Protein Physics by Advanced Computational Techniques: Conformational Sampling and Folded State Discrimination

by

Pilar Cossio Tejada

Ph.D. Thesis

Supervisors

Prof. Alessandro Laio

Prof. Flavio Seno

Dr. Antonio Trovato

12th October 2011

Contents

Acknowledgments

I believe that these four years at SISSA have been the best of my life. During my Ph.D., I have learned how to think: how to understand difficult problems and how to solve them. I have seen the world of biophysics, which has fascinated and enchanted me with its mysteries. I have learned how to use computers and simulations as fundamental tools in science. I have met incredible researchers and professors which have taught me a lot. Fortunately, there are a lot of people that I have to acknowledge.

First of all, I would like to thank SISSA for giving me the possibility of pursuing my Ph.D., and for believing in young foreign scientists. I would like to thank the Sector of Statistical and Biological Physics, its directors and members. Which have fully supported and encouraged me, with conferences, seminars and lectures, to continue my research. I feel very proud to be ‘scientifically born’ in this Italian institution of science, and very honored to be of such great lineage of thought. The people that have constructed this school have inspired and secretly challenged me to follow their examples.

I have no words to express my gratitude towards Prof. Alessandro Laio, my guide and excellent supervisor at SISSA. He is an outstanding professor, with his inteligence and curiosity for science, he has brought always knowledge, joy, and enthusiasm to my carrier. With his pragmatism and brilliant way of thinking, he has taught me how to find different perspectives in solving a problem. With his experience, knowing which are the relevant issues that are worth thinking about (and which are not), he has helped me to maturate my own thoughts and ideas. I have to say that it is not only because of his great intellect but also because of his humanity that I have been so happy following my Ph.D. I feel very fortunate, because I

have not only found in Prof. Laio an excellent scientific colleague but also a friend for life.

I want to thank Prof. Flavio Seno and Dr. Antonio Trovato from the University of Padova, Italy. They have been my supervisors outside SISSA, and have provided me with the proper theoretical background that I needed in biophysics. With their clear and robust knowledge, they have been an excellent complement to our computational work. None of the research that we have done would of been possible without their help and support. It has been a great pleasure for me to work with them, they are not only great researchers but also great people.

I want to thank Dr. Fabio Pietrucci, which was a postdoc in the SBP group during the first two years of my Ph.D.. Fabio, patiently, taught me how to program, use computers to their maximum potential and solve the problems of biology with a physicist's perspective. After he left SISSA, I visited him several times in EPFL, Switzerland and we have done some very interesting research collaborations. I believe that postdocs are essential for guiding Ph.D. students. I have learned a lot form Fabio, and I feel very lucky to have had him as a guide with my work.

I would like to acknowledge all the people at SISSA that in one way or another have helped me with my research. In particular, Dr. Fabrizio Marinelli, Dr. Xevi Biarnes, Dr. Rolando Hong, Dr. Giulia Rossetti, Prof. Paolo Carloni, Fahimeh Baftizadeh and Danielle Granata.

I thank my former university, Universidad de Antioquia, and my physics professors. Which taught me all the scientific basis that I needed, in order to abroad the research problems that I faced in my PhD. I feel that my academic background has been excellent. Moreover, they have always motivated and supported me in pursuing my scientific carrier.

I have to say that, these four years in Trieste, have not only been good because of the scientific quality of SISSA, but also because of the amazing people that I have met through the years. I thank my friends for bringing me the pleasure of enjoying life. Moreover, I thank Gareth Bland for the unconditional love and support he has given me. Without them my life would have been very dull.

Lastly, I would like to thank my country, Colombia, land of beauty and contradictions, giving me my heritage and making me what I am. I thank my wonderful parents Jorge Cossio

and Débora Tejada, and also my family, for setting the example of how to love, be good and follow happiness.

Chapter 1

Introduction

1.1 Why Proteins?

Proteins are essential parts of organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle that are in charge of motion and locomotion of cells and organisms. Others proteins are important for transporting materials, cell signaling, immune response, and several other functions. Proteins are the main building blocks of life.

A protein is a polymer chain of amino acids whose sequence is defined in a gene: three nucleo type basis specify one out of the 20 natural amino acids. All amino acids possess common structural features. They have an α-carbon to which an amino group, a carboxyl group, a hydrogen atom and a variable side chain are attached. In a protein, the amino acids are linked together by peptide bonds between the carboxyl and amino groups of adjacent residues. The side chains of the standard amino acids, have a great variety of chemical structures and properties. It is the combined effect of all the amino acids in the sequence and their interactions with the environment that determines the stable structure of the protein (or native state). The stable tridimensional structure, in almost all cases, determines the functionality of a protein.

