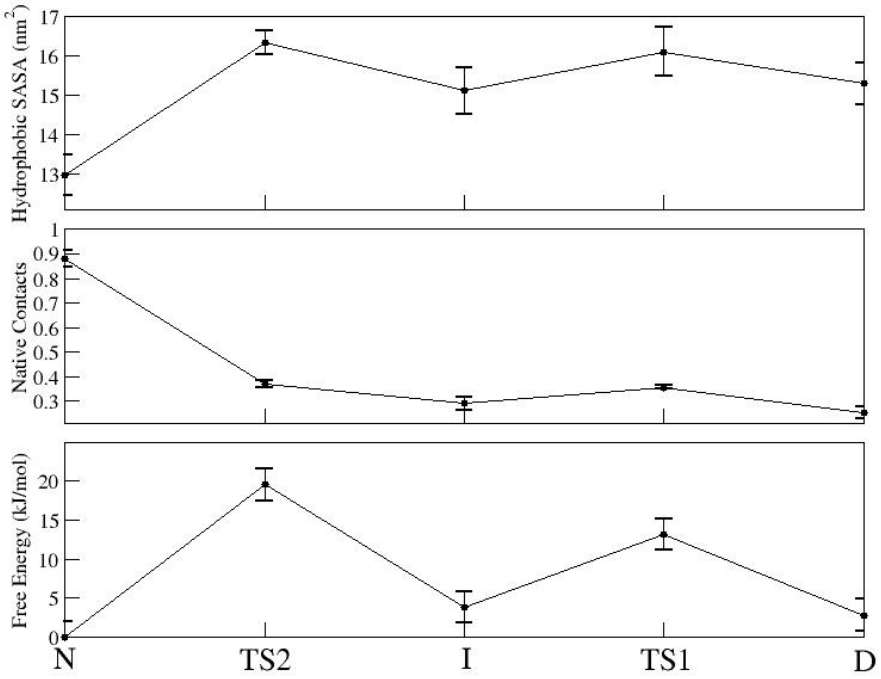

Figure 5.2: Hydrophobic solvent-accessible surface area (SASA), relative number of native contacts and free energy along the folding pathway.

the folding mechanism of a protein which folds on the millisecond time scale [169]. In Fig. 5.1 the free energy landscape is represented as a function of three collective variables, the fraction of antiparallel β-sheet, the fraction of parallel β-sheet, and the coordination number between the hydrophobic side chains (Fig. 5.1). This representation reveals the organization of the free energy landscape, with a deep minimum corresponding to native-like structures, separated by a relatively high barrier from other minima. The lowest free energy minimum includes configurations very similar to the reference structure (on average, at 1.3 Å RMSD). This result is confirmed by the analysis of the deviations of the calculated chemical shifts from the corresponding experimental values, both for the reference structure (PDB code 2OED) and for the structures belonging to the free energy minimum (Fig. 3.1). The agreement is excellent in both cases, thus confirming that by our procedure we were able to find structures that are very close to the X-ray structure. These results provide also evidence of the excellent quality of the Amber99SB-ILDN [127] force field that we used for modeling GB3.

The shallow minimum immediately after the free energy barrier that separates the folded state from the rest of the conformational space includes compact structures with a high secondary structure content, but with a fold that is rather different from the native, as will be discussed in detail below. This second minimum is separated by another free energy barrier from a further minimum, which includes more disordered structures with a much smaller secondary structure content. In these conformations, the native C-terminal β-hairpin appears to be present, confirming its high stability, while the helix and the N-terminal β-hairpin are completely disrupted [175, 176, 177]. The folded-like and unfolded-like states have a free energy difference of only 2.3 kJ/mol, which is comparable with the error of our free energy estimates. The similarity in the free energies of the folded and unfolded states reflects the conformational properties of the protein at the temperature at which the simulation has been performed (330 K), which is about 30 K below the experimental melting temperature of GB3 [36].

5.1.1 An Intermediate State in the Folding of GB3

The free energy landscape that we calculated illustrates explicitly the presence of three distinct states of GB3. In addition to the native (N, in dark green in Fig. 5.1) and unfolded (U, in yellow in Fig. 5.1) states, we identified the presence of an intermediate state (I, in red in Fig. 5.1) with a free energy 3.8 kJ/mol higher than the N state. From the relative free energies we calculated the populations of the three states at 330 K, which are 59% for N, 14% for I and 26% for U. An unbiased molecular dynamics simulation of 200 ns starting from a structure corresponding to the intermediate free energy minimum remains extremely stable, with an average RMSD of 2.4 Å from the equilibrated initial structure. These results are consistent with the observation of the presence of an intermediate state of GB1 [178, 179], which shares a 88% of sequence identity with GB3. In particular the latter work, which was based on the measurement of the kinetic folding constant as a function of the pH and denaturant concentration, reported a folding behaviour consistent with the presence of an on-pathway intermediate and two different transition states (TS). However, the structure of the intermediate of GB1 is more native-like than the one that we find here. The ensemble of

conformations making up the intermediate state characterized by our approach contains compact structures, which do share specific secondary elements with the native state, including the C-terminal β-hairpin. The N-terminal extension is instead less structured with only an incipient parallel pairing of the first β-strand [178] and the N-terminal region of the α-helix (residues 22-30). In addition, the C-terminal part of the α-helix exhibits a non-native configuration by forming an anti-parallel β-strand paired with the third β-strand of the protein (residues 41-47).

5.1.2 Identification of the Transition States

we tried to provide a qualitative characterization of the protein kinetics out of free energy landscape reconstructed by our procedure. This has been done by means of a kinetic Monte Carlo approach [180]. Based on our multidimensional reconstruction, at each step the system jumps from a microstate to another, based on the free energy difference with the nearest neighbors in the CV space. In this way it is possible to explore the connectivity between the different states on the free energy landscape, evaluating also the most probable pathways.

In order to better characterize the folding mechanism of GB3, we simulated by a kinetic Monte Carlo approach [180] the dynamics on the multi-dimensional free energy landscape reconstructed by our procedure. All the trajectories connecting the folded and the unfolded states go through the intermediate state, confirming it is an on-pathway intermediate, like the one observed in GB1 [179]. The black dashed line in Fig 5.1 represents the three-dimensional projection of the trajectory of highest probability connecting the folded and the unfolded state. Consistently with this topology, the trajectory crosses two transition states, TS1 between the unfolded state and the intermediate (in cyan in in Fig. 5.1), and TS2 between the intermediate and the native state (in blue in Fig. 5.1). The rate limiting step is represented by TS2, with a barrier of 19.5 kJ/mol from the native state, while TS1 is at a free energy of 12 kJ/mol.

The hydrophobic solvent-accessible surface area (SASA) reveals how the two TSs are less compact than the N and I states, but still quite structured (Fig. 5.2). A similar conclusion was reached by the experimental Tanford β-values

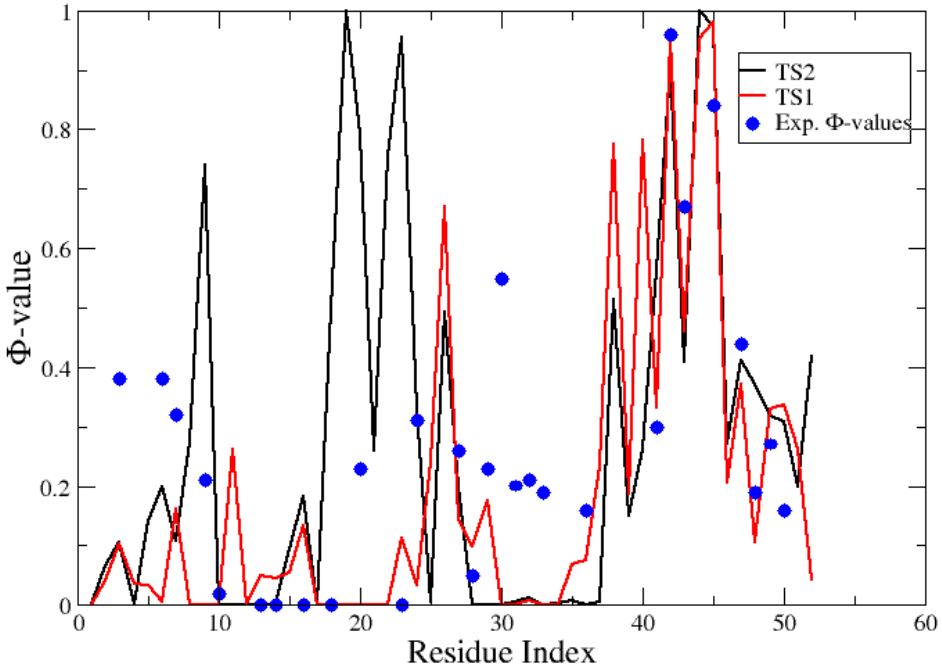

Figure 5.3: Comparison of the experimental ϕ-values (blue circles) of GB1 [169] and of the ϕ-values for GB3 calculated from the TS2 (black line) e TS1 (red line) structures determined in this work.

for the two transition states of GB1, $\beta_{TS1} = 0.76 \pm 0.04$ and $\beta_{TS2} = 0.93 \pm 0.04$ [179]. These values are consistent with those computed by the ratio of the total SASA between N and the corresponding TS (as proposed in [179]) obtained in the present study for GB3, $\beta_{TS1} = 0.82 \pm 0.03$ and $\beta_{TS2} = 0.91 \pm 0.03$.

We found that the TS2 of GB3 is more compact than TS1 (Fig. 5.1), at least in part because of the presence of a native salt bridge between Lys-10 and Glu-56 that is missing in TS1. This aspect was also suggested in the case of GB1 [179] to explain the differences in the pH-dependence for the unfolding rate constant of the two transition states. Indeed a visual inspection of the TS1, I and TS2 structures reveals how this salt bridge can trigger the correct arrangement between the C-terminus and the first β-strand (residues 1-10). The formation of

the salt bridge, which is absent in TS1, in I acts as an anchor that allows the parallel pairing of the first β-strand, increasing the fraction of native contacts from 29% in I to 37% in TS2. On this view, the second β-hairpin represents the initial native element in the folding process, followed by the formation of the N-terminus of the native helix and the parallel pairing of the first β-strand to the C-terminus β-hairpin, in order to stabilize then the formation of the first-hairpin.