Biochemists often refer to four distinct aspects of a protein's structure: the *primary structure* is the amino acid sequence, the *secondary structure* are the regularly repeating structures stabilized by hydrogen bonds (i.e. α helix, β sheets and turns), the *tertiary structure* is the overall shape of a single protein molecule, and the *quaternary structure* is the structure formed by several proteins, which function as a single protein complex. In figure 1.1 an example of a prototypical protein structure is shown.

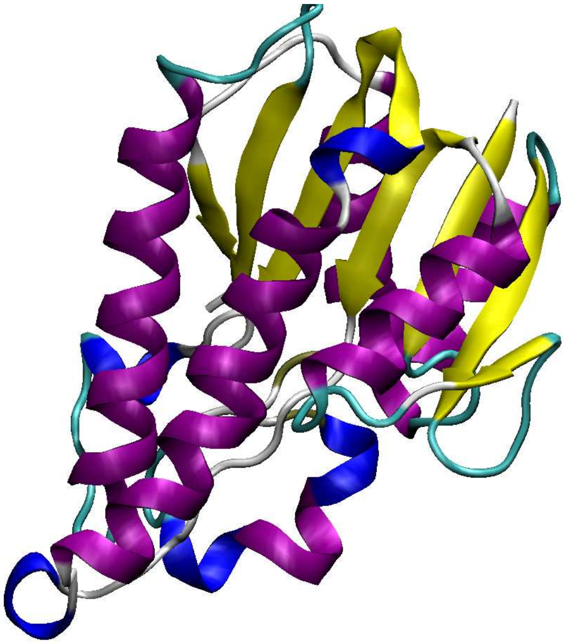

Figure 1.1: Example of a protein structure rich in α and β secondary structure. PDB code: 3D7L.pdb.

Discovering the tertiary structure of a protein, or the quaternary structure of its complexes, can provide important clues about how the protein performs its function, with which other proteins it interacts, and information about the biological mechanisms in which it is involved. Moreover, the knowledge of the structures, complexes and interactions between proteins involved in deadly pathologies will help designing drugs and curing diseases, such as cancer, neurodegenerative diseases (Alzheimer, Huntington, etc), virus infections, and much more. That is why finding out a protein's structure and functionality is one of the most important challenges in Biophysics.

1.2 Some Open Problems in Protein Physics

Most proteins are biomolecular machines essential for cell life, but some of them can also play crucial roles in lethal diseases. During the last century, biologists have made huge progress in understanding the role and function of different proteins, and the biomolecular mechanism that they pursue. Different pathologies have been understood and new drugs for their cures have been created. Much progress in the experimental part has been made, but these techniques are expensive and time consuming. Nowadays, it is also possible to study proteins with the aid of computer simulations, accurate methodologies are currently used to study diverse biological phenomena.

Currently, there are still several fundamental questions in protein physics, that still wait for a quantitative theory. In this section, we will list some of their questions together with the theoretical methods one can use to address them. We will present just a few examples of open problems in protein physics. Other important examples, like how to understand protein aggregation [1, 2], or protein DNA/RNA interactions [3, 4], will not be described.

1.2.1 The protein folding problem

Each protein exists as an unfolded polypeptide when translated from a sequence of mRNA to a linear chain of amino acids. At this first stage the polypeptide lacks any developed tridimensional structure. Due to the physical forces, amino acids interact with each other to produce some well-defined stable conformations. The physical process by which a polypeptide chain folds from a random coil into a functional tertiary structure is called protein folding. The correct tridimensional structure is essential for function, although some parts of functional proteins may remain unfolded. Failure to fold into the native structure produces inactive proteins that are usually toxic and lethal.

Decades of experimental work have led to a deep understanding of the folding process, and a large number of structures have been resolved with atomic resolution by using X-ray crystallography [5] or Nuclear Magnetic Resonance (NMR) [6]. Even though these methods are quite successful, they are expensive and time consuming. Out of millions of different active proteins sequences, from all the different species, only 70,000 structures have been solved experimentally [7]. It is also important to observe that in these cases, protein structures are measured in specific experimental conditions, like an unphysical tight packing in the crystal. Moreover, X-ray crystallography does not give much insight on how the proteins respond to the environment. Nor do they provide realible information to which are the metastable conformations of the system nor which is the folding mechanism.

Nowadays, for theoreticians it is still a huge challenge to determine the protein's tertiary structure from only the simple knowledge of the amino acid sequence. Progress has been made with bioinformatics, that uses comparative protein modelling with experimental structures as templates, in order to predict the native conformation. There are mainly two types of methodologies: *homology modeling* [8, 9] which is based on the assumption that two proteins of homologous primary sequence will share similar structures, and *protein threading* [10, 11] that scans the amino acid sequence of an unknown structure against a database of solved structures. These methods are well developed, and have been quite successful if the sequence of the protein has a high similarity with an existing sequence in the PDB.

Sometimes this is not the case or one needs information about the metastable states of the system. If this is the case, a Hamiltonian is needed to describe the system and the laws of Physics must be used. Solving the equations of motion exactly is the first thing that comes to mind, but since a protein is composed of thousands of atoms, this task is far from trivial [12]. During the last four decades, scientist have used computers to solve Newton's equations numerically in order to observe protein folding. In 1971 it was possible to simulate, for the first time, water in liquid form [13]. Four decades after, some small and fast folding peptides have been studied using computational techniques. A milestone in this field was the work by Duan and Kollman, in which the villin headpiece (a 36-mer) was folded in explicit solvent using a super-parallel computer [14]. With the help of worldwide-distributed computing, this small protein was also folded by Pande and co-workers [15, 12] using implicit and explicit solvent within a total simulation time of 300 μs, and 500 μs, respectively. Recently, with the use of super-computers Shaw *et al* [16] did a $1ms$ simulation of a WW protein domain that captured multiple folding and unfolding events.

Unfortunately, it is still very difficult to deal with the huge complexity of the conformational space of peptides whilst treating with good accuracy the physical interactions in the system (see Section 1.3). Studying by computer simulation the folding process of large proteins ($L > 80$) or slow folders are problems that remain to be fully addressed.

1.2.2 The universe of protein structures

Understanding the universe of all the possible protein folds is of fundamental and practical importance. Defining a fold as the tertiary structures adopted by protein domains, a number of key questions must be addressed. How many different folds are there? Is it essentially infinite, or is it a limited repertoire of single-domain topologies such that at some point, the library of solved protein structures in the Protein Data Bank (PDB) would be sufficiently complete that the likelihood of finding a new fold is minimal? These questions mainly arise from the fact that while the number of experimentally solved protein sequences grows linearly every year, in the last years, very few (almost none) new protein folds have been found [7]. Then, if the number of folds is finite, how complete is the current PDB library? That is, how likely is it that a given protein, whose structure is currently unknown, will have an already-solved structural analogue? More generally, can the set of existing protein folds and its degree of completeness be understood on the basis of general physical chemical principles? The answer to these questions is not only of theoretical importance, but has practical applications in structural genomic target selection strategies [17].

These issues have been addressed in many manners, most commonly with the aid of computational techniques. The first contribution in this direction was achieved in 1992 by Head-Gordon *et al* [18], who demonstrated that tertiray protein structures are stable even if all their side chains are replaced by alanine. In 2004, Hoang *et al* [19] demonstrated that all the secondary structure elements of proteins could be obtained by using a simple Hamiltonian with a tube-like model. In 2006, by studying the completeness of a library of compact homopolypeptides, that contain a protein-like distribution, Skolnick *et at* [20] have shown that the resulting set of computer-generated structures can be found in the PDB and viceversa, *i.e.* the PDB is complete. This idea is currently the most accepted one in the scientific community. Even though it seems convincing, it is still debated, for at least three reasons: *i*) Due to the large dimensionality of the conformational space, one expects that the number of possible protein-like conformations that a peptide can take is huge. *ii*) The number of distinct protein sequences which have been experimentally solved is tiny compared to the amount of genome-wide protein sequences that exist [21, 22]. Thus maybe it is the experimental techniques that limit the current existing library of folds. *iii*) Recently some experimental groups have been able to design, with computational methods, new folds, not previously observed in nature [23].

1.2.3 Protein design

As we mentioned there is a large but finite number of protein folds observed so far in nature, and it is not clear whether the structures not yet observed are physically unrealizable or have simply not yet been sampled by the evolutionary process. Methods for 'de novo' design of novel protein structures provide a route to solving this question and, perhaps more important, a possible route to design protein machines and therapeutics.

Creating a new protein from scratch is a very difficult problem. Given a structure, the objective is to find the amino acid sequence that has this structure as native conformation. The difficulties are in the fact that the space of possible sequences that one has to explore is huge (*e.g.* if one wants to design a 30 amino acid structure one would have to explore 30^{20} possibilities). Moreover, it is not even known whether the target backbone is really designable. Due to these difficulties the computational design of novel protein structures is incredibly expensive and is a sever test of current force fields and optimization methodology.