These findings are consistent with the ϕ-values measured for GB1 [169]. A comparison between the experimental ϕ-values of GB1 and those calculated for GB3 for TS1 and TS2 is presented in Fig. 5.3 through the fraction of native contacts of amino acid side-chains [22, 23]. Despite the differences in sequence between GB1 and GB3, the structure of the TS2 of GB3 exhibits a pattern approximately consistent with experimental ϕ-values of the transition state of GB1 (Fig. 5.3), especially in the two β-hairpin regions. These results, which are consistent with previous conclusions [169], indicate that in the transition state the C-terminal harpin is completely formed as well as the parallel pairing of the first β-strand. Instead the ϕ-values in the α-helical region shows a more complex behaviour compatible with a variety of conformations in the transition ensemble, and so less native-like.

Chapter 6

Free Energy Landscape of an Intrinsically Disordered Protein

The results shown in Chapter 5 motivated the attempt to use NMR-guided metadynamics to explore the conformational space of an intrinsically disordered protein. As we mentioned in Section 1.1.2 intrinsically disordered proteins are characterized by the absence of a well- defined three dimensional structure under native conditions [33, 34, 35]. Despite the great challenge to probe experimentally their structural features, by bringing together experiments and computations it has been recently possible to make progress in determining the conformational space available to them [181, 182, 183, 184, 185, 186, 187]. However, characterizing the free energy landscape of these proteins is still an open problem.

Exploiting our methodology, we are able to provide a structural and energetic characterization for the 40-residue form of Amyloid beta (we will refer to it as Aβ40), involved in Alzheimer's disease because prone to aggregate in oligomers and fibrils, which can be toxic for the cells. We found that the free energy landscape of this peptide is inverted with respect to that of folded proteins, as it is characterized by a global minimum consisting of highly disordered structures and by a series of local minima corresponding to partially folded conformations. These structures are kinetically committed to the disordered state, but their free energies are quite low, indicating that they are transiently explored even at room temperature and may be involved in the aggregation process. The presence of

an inverted free energy landscape suggests that intrinsically disordered proteins should become more compact when the temperature is increased.

6.1 Thermodynamics of an Intrinsically Disordered Protein: Abeta40

We apply the NMR-guided metadynamics strategy introduced in this thesis (Section 3.1) to characterize the structural ensembles and the free energy landscape of Abeta40 (see Section 3.3.2 for simulation detail). Since in this approach the chemical shifts are not used as structural restraints, the resulting free energy landscape represents the Boltzmann distribution of the force field used in the simulations. We performed our simulations in explicit solvent at 350 K using the Charmm22* force field [11]. The simulation has been run for 310 ns on 8 replicas ($\sim 2.5\mu$s in total) each biased by a different history-dependent potentials acting on a specific collective variable: two replicas use the CamShift CV, which measures the deviations between experimental (bmrb file 17796 [168]) and calculated chemical shifts; three for the exploration of the secondary structure elements (α-helix, parallel and anti-parallel β-sheet) ; the number of hydrophobic contacts, which counts the number of contacts between the heavy atoms of the hydrophobic residues and measures the degree of compactness of the protein; two AlphaBeta similarities which evaluates the deviation from the average χ_1 and χ_2 torsion angles values for the hydrophobic and polar residues side chains.

The simulation showed a high level of convergence in the free energy, verified by calculating the differences in free energy of the two halves of the trajectories for each replica after the equilibration phase, around 150 ns. In Fig. 6.1 we report the free energy profiles for all the 8 CV variables used in the sampling. From these calculations we estimate that up to 30 kJ/mol the free energies that we obtained are accurate within 2-3 kJ/mol.

6.1.1 Free Energy and Structural Characterization

Also in this case the procedure is able to provide a comprehensive representation of the conformational space of Aβ40, producing a wide range of structures, from

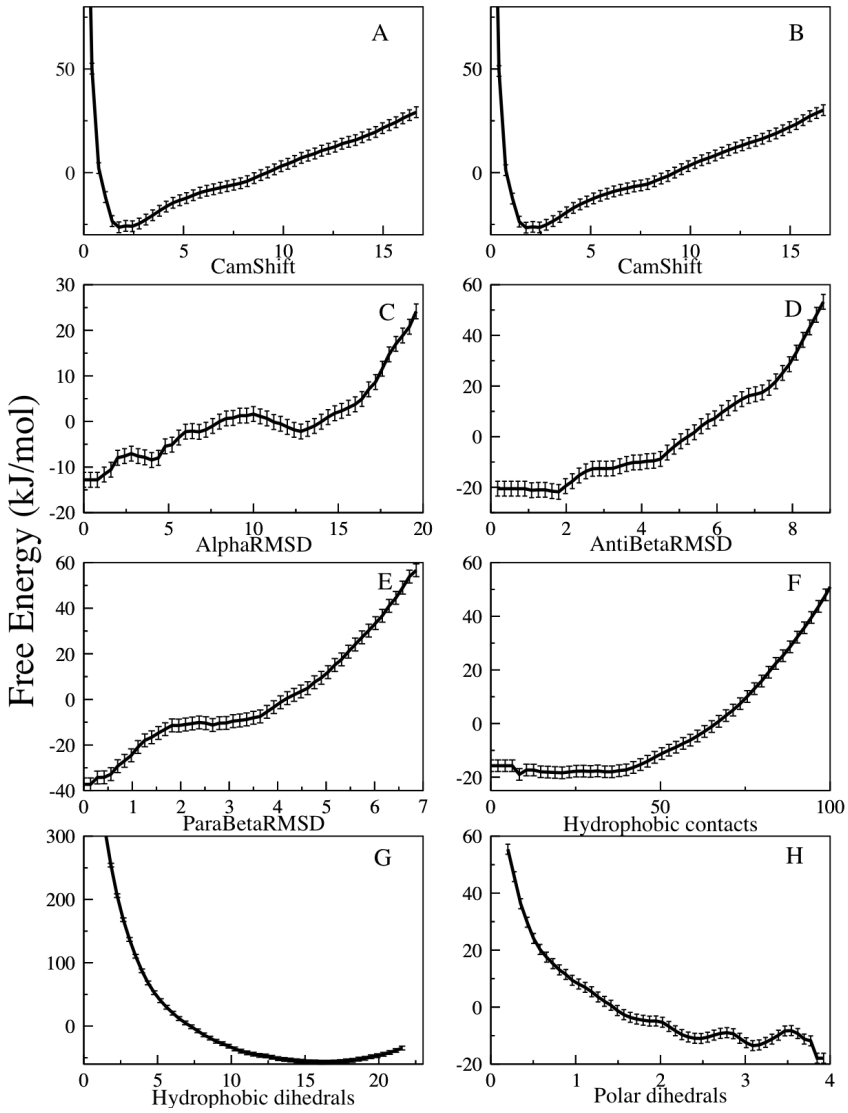

Figure 6.1: Free energy landscapes for the eight CVs used in the simulations: (A,B) Two replicas use the CamShift CV, which measures the deviation between experimental and calculated chemical shifts; (C) AlphaRMSD; (D) AntiBetaRMSD; (E) and ParaBetaRMSD (E) for the corresponding secondary structure content; (F) Hydrophobic contacts, which counts the number of contacts between the heavy atoms of the hydrophobic residues; (G,H), deviation from the average χ_1 and χ_2 torsion angles values for the Hydrophobic and Polar residues side chains. All the free energy landscapes are shown with the relative error bars, which are in all cases below 2.9 kJ/mol.

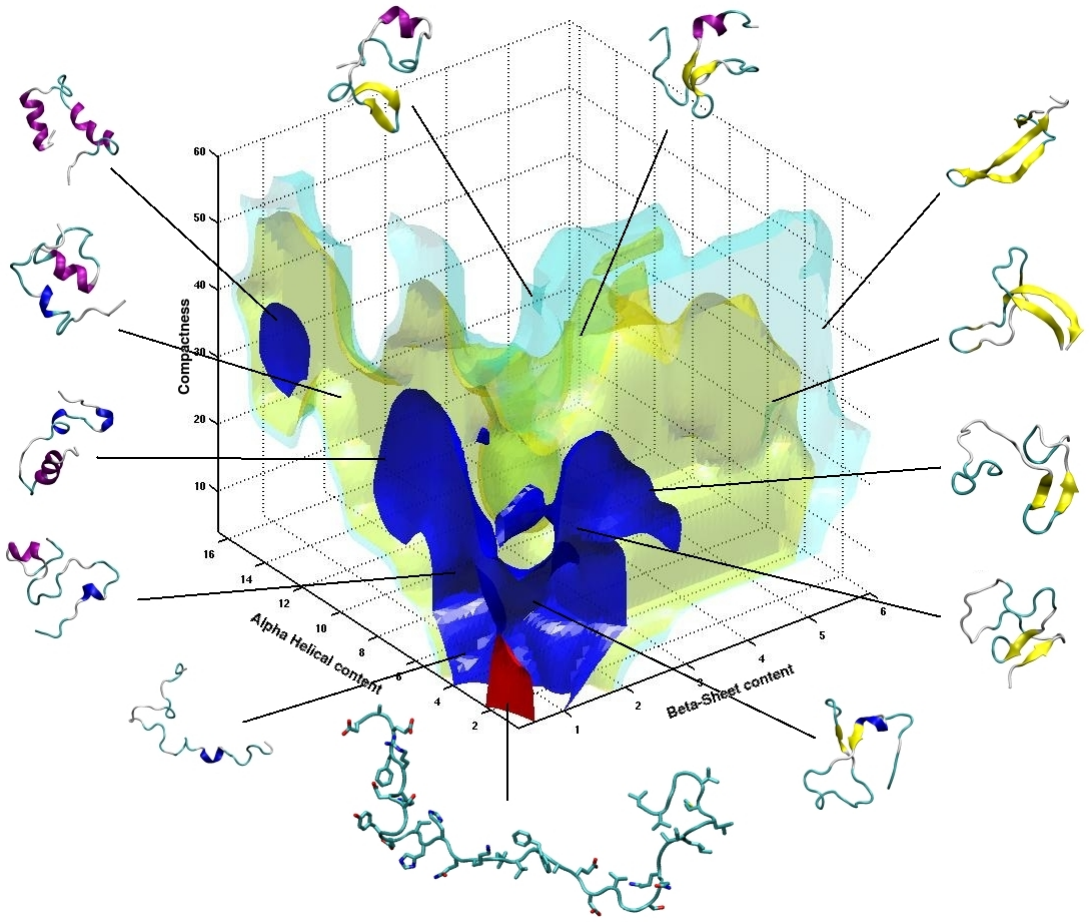

Figure 6.2: Inverted free energy landscape of the Aβ40 peptide. The free energy landscape is shown as a function of three of the eight collective variables used in the NMR-guided metadynamics simulations: anti-parallel *beta*-sheet content (X-axis), α-helical content (Y-axis) and degree of compactness (Z-axis). Isosurfaces are shown at 5 (red), 10 (blue), 18 (yellow) and 25 kJ/mol (cyan); white regions are not visited as they have higher free energies. Representative structures sampled during the simulation are also shown.

completely open to compact ones, as it is pointed out in Fig. 6.2. In this figure we represent the free energy of the system projected in the space defined by three variables, the anti-parallel β-sheet content, the α-helical content and the number of hydrophobic contacts. As expected from the intrinsically disordered nature of the Aβ40 peptide and from previous studies, the global free energy minimum in this landscape corresponds to an ensemble of highly disordered structures. At higher

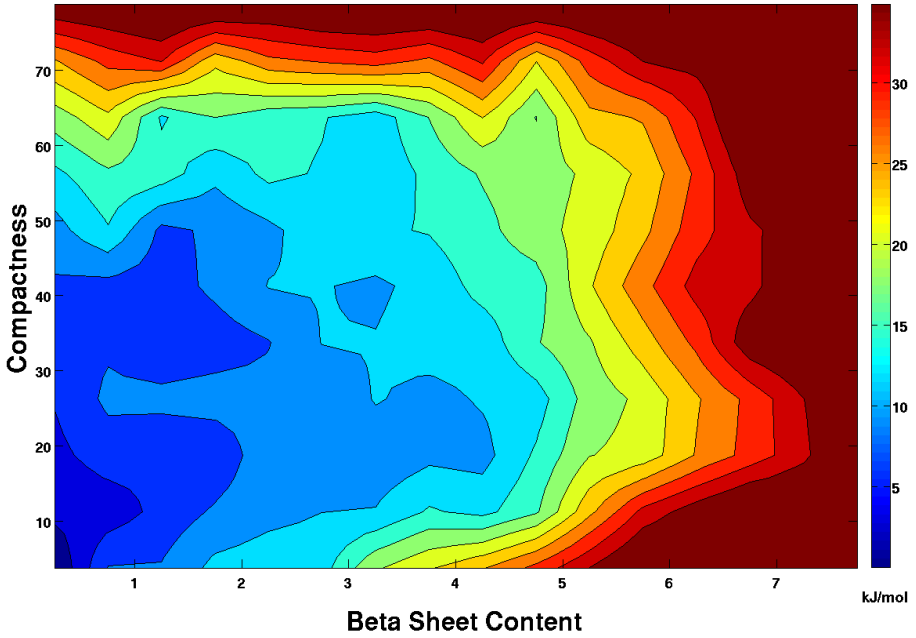

Figure 6.3: Two dimensional projection of the free energy of Abeta40 as a function of antiparallel beta-sheets content (X) and degree of compactness (Y) of the structure. The isoline are drawn every kT.

free energies, the landscape includes a large amount of partially structured conformations. These structures are kinetically committed to the disordered state, but their free energy is in many cases only 5-10 kJ/mol higher, indicating that they are transiently explored even at the simulation temperature.

In Fig. 6.3 we plot the free energy as a function of the β-sheet content and compactness. We observe just a funnel towards the global disordered minimum, without any other local minima. In fact up to 10 kJ/mol only short β structures are observed, with two or three β-bridges formed. Extended β-strand have a free-energy higher than 30 kJ/mol.

The projection on the alpha-helical content in Fig. 6.4 shows instead a different picture. Beside the funnelled shape with only short helical structures, we observe a local minimum with a higher helical content, with a free energy around 8-10 kJ/mol. This region is populated by several structures similar to the pdb 2LFM [188] with an RMSD of 1.3 A for the central 10-residue long helix (Fig.

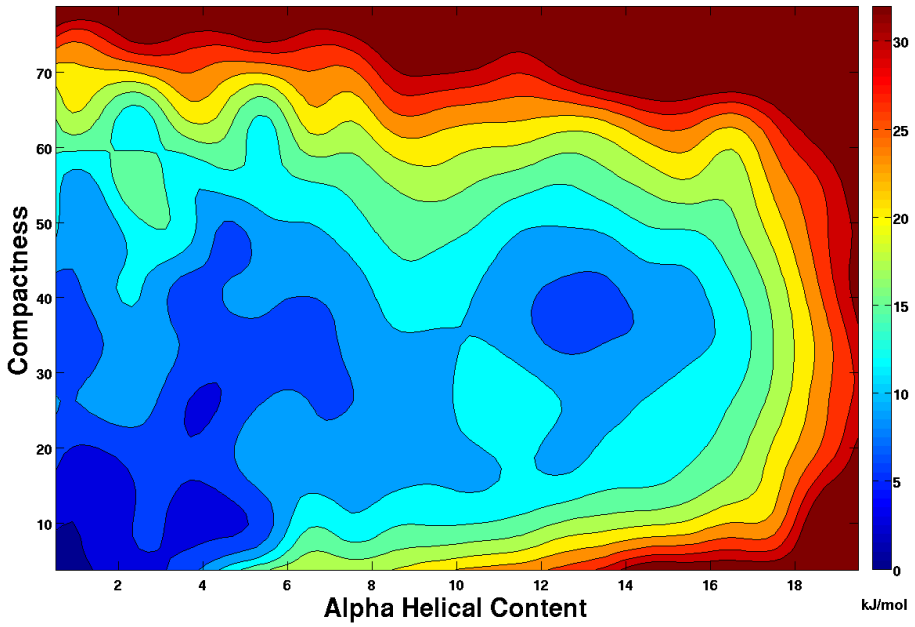

Figure 6.4: Two dimensional projection of the free energy landscape of Abeta40 as a function of the degree of compactness(X) and the helical content (Y) of the structure. The isoline are drawn every kT.

6.5).

In order to benchmark the goodness of our structural exploration and the consistency with experimental structure, we compute the chemical shifts of our ensemble by the averaging procedure described in Section 2.4.3 (eq. 2.19): the shifts for all the backbone atoms are computed with a simple average over all the structure of each microstate and then weighted according the free energy of the microstate. Chemical shifts were calculated by a different chemical shift predictor, SPARTA+ [155], to avoid possible systematic errors from Camshift predictor. As it reported in the first column of Table 6.1 and in Fig 6.6, the agreement between the back calculated and the experimental chemical shifts [168] is remarkable. Moreover the chemical shifts of our ensemble are consistent with a random coil conformation: in fact comparing the theoretical values corresponding to a random coil configuration with the experimental ones (second column of Table 6.1) we obtain the same agreement. These results shows that the force-field is overall not

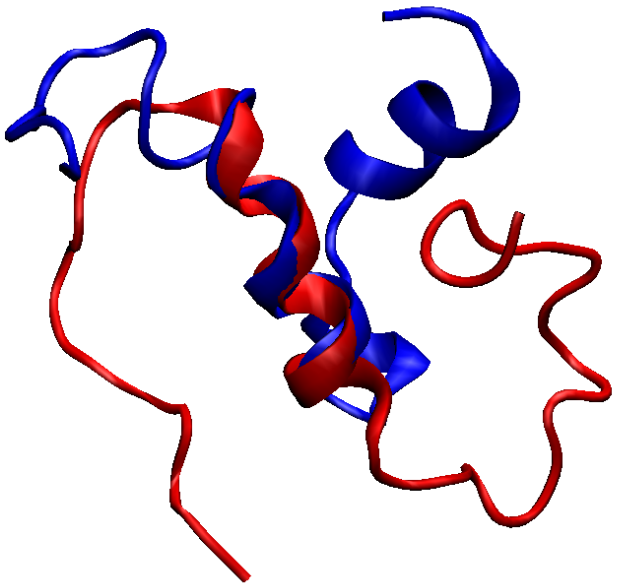

Figure 6.5: Alignment of a sampled structure (in blue) in the local minimum with an NMR solved structure (pdb code 2LFM [2], in red), with a RMSD of 1.3 Å for the central 10-residue long helix.

over-structuring the disordered state of Aβ40.

In Fig. 6.7 we report a more quantitative analysis of the secondary structure content on the reconstructed landscape. The secondary conformation of each residue is again computed first averaging over all the structure of each microstate and then weighting the obtained average with the corresponding free energy. At low free energy (panel b) and c)) helical configurations are preferred with respect to β-bridges; the picture changes at higher free energy where the β-sheets population increases notably. These results show that few kcal/mol above the completely opened and disordered global minimum, the system populates more easily helical rather than β-sheet conformation. Fibril-like structure with extended strand appears only at higher free-energy.

Atom	Ensemble CS	Coil CS	Pred. error
HA	0.10	0.08	0.25
HN	0.30	0.28	0.46
N	1.25	1.45	2.36
CA	0.35	0.36	0.88
CB	0.45	0.28	0.97

Table 6.1: Average deviation of the ensemble averaged chemical shifts calculated by SPARTA+ [155] from the experimental chemical shifts [168]. In the second column the same comparison is performed based only on the sequence, considering the residue to which the atom belong in a random coil conformation. The expected error of the predictor is also reported. All the numerical values are expressed in ppm

6.1.2 An Inverted Free Energy Landscape

The most striking feature that can be extracted from Fig. 6.2 and 6.7, is that the overall shape of the free energy landscape of the Aβ40 peptide seems inverted with respect to that of globular folded proteins. The latter ones usually show a global minimum with a well-defined secondary and tertiary arrangements, surrounded by slightly defective states which shares however the same topology of the folded one; as the free energy increases, the native structure is gradually lost in the metastable states, determining progressively misfolded and completely unfolded configurations, as depicted in chapter 5 by the free energy landscape of GB3.