A pioneering work in this direction, was the complete design of a zinc finger protein by Mayo and co-workers [24]. Another important work was done in 2003, by Baker *et al* [23] that introduced a computational strategy that iterates between sequence design and structure prediction to design a 93-residue α/β protein with a novel sequence and topology.

However, due to the vast size and ruggedness of the conformational space to be searched and the limited accuracy of current potential functions, protein design is still an open problem. Knowing the universe of possible protein structures can be of great for designing new proteins. This problem is fundamental because its understanding could bring mayor progress in protein therapeutics and designing the molecular machines of the future.

1.2.4 Protein - protein interactions

Protein-protein interactions occur when two or more proteins bind together to carry out their biological function. Many of the most important molecular processes in the cell, such as DNA replication, are carried out by large molecular machines that are built from a large number of protein components organized by protein-protein interactions. Indeed, protein-protein interactions are at the core of the machinery on which cells are based. Proteins might interact for a long time to form part of a protein complex, a protein may carry another protein (for example, from cytoplasm to nucleus or viceversa), or a protein may interact briefly with another protein just to modify it (for example, a protein kinase will add a phosphate to a target protein). Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches [25].

In this context, the two most important questions are the following: *i*) understanding if two proteins interact, *ii*) finding out which is the quaternary conformation that two proteins have when they interact. The first issue has been studied from the perspectives of biochemistry, quantum chemistry, signal transduction and metabolic or genetic/epigenetic networks [26]. As complementary ways, various bioinformatics methods have been developed to cope with this problem, ranging from the sequence homology-based to the genomic-context based [25]. These methods for example, integrate the data from different methods to build the protein-protein interaction network, and to predict the protein function from the analysis of the network structure [26, 27]. These techniques have revealed the existence of hundreds of multiprotein complexes. Due to the huge space of conformations and possible binding positions, finding the quaternary conformation has been much more difficult. Some progress has been made with tools like docking [28]. However, much more work is needed to provide a quantitative picture of the binding interaction sites and affinities that two or more proteins can have.

In this section, we have listed some fundamental issues that need to be addressed in protein physics. We believe that new computational techniques and methodologies will not only help addressing these problems, but can become a fundamental tool for understanding biological systems and for designing new cures for pathologies.

1.3 How to address these problems by simulation?

Not only protein folding, but also the great majority of the relevant biological processes start at the macromolecular level. Examples are enzymatic reactions, protein-protein interactions, nucleic acid packaging and membrane formation. Such phenomena take place on time scales ranging from pico seconds to seconds, and are strongly coupled to atomistic dynamics (*e.g.* fluctuations in side chain and backbone atom conformations, etc). In order to study and understand these phenomena, computational models have been often used. The quality of these models depends on the accuracy with which two issues are addressed, the description of the interactions and the sampling of the configurations of the system [29].

1.3.1 Accuracy of the physical interactions

The methods used to evaluate the interactions in the system are several, differing by the resolution level. Nowadays, purely 'ab-initio' simulations, where both the nuclei and electrons are treated with quantum mechanics can be afforded only for small molecules. Commonly, the Born-Oppenheimer approximation is used assuming that the electrons relax much faster than the atomic nuclei. Thus, at any time, the Schrodinger equation for the electron system is solved by considering the external field generated by the atomic nuclei as frozen, and one is left with a nuclear-configuration dependent set of energy eigenvalues that define the ground and excited states of the system. Unfortunately, also this 'simplified' approach is computationally quite expensive, limiting its usefulness only for studying phenomena that involve tens of atoms and that occur in time scales of the order of few tens of picoseconds.

In order to study larger systems, one must abandon the quantum mechanics (QM) approach in favor of an empirical description of the interatomic forces, also called the molecular mechanics (MM) scheme. The elimination of the electronic degrees of freedom enables enormous savings of computational cost, allowing the simulation of fully hydrated biological systems. In passing from QM to MM, one has to choose an empirical form of the potential energy surface, that is normally approximated with a sum of analytical terms describing the chemical bonds, the van der Waals and the electrostatics interactions, globally called the force field (FF). The FF parameters are often fitted on the potential energy surface evaluated with QM approaches in small molecules representing parts of the system. The first macromolecular simulation of a protein, done in 1977, was a 9.2ps long molecular dynamics simulation of a bovine pancreatic trypsine inhibitor with only 500 atoms [30]. Nowadays one can reach, with the aid of super computers, at most the milliseconds in time scale [16] for systems of several thousands of atoms. As was mentioned in Section 1.2, by using pure molecular dynamics some small protein structures have been predicted, and their folding mechanism has been understood [12, 14, 16]. Moreover, phenomena like transitions of ions through channels [31], enzymatic reactions [32], and many others have been studied with all atom molecular dynamics simulations.