Conversely, in our reconstruction of the Aβ40 free energy landscape the global minimum consists of an ensemble of highly unstructured conformations, rather than of a well-defined structure. At higher free energies microstates with relatively compact and structured configurations appear. As it is shown in Fig. 6.7, the amount of structure gradually increases with the free energy, in a manner that is opposite to what happens in folded proteins. The secondary structure populations indeed grow progressively from the lower free energy region (Fig. 6.7b) to the higher ones of the free energy landscape (Fig. 6.7c-e), until 18-24 kJ/mol.

These results suggest a remarkable phenomenon: at increasing temperatures the Aβ40 peptide becomes more structured and hence more compact, as it becomes possible for it to explore the regions of higher free energy. At even higher

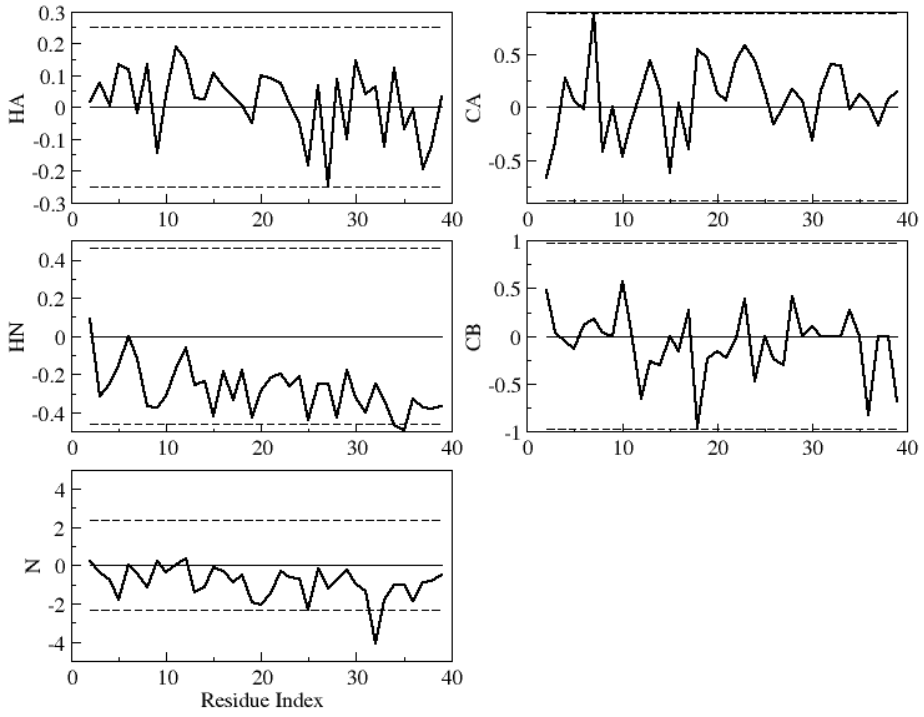

Figure 6.6: Difference between the experimental chemical shifts and the ones obtained by the ensemble average of our sampled structure. The dashed line indicate the expected error of the predictor, SPARTA+ [155]. The chemical shift value for C' atoms were missing in the experimental data (bmrb file 17796). All the values on the y-axis are expressed in ppm.

free energies, above 24kJ/mol, the secondary populations begin to decrease (Fig. 6.7f). This behaviour is reminiscent of the cold denaturation of structured proteins, which in the case of intrinsically disordered proteins we predict to take place at high temperatures. In order to prove this scenario, we computed the average number of hydrophobic contacts and of the gyration radius, based on the reconstructed landscape, as a function of the temperature with a quadratic approximation around the simulation temperature $T_0 = 350$ K. For the number of hydrophobic contacts n we have

$$n(T) = n|_{T_0} + \left.\frac{dn}{dT}\right|_{T_0} (T - T_0) + \frac{1}{2} \left.\frac{d^2n}{dT^2}\right|_{T_0} (T - T_0)^2 \tag{6.1}$$

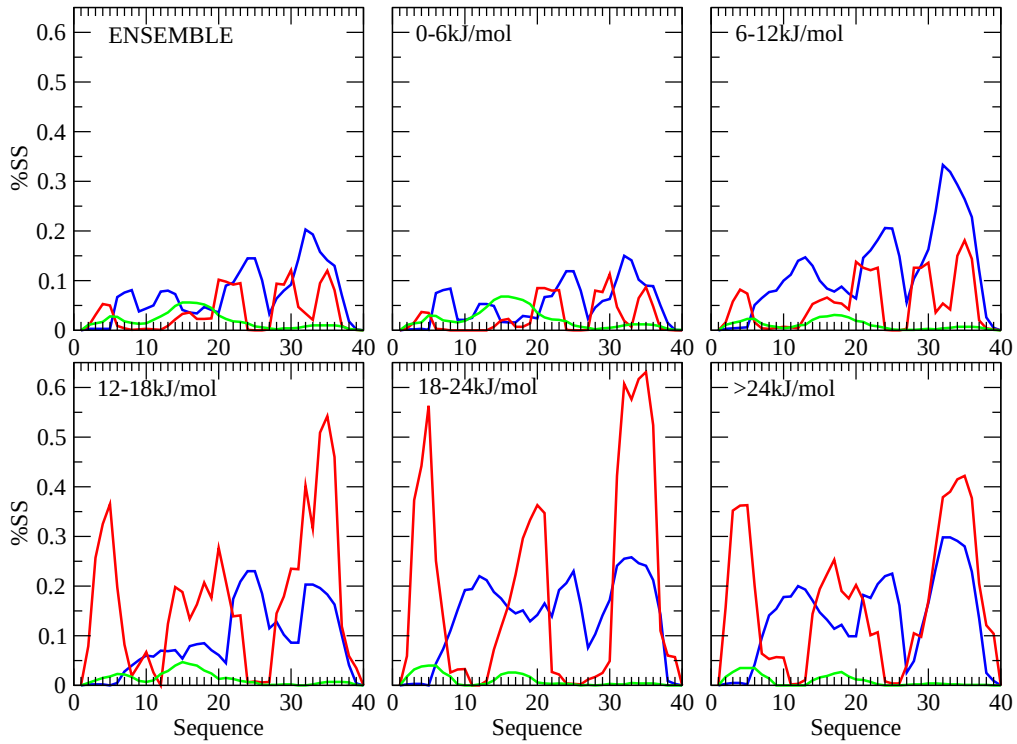

Figure 6.7: Secondary structure populations of the Aβ40 peptide at increasing values of the free energy. The lines indicate different secondary structure types: sheets are shown in red, α-helices in blue and polyproline II in green. The different panels report the secondary structure populations corresponding to different slices of the free energy landscape: (a) entire free energy landscape, (b) lower region of the free energy landscape (0-6 kJ/mol), (c) 6-12 kJ/mol region, (d) 12-18 kJ/mol region, (e) 18-24 kJ/mol region and (f) higher region (above 24 kJ/mol).

where

$$n = \langle n \rangle = \frac{\sum_\alpha n_\alpha e^{-F_\alpha/T}}{\sum_\alpha e^{-F_\alpha/T}} \tag{6.2}$$

and using eq. 2.19 it is possible to show that:

$$\frac{dn}{dT} = \frac{1}{T^2} \left(\langle nV \rangle - \langle n \rangle \langle V \rangle \right) \tag{6.3}$$

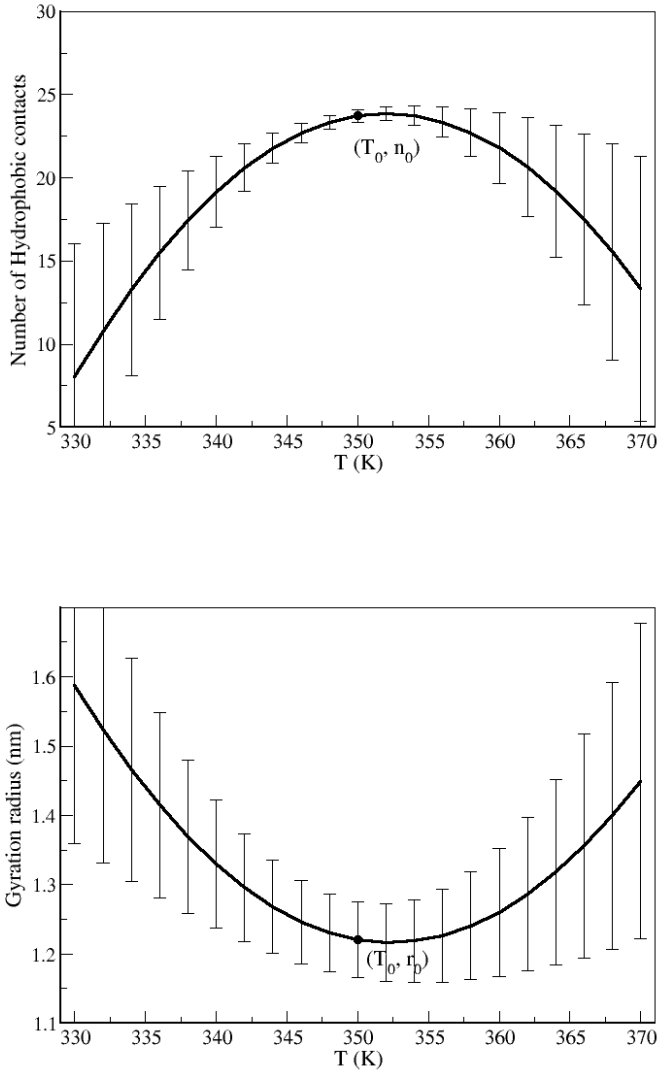

Figure 6.8: The two panels show respectively the behavior of the number of hydrophobic contacts (above) and of the radius of gyration (below) as a function of the temperature, with a parabolic approximation around the temperature of the simulation (350 K). This allows capturing the variation of the compactness of the system. The error is estimated comparing the two parabolas obtained computing the three relative coefficients on the first and the second halves of the trajectory after the equilibration time (150 ns)

$$\frac{d^2n}{dT^2} = \frac{1}{T^4}\left(\langle nV^2\rangle - \langle n\rangle\langle V^2\rangle\right) - \frac{2}{T^2}(T + \langle V\rangle)\frac{dn}{dT} \tag{6.4}$$

where E is the potential energy. The previous equations are valid for any other observable, including the gyration radius. The errors are estimated from the difference between the two parabolae evaluating the corresponding three coefficients on the two halves of the trajectory after the equilibration time.

Fig. 6.8 confirms our hypothesis: as the temperature decreases towards the room temperature, the size of the hydrophobic core decreases (and correspondingly the gyration radius increases).

Based on the results of our simulations, at room temperature the protein is predicted to visit less structured and more open conformations; at higher temperature, the protein access configurations which are more compact and with an higher content in secondary structure. Increasing again the temperature, the open conformations returns to be favored by the entropic contribution.

Chapter 7

Conclusions and Perspectives

Combining the information contained only in the backbone chemical shifts with an advanced conformational search technique we have been able to provide the structural characterization for different proteins, without any other knowledge on the target structure. Exploiting the ability of our approach to estimate the free energy of the visited states, we have been also able to reconstruct with high accuracy the folding landscape for a well-structured globular and for an intrinsically disordered protein.

The introduction of the new collective variable, Camshift CV, which measures the difference between experimental and calculated chemical shifts, helped the protein to find the route to its native state. The bias introduced by this collective variable allows the system to form rapidly the correct local topology encoded in the chemical shifts. The consequent formation of critical native contacts coupled with the fast exploration of the secondary structure and of the tertiary assembly induced by the other collective variables, allows mimicking the fundamental mechanisms of the folding theory, described in Fig. 1.3, and explains the success of our methodology.

Thanks to the molecular dynamics framework, this kind of approach can be generalized by incorporating other experimental data in a metadynamics framework, including NOEs, J-couplings, and residual dipolar couplings, or indeed data from other experimental techniques, such as SAXS or FRET methods. We anticipate that these developments will enable obtaining molecular dynamics descrip-

tions of the behaviour of a variety of proteins for which only sparse experimental data are available.

NMR Structure Determination

As we have shown in Chapter 4, the NMR-guided metadynamics is able to construct highly accurate structural models of proteins, based only on the chemical shifts assignment. Both for GB3 protein and for the blind test from the CASD-NMR experiment (in Section 4.1 and 4.3.2), we produced structures with a root-mean-square deviation lower than 1 Å from the experimental solved ones .

The use of Camshift as a scoring function allows to rank and distinguish even small structural differences (Section 4.2), assessing as native the state with the lowest mean value in the CamShift in a short plain molecular dynamics simulation. This result confirms the methodological importance of evaluating a structural-based score not only on a single structure but on an ensemble, as we also proposed in [96].

In perspective, the computational cost reduction introduced by exploiting implicit solvent simulations and the Restrained Metadynamics approach (Section 4.3) can lead to a new automatic and fast method for structure determination from NMR spectroscopy. As schematically shown in Fig. A 7.1, after the building of the starting configuration by comparative modeling (either homology- or threading-based), a short NMR Restrained metadynamics simulation, based only on the known chemical shifts, can be run for 5 ns. A configuration found in the lowest Camshift minimum can then be used to start a new run, reiterating the procedure until several sampled minima converge towards the same topology. After a clustering in RMSD, the best candidate can be selected ranking the different possible models based on the corresponding mean CamShift value on a 1 ns plain MD run.

Since only the chemical shifts assignment stage is required, our methods could simplify the standard NMR structure determination protocols avoiding lots of resource- and time-consuming experiments such as NOESY, correlation spectroscopy, which becomes more and more complicated as the protein size increases, due to the superimposition and broadening of the resonance peaks, allowing the

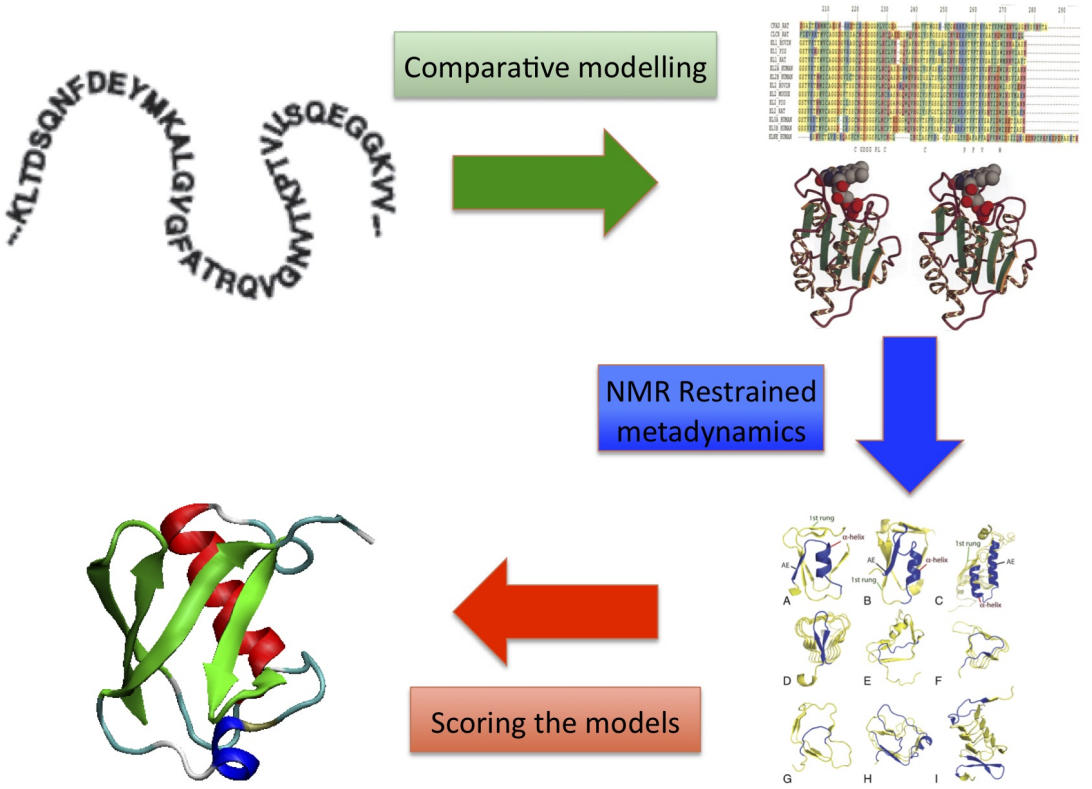

Figure 7.1: Structure determination protocol by NMR Restrained metadynamics. Starting only from the prior knowledge of the amino acids sequence and of the chemical shifts assignment, an initial configuration for the simulation is built by comparative modeling. Several stages of NMR Restrained metadynamics are then run improving progressively the structure until the Camshift minima converge towards similar structures. The differerent models can be finally scored to find the one which overlap better the chemical shifts.

investigation of larger systems.

Free Energy Landscape of Proteins

The possibility to characterize the free energy landscape of proteins by explicit solvent molecular dinamics simulations is the most important achievement of this thesis. The accurate description of the folding landscape of GB3 protein shown in Fig. 5.1, is obtained by limited computational resources (with a total simulation time of 2.7 μs) for a protein which folds on the millisecond timescale. It exhibits a

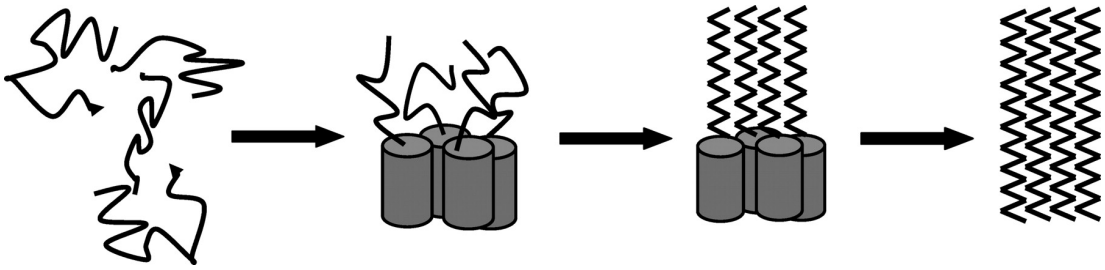

Figure 7.2: Schematic diagram of how α-helical intermediates might promote amyloid formation. Since in our landscape reconstruction the helical conformations (schematized as cylinders) are more abundant than β-strands ones (zigzagged lines). Initial oligomerization can be driven by the interaction between self-association and helix formation. This generates the high local concentration of regions of the peptide chain which has a high propensity to adopt β-structure. The consequent propagation of β-structure leads to the formation of β-sheet-rich assemblies. The figure is adapted from [189]

global free energy minimum consisting of structures effectively indistinguishable from the experimental reference, providing evidence of the excellent quality of current force fields. Our calculations revealed also the presence of an on-pathway partially non-native intermediate state and allowed the characterization of the two transition states, all consistent with recent experimental results about the system [179].

In the application to well-structured protein, our method can investigate protein states otherwise invisible both to simulations and experiments, including intermediates, denaturated and transition state ensembles, describing also the relative interconversion pathways. The approach can be used in particular when the chemical shifts of a state different from the native one are know, or more extensively to characterize the free energy landscape of already solved structures, calculating "synthetic" chemical shifts on the native state by predictors such as Sparta+ [155], ShiftX [153] or CamShift [152] itself, and then using them in a NMR-guided metadynamics.

The ability to reproduce landscapes highly consistent with the underlying experimental data, appears to be particularly crucial to describe quantitatively the structural ensembles of intrinsically disordered proteins. The calculations performed on the monomeric form of Aβ40 peptide, depicted a free energy landscape

with a global minimum composed of highly disordered and open conformations. At higher free energy we observed several partially folded conformations with a relatively high secondary structure content and a free energy only 5-10 kJ/mol higher than the disordered minimum. While conformations rich in β-strands arrangement compatible with oligomers and fibrils formation possess a quite high free energy, the great abundance of helical states described by our reconstruction suggests a possible explanation for the aggregation process. Following the hypothesis proposed in [189] and shown in Fig. 7.2, the transient formation of helical intermediates can favor the interaction and the initial oligomerization of several Aβ peptides. This high localization of specific region of the protein chain can indeed be the crucial seed for the further β-sheet assembly. Moreover, since with our approach we assessed the relative free energy of partially structured metastable states, we are able to define a library of structure that could be used to obtain a list of targets for docking studies to design inhibitors to bind the Aβ40 peptide in partially structured conformations in order to prevent the aggregation process.