If one wants to analyze the biomolecular structure formation or self-assembly of supramolecular complexes that involve hundred of thousands of atoms and time scales larger than the millisecond, pure all atom molecular dynamics cannot yet provide a truly quantitative and statistically reliable description of the system.

In order to cope with the huge size of the systems, coarse-grained (CG) force fields have been developed. The idea of coarse-graining, is condensing groups of atoms into single interacting sites. Very different levels of CG have been considered, ranging from the 'united atoms' approach [33], where the atoms in an amino acid are represented as one or two beads, to mesoscale models with interacting sites representing whole proteins [29]. A critical issue here is the combination of accuracy and predictive power in the CG model. One can generally recognize three different parametrization methodologies: *i*) structure based [34, 35], *ii*) statistical/empirical potentials [36, 37, 38], or *iii*) force (or free energy) matching [39, 40]. In method (*i*), the equilibrium values of the FF terms are based on a single (experimental) structure that is by construction an equilibrium one. The few free parameters are fitted on experimental data (amplitude of thermal fluctuations, specific heat of a transition, relative equilibrium concentrations, etc). In this class, a number of popular models fall, such as the elastic network models [41]. These approaches are successful when analyzing the slow motions near the native conformation. Transferability and predictability is improved in class (*ii*), where a statistical set of structures is used to fit FF parameters with procedures based on the Boltzmann inversion [42] often integrated including empirical or experimental data. This method generates models that are thermodynamically consistent with the data set. Some popular and successful approaches, such as the CG potential generated by Bahar and Jernigan [43] belong to this class. Statistical mechanical consistency is inherent to method (*iii*), which consists of fitting the CG forces on those obtained from sets of trajectories from all-atom molecular dynamics simulations [44].

In conclusion, the choice of the potential energy function will depend dramatically on the size and time scales of the phenomena one is interested to study. The chosen accuracy level of the physical interactions, allows to cope with the computational cost of simulation whilst providing a reliable physical description of the system.

Up to now, we have described different levels of accuracy of the interactions in simulations of biological systems. But another problem in these simulations is the sampling of the conformational space. If one has a good Hamiltonian, but one cannot sample all physically relevant conformations, the simulation will not be able to provide any statistically reliable information about the process one is interested in studying. In the following section, we will describe methods that deal with this issue.

1.3.2 Conformational space search

The computational cost, in all atom molecular dynamics simulations that perform an extensive conformational space search is still very significant. Nowadays, it is possible, with the aid of super computers, to simulate thousands of atoms for several microseconds. But what if one is interested in simulating the system, with the same level of accuracy, for a longer time?

A possible manner of coping with this problem is to rely on some methodology for accelerating rare events, *i.e.* conformational changes that involve the crossing of large free energy barriers. Using these approaches, notable success has been achieved in several fields, ranging from solid state physics to biophysics and quantum chemistry [45]. Broadly speaking these methods can be classified in a few categories, according to their scope and range of applicability: *i*) Methods aimed at reconstructing the probability distribution or enhancing the sampling as a function of one or a few predefined collective variables (CVs). Examples of these methods include thermodynamic integration [46], free energy perturbation [47], umbrella sampling [48], and weighted histogram techniques [49]. These approaches are very powerful but require a careful choice of the CVs that must provide a satisfactory description of the reaction coordinate. If an important variable is forgotten they suffer hysteresis and lack of convergence. *ii*) Methods aimed at exploring the transition mechanism and constructing reactive trajectories, such as finite-temperature string method [50], transition path sampling [51] or transition interface sampling [52]. These approaches are extremely powerful but they can be applied only if one knows in advance the initial and final states of the process that has to be simulated. This is not the case, in several problems of biophysical interest. *iii*) Methods in which the phase space is explored simultaneously at different values of the temperature, such as parallel tempering [53] and replica exchange [54]. These method are very successful, and have been used together with potentials of various accuracies, to fold small globular proteins [33, 55, 56]. However, these methods require a great number of replicas, and work only if the temperature distribution is carefully chosen [53]. This has so far limited the scope of this otherwise extremely powerful methodologies.