The characterization of the free energy landscape of this disordered protein gave us the opportunity to hypothesize a remarkable phenomenon about the folding theory of this kind of proteins. The temperature-induced collapse is just a direct consequence of the picture described in Fig. 6.2: the raise in temperature, together with the absence of any significant energy barrier, provide an easy access to the plateau-like scenario of ordered states just few kcals/mol above the global disordered minimum. This temperature effect is indeed a strong signature of the inversion of the free energy landscape for intrinsically disordered proteins, that we hypothesize: their native states, in thermodynamic sense, tend to be stabilized by entropy, while those of ordered ones by enthalpy. Consequently while the high-energy states of ordered proteins are partially disordered, those of disordered proteins are partially ordered.

Completing the Protein Paradigm

The inversion of the landscape is therefore based on the reversal of the role of entropy and enthalpy between ordered and disordered proteins. Revising the rough

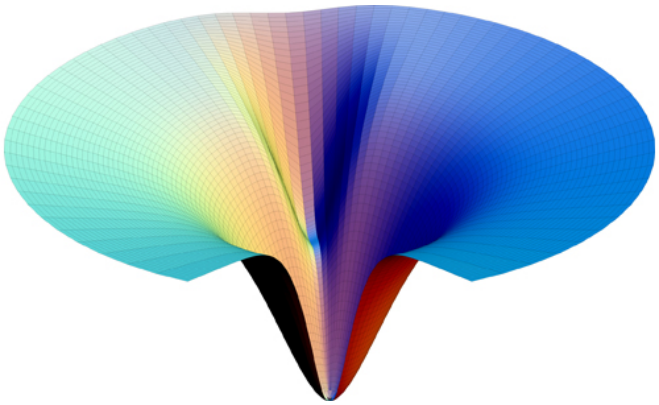

Figure 7.3: Representation of the free energy landscape of an intrinsically disordered protein according our hypothesis, characterized by an entropic dominated minimum at the bottom of wide and barrierless funnel composed at the top of an almost flat region populated by more structured metastable states.

comparison between the free energy of well- and un-structured proteins shown in Fig. 1.7, the landscape of an intrinsically disordered protein resembles more to a very wide and smooth funnel, as the one schematized in Fig. 7.3. Although the energetics of the disordered proteins is usually depicted as highly rugged and locally frustrated like in Fig. 1.7b, the case is actually the opposite: the entropic predominance removes any energy barriers determining a low global frustration in the system which is therefore able to change conformation rapidly. This provides to disordered proteins the properties to perform their peculiar functions of signaling and regulation inside the cell, which would be hard to address with the high stability of a well structured protein.

By taking a general view about the protein folding problem, nature has been able to finely tune the protein properties though the selection of the amino acid sequence. Changing the relative composition in charged and hydrophobic residues, it is able to change as needed the balance between entropy and enthalphy, shaping the final free energy scenario, ranging in a continuous manner from enthalphy-dominated, as in the case of enzymes, to entropic-dominated landscape, as in the case of intrinsically disordered proteins, in order to determine the immense repertoire of biological functions needed for life.

Bibliography

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[2] A. R. Fersht. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. *WH Freeman*, 1998.

[3] V. Munoz and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proceedings of the National Academy of Sciences*, 96(20):11311–11316, 1999.

[4] C. M. Dobson, A. Šali, and M. Karplus. Protein folding: a perspective from theory and experiment. *Angewandte Chemie International Edition*, 37(7):868–893, 1998.

[5] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[6] K. A. Dill, H. S. Chan, *et al.* From Levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, 1997.

[7] F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus, P. Crumley, A. Curioni, M. Denneau, W. Donath, M. Eleftheriou, B. Flitch, B. Fleischer, C. J. Georgiou, R. Germain, M. Giampapa, D. Gresh, M. Gupta, R. Haring, H. Ho, P. Hochschild, S. Hummel, T. Jonas, D. Lieber, G. Martyna, K. Maturu, J. Moreira, D. Newns, M. Newton, R. Philhower, T. Picunko, J. Pitera, M. Pitman, R. Rand, A. Royyuru, V. Salapura, A. Sanomiya, R. Shah, Y. Sham, S. Singh, M. Snir, F. Suits, R. Swetz, W. C. Swope,

N. Vishnumurthy, T. J. C. Ward, H. Warren, and R. Zhou. Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, 40(2):310–327, 2001.

[8] G. Kikugawa, R. Apostolov, N. Kamiya, M. Taiji, R. Himeno, H. Nakamura, and Y. Yonezawa. Application of MDGRAPE-3, a special purpose board for molecular dynamics simulations, to periodic biomolecular systems. *Journal of Computational Chemistry*, 30(1):110–118, 2009.

[9] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. Richard Ho, D. J. Ierardi, I. Kolossvary, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[10] M. Shirts and V. S. Pande. Screen savers of the world unite! *Science*, 290(5498):1903–1904, 2000.

[11] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *et al.* How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):47–49, 2011.

[12] A. D. MacKerell, N. Banavali, and N. Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2001.

[13] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, 2002.

[14] V. S. Pande, I. Baker, J. Chapman, S. Elmer, S. Khaliq, S. Larson, Y. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, *et al.* Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109, 2002.

[15] R. B. Best. Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology*, 22(1):52–61, 2012.

[16] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[17] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[18] A. R. Fersht and V. Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573–582, 2002.

[19] J. I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26):10340–10345, 2012.

[20] A. H. Ratje, J. Loerke, A. Mikolajka, M. Brünner, P. W. Hildebrand, A. L. Starosta, A. Dönhöfer, S. R. Connell, P. Fucini, T. Mielke, *et al.* Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. *Nature*, 468(7324):713–716, 2010.

[21] J. Velázquez-Muriel, K. Lasker, D. Russel, J. Phillips, B. M. Webb, D. Schneidman-Duhovny, and A. Sali. Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *Proceedings of the National Academy of Sciences*, 109(46):18821–18826, 2012.

[22] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–645, 2001.

[23] E. Paci, M. Vendruscolo, C. M. Dobson, and M. Karplus. Determination of a transition state at atomic resolution from protein engineering data. *Journal of Molecular Biology*, 324(1):151–163, 2002.

[24] J. Gsponer and A. Caflisch. Molecular dynamics simulations of protein folding from the transition state. *Proceedings of the National Academy of Sciences*, 99(10):6719–6724, 2002.

[25] A. De Simone, R. W. Montalvao, and M. Vendruscolo. Determination of conformational equilibria in proteins using residual dipolar couplings. *Journal of Chemical Theory and Computation*, 7(12):4189–4195, 2011.

[26] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, 2005.

[27] P. Robustelli, K. Kohlhoff, A. Cavalli, and M. Vendruscolo. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure*, 18(8):923–933, 2010.

[28] P. Neudecker, P. Robustelli, A. Cavalli, P. Walsh, P. Lundström, A. Zarrine-Afsar, S. Sharpe, M. Vendruscolo, and L. Kay. Structure of an intermediate state in protein folding and aggregation. *Science*, 336(6079):362–366, 2012.

[29] C. Camilloni, P. Robustelli, A. De Simone, A. Cavalli, and M. Vendruscolo. Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. *Journal of American Chemical Society*, 134(9):3968–3971, 2012.

[30] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[31] A. Laio and F. L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71:126601, 2008.

[32] S. Piana and A. Laio. A bias-exchange approach to protein folding. *Journal of Physical Chemistry B*, 111(17):4553–4559, 2007.

[33] V. N. Uversky and A. K. Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1804(6):1231–1264, 2010.

[34] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208, 2005.

[35] P. Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, 2002.

[36] P. Alexander, J. Orban, and P. Bryan. Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry*, 31(32):7243–7248, 1992.

[37] A. Rosato, A. Bagaria, D. Baker, B. Bardiaux, A. Cavalli, J. F. Doreleijers, A. Giachetti, P. Guerry, P. Güntert, T. Herrmann, *et al.* CASD-NMR: critical assessment of automated structure determination by NMR. *Nature methods*, 6(9):625–626, 2009.

[38] E. Fischer. Einfluss der konfiguration auf die wirkung der enzyme. *Berichte der Deutschen Chemischen Gesellschaft*, 27(3):2985–2993, 1894.

[39] A. E. Mirsky and L. Pauling. On the structure of native, denatured, and coagulated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 22(7):439, 1936.

[40] M. Sela, F. White Jr, and C. B. Anfinsen. Reductive cleavage of disulfide bridges in ribonuclease. *Science*, 125(3250):691–692, 1957.

[41] B. Gutte and R. B. Merrifield. Total synthesis of an enzyme with ribonuclease A activity. *Journal of the American Chemical Society*, 91(2):501–502, 1969.

[42] C. B. Anfinsen. Principles that govern the folding of polypeptide chains. *Science*, 181:223–230, 1973.

[43] M. L. Anson and A. E. Mirsky. On some general properties of proteins. *The Journal of General Physiology*, 9(2):169, 1925.

[44] `http://www.t3portal.org/T3_Portal_v1/!SSL!/WebHelp/ales_vancura/proteins.htm`. Online; accessed on 16th October 2013.

[45] A. D. Miranker and C. M. Dobson. Collapse and cooperativity in protein folding. *Current Opinion in Structural Biology*, 6(1):31–42, 1996.

[46] A. R. Fersht. Nucleation mechanisms in protein folding. *Current Opinion in Structural Biology*, 7(1):3–9, 1997.

[47] W. A. Eaton, V. Muñoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter. Fast kinetics and mechanisms in protein folding. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):327–359, 2000.

[48] V. Daggett and A. R. Fersht. Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences*, 28(1):18–25, 2003.

[49] C. L. Brooks, J. N. Onuchic, and D. J. Wales. Taking a walk on a landscape. *Science*, 293(5530):612–613, 2001.

[50] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[51] C. M. Dobson. Protein folding and misfolding. *Nature*, 426(6968):884–890, 2003.

[52] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.

[53] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48(1):545–600, 1997.

[54] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[55] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992.

[56] C. B. Levinthal. Are there pathways for protein folding. *Journal de Chimie Physique*, 65(1):44–45, 1968.

[57] V. N. Uversky. Intrinsically disordered proteins from a to z. *The International Journal of Biochemistry and Cell Biology*, 43(8):1090–1103, 2011.

[58] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 A resolution, obtained by X-ray analysis. *Nature*, 185(4711):416–422, 1960.

[59] R. E. Dickerson, J. C. Kendrew, and B. E. Strandberg. The crystal structure of myoglobin: Phase determination to a resolution of 2 A by the method of isomorphous replacement. *Acta Crystallographica*, 14(11):1188–1195, 1961.

[60] C. Blake, D. Koenig, G. Mair, A. North, D. Phillips, and V. Sarma. Structure of hen egg-white lysozyme: A three-dimensional Fourier synthesis at 2 A resolution. *Nature*, 206(4986):757–761, 1965.

[61] A. C. Bloomer, J. N. Champness, G. Bricogne, R. Staden, and A. Klug. Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. *Nature*, 276(5686):362, 1978.

[62] W. Bode, P. Schwager, and R. Huber. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding: The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 A resolution. *Journal of Molecular Biology*, 118(1):99–112, 1978.

[63] R. J. Williams *et al.* The conformational mobility of proteins and its functional significance. *Biochemical Society Transactions*, 6(6):1123, 1978.

[64] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review of Biophysics*, 37:215–246, 2008.

[65] D. Dolgikh, R. Gilmanshin, E. Brazhnikov, V. Bychkova, G. Semisotnov, S. Venyaminov, and O. Ptitsyn. Alpha-lactalbumin: compact state with fluctuating tertiary structure? *FEBS Letters*, 136(2):311–315, 1981.

[66] Z. Shi, R. W. Woody, and N. R. Kallenbach. Is polyproline II a major backbone conformation in unfolded proteins? *Advances in Protein Chemistry*, 62:163–240, 2002.

[67] B. Mészáros, I. Simon, and Z. Dosztányi. The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Physical Biology*, 8(3):035003, 2011.

[68] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are natively unfolded proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics*, 41(3):415–427, 2000.

[69] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.

[70] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, and Y. Zhou. SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of Biomolecular Structure and Dynamics*, 29(4):799–813, 2012.

[71] L. P. Kozlowski and J. M. Bujnicki. MetaDisorder: A meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, 13(1):111, 2012.

[72] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12(1):54–60, 2002.

[73] P. E. Wright and H. J. Dyson. Linking folding and binding. *Current Opinion in Structural Biology*, 19(1):31–38, 2009.

[74] K. Gunasekaran, C. J. Tsai, S. Kumar, D. Zanuy, and R. Nussinov. Extended disordered proteins: Targeting function with less scaffold. *Trends in Biochemical Sciences*, 28(2):81–85, 2003.

[75] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proceedings of the National Academy of Sciences*, 93(21):11504–11509, 1996.

[76] A. K. Dunker and Z. Obradovic. The protein trinity-linking function and disorder. *Nature Biotechnology*, 19(9):805–806, 2001.

[77] J. Kraut. How do enzymes work? *Science*, 242:533–540, 1988.

[78] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[79] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank. *European Journal of Biochemistry*, 80(2):319–324, 2008.

[80] `http://www.ruppweb.org/xray/tutorial/maps/bundle_map_bonds.GIF`. Online; accessed on 16-October-2013.

[81] `http://upload.wikimedia.org/wikipedia/en/0/04/Ensemble_of_NMR_structures.jpg`. Online; accessed on 16th October 2013.

[82] W. L. Bragg. X-ray analysis of proteins. *Proceedings of the Physical Society. Section B*, 65(11):833, 1952.

[83] K. Wüthrich. NMR in biological research: peptides and proteins. *North-Holland Publishing Company Amsterdam*, 1976.

[84] J. Jeener, B. H. Meier, P. Bachmann, and R. R. Ernst. Investigation of exchange processes by two-dimensional NMR spectroscopy. *The Journal of Chemical Physics*, 71(11):4546–4553, 1979.

[85] K. Wuthrich. NMR of proteins and nucleic acids. *Wiley*, 1986.

[86] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall. Cryo-electron microscopy of viruses. *Nature*, 308:32–36, 1984.

[87] T. Gonen, Y. Cheng, P. Sliz, Y. Hiroaki, Y. Fujiyoshi, S. C. Harrison, and T. Walz. Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. *Nature*, 438(7068):633–638, 2005.

[88] K. D. Gibson and H. A. Scheraga. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proceedings of the National Academy of Sciences of the United States of America*, 58(2):420, 1967.

[89] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[90] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[91] J. Söding, A. Biegert, and A. N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(2):W244–W248, 2005.

[92] S. Tanaka and H. A. Scheraga. Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–950, 1976.

[93] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[94] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, 10(2):139–145, 2000.

[95] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Bioinformatics*, 34(1):82–95, 1999.

[96] P. Cossio, D. Granata, A. Laio, F. Seno, and A. Trovato. A simple and efficient statistical potential for scoring ensembles of protein structures. *Scientific Reports*, 2, 2012.

[97] J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995.

[98] C. Chothia and A. V. Finkelstein. The classification and origins of protein folding patterns. *Annual Review of Biochemistry*, 59(1):1007–1035, 1990.

[99] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[100] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, and A. Laio. Exploring the universe of protein structures beyond the Protein Data Bank. *PLoS Computational Biology*, 6(11):e1000957, 2010.

[101] A. Šali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus. Evaluation of comparative protein modeling by Modeller. *Proteins: Structure, Function, and Bioinformatics*, 23(3):318–326, 1995.

[102] N. Eswar, B. Webb, M. A. Marti-Renom, M. Madhusudhan, D. Eramian, M.-y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, pages 5–6, 2006.

[103] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1029–1034, 2005.

[104] J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.

[105] J. Xu, M. Li, and Y. Xu. Protein threading by linear programming: theoretical analysis and computational results. *Journal of Combinatorial Optimization*, 8(4):403–418, 2004.

[106] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.

[107] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383:66–93, 2003.

[108] S. M. Larson, C. D. Snow, M. R. Shirts, and V. S. Pande. Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology. *Computational Genomics*, 2002.

[109] R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@ Home. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):118–128, 2007.

[110] W. P. Aue, E. Bartholdi, and R. R. Ernst. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *The Journal of Chemical Physics*, 64:2229, 1976.

[111] A. Rosato, J. M. Aramini, C. Arrowsmith, A. Bagaria, D. Baker, A. Cavalli, J. F. Doreleijers, A. Eletsky, A. Giachetti, P. Guerry, *et al.* Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure*, 20(2):227–236, 2012.

[112] R. A. Laskowski, J. A. C. Rullmann, M. W. MacArthur, R. Kaptein, and J. M. Thornton. AQUA and PROCHECK-NMR: Programs for checking

the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, 8(4):477–486, 1996.

[113] A. Bhattacharya, R. Tejero, and G. T. Montelione. Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure, Function, and Bioinformatics*, 66(4):778–795, 2007.

[114] J. F. Doreleijers, W. F. Vranken, C. Schulte, J. L. Markley, E. L. Ulrich, G. Vriend, and G. W. Vuister. NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwpdb. *Nucleic Acids Research*, 40(D1):D519–D524, 2012.

[115] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T. E. Malliavin, and M. Nilges. ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382, 2007.

[116] P. Güntert. Automated structure determination from NMR spectra. *European Biophysics Journal*, 38(2):129–143, 2009.

[117] T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319(1):209–227, 2002.

[118] S. Raman, Y. J. Huang, B. Mao, P. Rossi, J. M. Aramini, G. Liu, G. T. Montelione, and D. Baker. Accurate automated protein NMR structure determination using unassigned NOESY data. *Journal of the American Chemical Society*, 132(1):202–207, 2009.

[119] A. M. J. J. Bonvin, A. Rosato, and T. A. Wassenaar. The eNMR platform for structural biology. *Journal of Structural and Functional Genomics*, 11(1):1–8, 2010.

[120] T. A. Wassenaar, M. van Dijk, N. Loureiro-Ferreira, G. van der Schot, S. J. de Vries, C. Schmitz, J. van der Zwan, R. Boelens, A. Giachetti, L. Ferella, *et al.* WeNMR: structural biology on the grid. *Journal of Grid Computing*, 10(4):743–767, 2012.

[121] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo. Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23):9615–9620, 2007.

[122] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, 2008.

[123] D. Frenkel and B. Smit. Understanding molecular simulation: From algorithms to applications. *Academic Press*, 2001.

[124] J. Chen, C. L. Brooks III, and J. Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Current Opinion in Structural Biology*, 18(2):140–148, 2008.

[125] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, S. Swaminathan, M. Karplus, *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[126] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[127] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[128] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS One*, 7(2):e32131, 2012.

[129] L. Verlet. Computer "experiments" on classical fluids. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98, 1967.

[130] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, 126(1):014101–014101, 2007.

[131] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.

[132] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physics Review A*, 31(3):1695–1697, 1985.

[133] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984.

[134] B. A. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, 1992.

[135] E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chemical Physics Letter*, 156(5):472–477, 1989.

[136] G. Henkelman and H. Jónsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *Journal of Chemical Physics*, 111:7010, 1999.

[137] T. Huber, A. E. Torda, and W. F. van Gunsteren. Local elevation: A method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-Aided Molecular Design*, 8(6):695–708, 1994.

[138] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141–151, 1999.

[139] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calcula-

tions on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[140] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 16(11):1339–1350, 1995.

[141] B. Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 1995.

[142] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letter*, 86(10):2050–2053, 2001.

[143] Y. Crespo, F. Marinelli, F. Pietrucci, and A. Laio. Metadynamics convergence law in a multidimensional system. *Physical Review E*, 81(5):055701, 2010.

[144] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana. A kinetic model of Trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Computational Biology*, 5(8):e1000452, 2009.

[145] R. Martoňák, A. Laio, and M. Parrinello. Predicting crystal structures: the parrinello-rahman method revisited. *Physical Review Letters*, 90(7):075503, 2003.

[146] P. Cossio, F. Marinelli, A. Laio, and F. Pietrucci. Optimizing the performance of bias-exchange metadynamics: Folding a 48-residue LysM domain using a coarse-grained model. *Journal of Physical Chemistry B*, 114(9):3259–3265, 2010.