An alternative to the traditional equilibrium approaches is provided by a class of recently developed methods in which the free energy is obtained from non-equilibrium simulations: Wang-Landau sampling [57], adaptive force bias [58] and metadynamics [59]. In the latter approach, the dynamics of the system is biased by a history-dependent potential constructed as sum of Gaussians centered on the trajectory of a selected set of collective variables. After a transient time, the Gaussian potential compensates the free energy, allowing the system to efficiently explore the space defined by the CVs. This method allows an accurate free energy reconstruction in several variables, but its performance deteriorates with the number of CVs [60], limiting its usefulness for studying biological problems that occur in an intrinsically high dimension.

More recently, a method based on ideas from metadynamics and replica exchange, called bias-exchange metadynamics [61] was introduced. In this method, a large set of collective variables is chosen and several metadynamics simulations are performed in parallel, biasing each replica with a time-dependent potential acting on just one or two of the collective variables. Exchanges between the bias potentials in the different variables are periodically allowed according to a replica exchange scheme [54]. Due to the efficacy multidimensional nature of the bias, the method allows exploring complex free energy landscapes with great efficiency. This methodology has been proved to be quite efficient for studying complex processes like protein folding, ligand-binging, amyloid formation, and ion channeling [62], these applications demonstrate the potential of BE for studying protein related phenomena.

1.4 Outlook

We have seen that there are two major issues that govern the computational study of protein systems: sampling of the conformational space and the accuracy of the potential that describes the interactions. In this thesis, we will present two main results, one related to the sampling issue, another to the optimal choice of an interaction potential for protein folding.

First, we will use bias-exchange metadynamics in order to explore the conformational space of a peptide and to find how many different protein-like structures it can take. This will give new insights in the questions described in Section 1.2.2. We show that by using molecular dynamics together with bias-exchange metadynamics it is possible to generate a database of around 30,000 compact folds with at least 30% of secondary structure corresponding to local minima of the potential energy. This ensemble could plausibly represent the universe of protein folds of similar length; indeed, all the known folds from the PDB are represented in the set with good accuracy. However, we show that nature exploits a relatively small corner of the protein fold universe, where the known folds form a small subset which cannot be reproduced by choosing random structures in the database. *Natural* folds are indistinguishable in terms of secondary content and compactness from non-natural folds. But we find that *natural* and *possible* folds differ by the contact order [63], on average significantly smaller in the former. One can argue that, due to a higher beta content, structures with large contact order could have a higher tendency to aggregate. Another possible explanation relies on kinetic accessibility, as the contact order is known to correlate with the folding time of two-state globular proteins [63]. Evolution might have selected the known folds under the guidance of a simple principle: reducing the entanglement in the bundle formed by the protein in its folded state. Bundles with shorter loops might be preferable, as they are explored more easily starting from a random coil. This suggests the presence of an evolutionary bias, possibly related to kinetic accessibility, towards structures with shorter loops between contacting residues.

The availability of this new library of structures opens a range of practical applications including the identification and design of novel folds. Motivated by this, we find the necessity of developing a potential function which is efficient and accurate for estimating if a given structure is stable with a certain amino acid composition. The second part of this thesis, will be related to the development of a new knowledge based potential for native state protein discrimination. We will test it by measuring its ability to discriminate the native state over a set of decoy structures (different structures with the same sequence), and its performance will be compared with other state-of-the-art potentials. Mainly, our potential aims at reproducing the average propensity for pair residues of forming contacts, or forming secondary structure elements, or the propensity of a residue to be exposed to the solvent. Its parameters are not obtained by training the potential on any decoy set. Instead, they are learned on a relatively small set of experimental folded structures. Moreover, and possibly more importantly, the potential depends on a much smaller number of parameters than the competitors, making its evaluation extremely computationally efficient. We find that our potential achieves excellent results when tested both on traditional decoy sets, decoys generated by molecular dynamics and on the CASP decoy sets [64]. Not only it is the best in assigning to the native structure the lowest energy value, but it also gives the largest gap between the energy of the native and the mean of the set. We for see that due to its simplicity and efficiency, this semi-empirical potential will have a lot of practical applications in protein structure prediction, protein design, and protein-protein interaction.

Chapter 2

Theoretical Background

Computer simulations of biological systems offer the possibility of investigating the most basic mechanisms of life, but pose a formidable challenge for theoreticians, due to the level of complexity that arises from both the dimension of biomolecules and their heterogeneity. As a result, computer simulations are nowadays predictive only for phenomena that take place on a relatively short time scale or in a limited region of space. Major conformational changes that involve very complex free energy landscapes, like gating in ion channels, protein-protein interaction, and protein folding, are still very difficult to study with direct all atom molecular dynamics (MD) simulation. Normally, what happens is that the system gets trapped in a local free energy minimum and there are not enough computational resources to allow the system to explore all of its conformational space. This is a significant limitation of MD, because in the majority of cases the objective of simulations is to find the most probable conformation of a system, *i.e.*, the structure with lowest free energy. An alternative, to brute-force MD, is to rely on some methodology that is capable to accelerate rare events and that allows a faster exploration of the conformational space. In the following an introduction to MD simulations will be given, and the concepts of free energy and rare events will be introduced. Moreover, two enhanced sampling techniques: metadynamics and bias exchange metadynamics (BE), will be fully described. At the end of this chapter the collective variables, and optimal set of BE parameters to study protein systems will be presented.

2.1 Molecular Dynamics

Molecular dynamics is a form of computer simulation in which atoms and molecules are evolved, for a period of time, under the action of a potential that provides approximations of known physics. 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 (FF). The FF parameters are fitted on the potential energy surface evaluated with QM approaches in small molecules representing parts of the system or are derived from experimental data. During the last two decades, a lot of effort has been dedicated for the FF parameter optimization. AMBER [65] and CHARMM [66] are most commonly used. Even though these FF have been quite successful in reproducing a lot of experimental systems, we note that they still have important limitations [67].

In a normal MD simulation, given a certain force field U and the positions and velocities of the particles at time t as $\vec{r}(t)$, $\vec{v}(t)$, respectively, the accelerations over the particles are computed using $\vec{a} = -\vec{\nabla}U/m$, and then the equations of motion are integrated at a certain time step (Δt) as to find the final positions $\vec{r}(t+\Delta t)$ and final velocities $\vec{v}(t+\Delta t)$. The most commonly used integrator is the Velocity Verlet [68] that calculates the final coordinates as follows:

$$\vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)\Delta t^2, \tag{2.1}$$

$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{\vec{a}(t)+\vec{a}(t+\Delta t)}{2}\Delta t. \tag{2.2}$$

This procedure is done iteratively and the system is evolved in time. What is important here is the value of the time step Δt. If it is too large the results will not be realistic and the simulation will crash. The appropriate Δt for atomistic simulations is usually of the order of $1fs$.

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 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, the Nosé-Hoover thermostat [69], and the Berendsen thermostat [70].

The result of a MD simulation will depend crucially on the FF, system size, thermostat choice, thermostat parameters, time step and integrator. During the last decades, MD simulations 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 Δt. For example, if one would like to simulate a process that occurs every $1\mu s$ one would at least have to integrate the equations of motion 10^9 times, and even this would not provide a statistically reliable sampling of the phenomena. Thus, if one wants to study these systems, with the same level of accuracy one is forced to rely on others methodologies that can accelerate conformational transitions and rare events.

2.2 Rare Events

2.2.1 Metastability and dimensional reduction

Let us consider a system of particles of coordinates x, in space Ω, where x can include the normal coordinates ($\vec{r}$) but also generalized coordinates such as the box edge in Parrinello Rahman [71] or the volume. The system evolves under the influence of a potential $V(x)$ and is coupled to a thermostat bath of temperature T. According to the laws of thermodynamics, the system evolves following the canonical equilibrium distribution

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

where β is the inverse temperature and $Z = \int dx e^{-\beta V(x)}$ is the partition function of the system.

In normal biological systems, like proteins, there are of the order of 10^4 atoms. Thus $P(x)$ has an incredibly large dimensionality. In order to describe the phenomena in more simple terms, what is done is to consider the reduced probability distributions in terms of some collective variables or reaction coordinates $s(x)$. Namely, instead of monitoring the full trajectory $x(t)$ of the system, a reduced trajectory $s(t) = s(x(t))$ is analyzed. For an infinitely long trajectory the probability distribution $P(s)$ is given by the histogram of s:

$$P(s) = lim_{t\to\infty}\frac{1}{t}\int dt\delta(s - s(t)), \tag{2.4}$$

in real applications $P(s)$ is estimated as

$$P(s) \approx \frac{1}{n\Delta s}\sum_{t=1}^{n}\chi_s(s(t)), \tag{2.5}$$

where $\chi_s(x) = 1$ 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 s:

$$F(s) = -\frac{1}{\beta}ln(P(s)). \tag{2.6}$$

Qualitatively, a system will display metastability if the probability $P(s)$ is large in a set of disconnected regions separated by regions in which the probability is low. A system is to be considered metastable if $F(s)$ has a characteristic shape with wells and barriers and if it presents more than one sharp minimum in its free energy profile (see Figure 2.1). A metastable system resides for the big majority of time in disconnected regions of the space and it will take the system a long time to go from one minimum to another.