[147] F. Baftizadeh, P. Cossio, F. Pietrucci, and A. Laio. Protein folding and ligand-enzyme binding from bias-exchange metadynamics simulations. *Current Physical Chemistry*, 2:79–91, 2012.

[148] X. Biarnés, F. Pietrucci, F. Marinelli, and A. Laio. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Computational Physics Communication*, 183:203–211, 2012.

[149] W. Humphrey, A. Dalke, K. Schulten, *et al.* VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[150] D. Granata, C. Camilloni, M. Vendruscolo, and A. Laio. Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics. *Proceedings of the National Academy of Sciences*, 110(17):6817–6822, 2013.

[151] G. M. Clore, M. Nilges, D. K. Sukumaran, A. T. Brünger, M. Karplus, and A. M. Gronenborn. The three-dimensional structure of α1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics. *The EMBO Journal*, 5(10):2729, 1986.

[152] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of American Chemical Society*, 131(39):13894–13895, 2009.

[153] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, 2011.

[154] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR*, 44(4):213–223, 2009.

[155] Y. Shen and A. Bax. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 48(1):13–22, 2010.

[156] J. A. Pople. Molecular orbital theory of aromatic ring currents. *Molecular Physics*, 1(2):175–180, 1958.

[157] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, 2008.

[158] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, *et al.* BioMagResBank. *Nucleic Acids Research*, 36(suppl 1):D402–D408, 2008.

[159] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. Broglia, *et al.* PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computational Physics Communication*, 180(10):1961–1972, 2009.

[160] F. Pietrucci and A. Laio. A collective variable for the efficient exploration of protein beta-sheet structures: application to SH3 and GB1. *Journal of Chemical Theory and Computation*, 5(9):2197–2201, 2009.

[161] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[162] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, 134(8):3787–3791, 2012.

[163] T. S. Ulmer, B. E. Ramirez, F. Delaglio, and A. Bax. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *Journal of American Chemical Society*, 125(30):9179–9191, 2003.

[164] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79(2):926–935, 1983.

[165] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *Journal of Chemical Physics*, 103:8577–8593, 1995.

[166] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *et al.* LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[167] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51(1):129–152, 2000.

[168] L. Hou, H. Shao, Y. Zhang, H. Li, N. K. Menon, E. B. Neuhaus, J. M. Brewer, I.-J. L. Byeon, D. G. Ray, M. P. Vitek, *et al.* Solution NMR studies of the Aβ (1-40) and Aβ (1-42) peptides establish that the Met35 oxidation state affects the mechanism of amyloid formation. *Journal of the American Chemical Society*, 126(7):1992–2005, 2004.

[169] E. L. McCallister, E. Alm, D. Baker, *et al.* Critical role of beta-hairpin formation in protein G folding. *Nature Structural Biology*, 7(8):669–673, 2000.

[170] C. Camilloni, R. A. Broglia, and G. Tiana. Hierarchy of folding and unfolding events of protein G, CI2, and ACBP from explicit-solvent simulations. *Journal of Chemical Physics*, 134(4):045105, 2011.

[171] J. W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins. *Proceedings of the National Academy of Sciences*, 100(13):7587–7592, 2003.

[172] B. D. Bursulaya and C. L. Brooks. Comparative study of the folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models. *The Journal of Physical Chemistry B*, 104(51):12378–12383, 2000.

[173] `http://www.wenmr.eu/wenmr/casd-nmr-data-sets`. Online; accessed on 16th October 2013.

[174] http://lamp-lbi-43.rcs.le.ac.uk/casd/CASD-NMR-CING/data/m5/2m5o_Trieste_310/2m5o_Trieste_310.cing/2m5o_Trieste_310/HTML/summary.html#_top. Online; accessed on 16th October 2013.

[175] F. J. Blanco and L. Serrano. Folding of protein GB1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *European Journal of Biochemistry*, 230(2):634–649, 1995.

[176] G. Bussi, F. Gervasio, A. Laio, and M. Parrinello. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *Journal of American Chemical Society*, 128(41):13435–13441, 2006.

[177] C. Camilloni, D. Provasi, G. Tiana, and R. A. Broglia. Exploring the protein G helix free-energy surface by solute tempering metadynamics. *Proteins*, 71(4):1647–1654, 2008.

[178] S. H. Park, K. T. O'Neil, and H. Roder. An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry*, 36(47):14277–14283, 1997.

[179] A. Morrone, R. Giri, R. D. Toofanny, C. Travaglini-Allocatelli, M. Brunori, V. Daggett, and S. Gianni. GB1 is not a two-state folder: Identification and characterization of an on-pathway intermediate. *Biophysical Journal*, 101(8):2053–2060, 2011.

[180] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics*, 17(1):10–18, 1975.

[181] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society*, 127(2):476–477, 2005.

[182] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipo-

lar couplings and small-angle X-ray scattering. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47):17002–17007, 2005.

[183] T. Mittag and J. D. Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Current Opinion in Structural Biology*, 17(1):3–14, 2007.

[184] N. G. Sgourakis, Y. Yan, S. A. McCallum, C. Wang, and A. E. Garcia. The Alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: A combined MD/NMR study. *Journal of molecular biology*, 368(5):1448–1457, 2007.

[185] J. R. Allison, P. Varnai, C. M. Dobson, and M. Vendruscolo. Determination of the free energy landscape of α-synuclein using spin label nuclear magnetic resonance measurements. *Journal of the American Chemical Society*, 131(51):18314–18326, 2009.

[186] M. R. Jensen, P. R. L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, and M. Blackledge. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, 17(9):1169–1185, 2009.

[187] S. Côté, P. Derreumaux, and N. Mousseau. Distinct morphologies for amyloid beta protein monomer: Aβ1–40, Aβ1–42, and Aβ1–40 (d23n). *Journal of Chemical Theory and Computation*, 7(8):2584–2592, 2011.

[188] S. Vivekanandan, J. R. Brender, S. Y. Lee, and A. Ramamoorthy. A partially folded structure of amyloid-beta (1–40) in an aqueous environment. *Biochemical and Biophysical Research Communications*, 411(2):312–316, 2011.

[189] A. Abedini and D. P. Raleigh. A critical assessment of the role of helical intermediates in amyloid formation by natively unfolded proteins and polypeptides. *Protein Engineering Design and Selection*, 22(8):453–459, 2009.

Acknowledgement

This is doubtless the hardest part to write in this thesis. Because it's impossible to collect and thank all the people who have been important during these amazing four years only in few words.

At first I want to thank my supervisor at SiSSA, Prof. Alessandro Laio, to have been an excellent scientific mentor, but above all a great friend. I felt his trust in each step of my PhD, giving me complete freedom to try, fail and sometimes success also on my own ideas, and supporting me when I found difficulties. But more importantly he taught me to be always optimistic in science and life, adventures which are prohibitive without this fundamental ingredient. I hope to hold always with me this teaching as well as his friendship.

Another huge thank is for my external supervisor Prof. Michele Vendruscolo from the University of Cambridge, for the great opportunity to collaborate with him and his group. Every time that we met to discuss about the project he always enlarged incredibly its perspectives and boundaries, bringing me beyond the simple and strict view that I have on my results: I think I spent more time to understand what we really achieved in this work than in doing simulations. From his group I want to thank expecially Carlo Camilloni for his fundamental contribution to my research: he suffered all my tortures via mail and skype, remaining always patient and kind in front of my stupid errors and questions.

I want to thank also Prof. Flavio Seno and Dr. Antonio Trovato from the University of Padua, together with a previous SISSA student, Pilar Cossio, who got me involved in another project that I really loved because it opened me on other important aspects of theoretical and computational biophysics. Every meeting has been very stimulating and enjoyable, expecially the last one in Padua, with the two groups all together.

Thanks to Prof. Alberto Granato, who he gave me the opportunity to teach my first university course in Physics, and for the consideration he always demonstrates to me.

One of the things which maked my experience in SISSA so special is that I didn't have any colleagues, but just friends. Thanks to the nice atmosphere created by the Sector staff members, Alessandro, Prof Cristian Micheletti, Dr. Giovanni Bussi and Angelo Rosa, we often share very nice moments all together, without any distinction of group, age, role. This is one of the aspects that for sure I will miss more from Trieste. I really want to say "thank you" to each of them, and in particular to Sandro, Angelo, Francesca, Jessica, Alex, and to my officemates Edoardo and Paolo with whom I shared a lot of work, laughs and bad words in front of the computer. Another important thank goes to some past students and postdocs such as Pilar, Rolando, Trang, Emmanuela, Federica, and to Francesco and Salvatore, also for the time spent on our ICONA.

A special thanks to the students of my year, Gianpaolo, Giuseppe, Luca, Duvan and Xiao for all the way shared together, including hopes, doubts, problems, lunches and dinners, rides and trips. To this group I have to add also Olga who belongs to another sector but who was always with us; all these people constituted a sort of new family for me when I arrived in Trieste, and it would be impossible to be so happy here without them.

Another fundamental person for me in these years has been Fahimeh, who always treated me as her young brother, supporting me in every difficulties on work and everyday life.

I want to say thanks to all my past housemates Ciro, who adopted me when I arrive in Trieste, Massimo and especially to Barbara and his boyfriend, whose I can't really remember the name in this moment. It has been really important for me to get back a great friend!

Thanks to Emanuele, Andrea and Silvia and our meetings around the world. Thanks to be always with me, even if far away. Thanks to Eleonora who helps our African kids to have a better future, never forgetting to write me, and to Loredana for her demonstration on how life should be lived and faced.

Thanks to all my friends in Rome, because when I'm in Trieste and whenever I go back, they make me feel as I'm always with them and participating to

their lives. Thanks to Gabriel, Aurora, Simone, Cri, Andrea, Monica, Giulia, Maurizio, Massimo, Cristina, Checco, Claudia, Serena and in particular to the youngest ones, Camilla and Aulolina

Finally thanks to my true family, for always believing in me and taking care of me. Thanks to my mother, my father and my brother: what I am now, it's only thanks to them.

And thanks to Giorgia to support and share with me all the adventures that I want to face. Thanks to teach me how one can be happy in every day.

Grazie . . .