2.2.2 Methods for computing free energy

The free energy as a function of a relevant and smartly chosen set of variables provides very important insight on the equilibrium and metastability properties of the system. For instance, the minima in a free energy surface corresponds approximately to the metastable sets of a

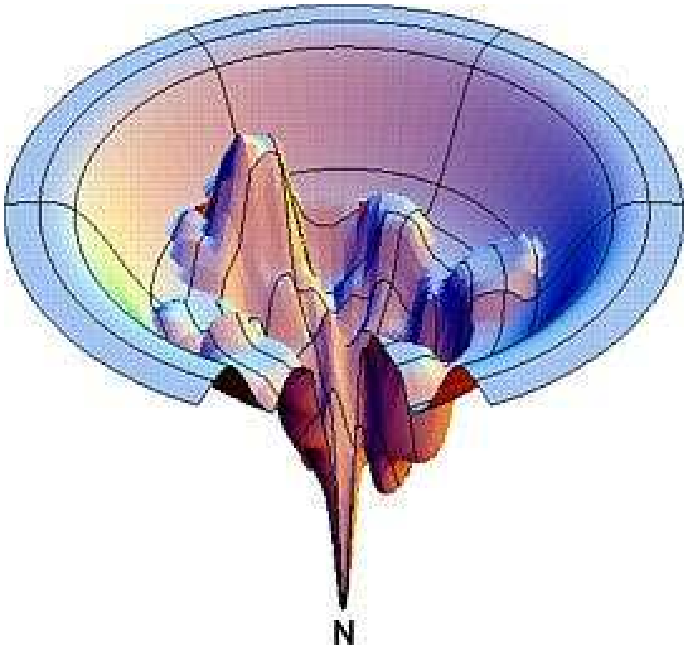

Figure 2.1: Free Energy profile of a metastable system.

system: the system spends by definition a lot of time in the minima and only rarely it visits the barrier regions in between. The free energy profiles can be used to estimate the transition time between two metastable states and can give accurate estimations of interaction energies. For instance, in chemistry one can estimate the free energy needed to break a bond in a chemical reaction by using, as a collective variable, the distance between two atoms and studying its free energy profile. In the past decades, different methods for computing free energy profiles have been developed. In the following, some of the methods will be explained.

Umbrella sampling [48] is a commonly used method 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 $s(x)$. The new biased probability distribution is

$$P^B(x) = \frac{1}{Z^B} e^{-\beta(V^B(s(x))+V(x))}, \tag{2.7}$$

where Z^B is the canonical partition function for the potential $V(x) + V^B(x)$. So, 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. It can be shown [48] that the optimal efficiency is obtained when the biased potential is $V^B(s(x)) = -F(s)$, but 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) [49], in which several histograms, constructed with different umbrellas $V^{B_i}(s(x))$, are combined in order to reconstruct a single estimate of $F(s)$. A typical bias potential would be of the form $V^{B_i}(s) = \frac{1}{2}k(s - s_i)^2$. The principal problem with this method is that the number of biasing potentials, that one has to use, scales exponentially with the dimensionality, so the computational price becomes expensive in $d > 2$.

These methodologies are based on studying the properties of the system in equilibrium. Lately, some new algorithms have been generated to exploit non equilibrium dynamics in order to compute equilibrium observables. One of these methods makes use of Jarzynski's equality [72]

$$< e^{-\beta W_t} >= e^{-\beta \Delta F}, \tag{2.8}$$

where W_t is the work performed on the system in a trajectory of time duration t. This equation provides an explicit expression for the free energy of the system in terms of the average of the exponential of the work performed on it. The main problem of this method is that the average value of $e^{-\beta W}$ is dominated by the rare trajectories for which W is small and thus $e^{-\beta W}$ is large. This hinders accuracy especially if the time duration of the trajectory is short.

In the next section we will explain a powerful methodology called metadynamics [59], in which the dynamics of the system is biased with a history-dependent potential that brings the system out of equilibrium but provides a full description of the system's free energy.

2.3 Metadynamics

In metadynamics, the dynamics in the space of the chosen CVs is driven by the free energy of the system and is biased by a history-dependent potential, $F_G(s, t)$ constructed as a sum of Gaussians centered along the trajectory followed by the collective variables up to time t. This history-dependent potential is expressed as

$$F_G(s(x),t) = \frac{w}{\tau_G}\int_0^t e^{-\frac{(s(x)-s(x(t')))^2}{2\delta s^2}}dt', \tag{2.9}$$

where τ_G is the rate at which the Gaussians are introduced and $w, \delta s$ represent the height and width of the Gaussian, respectively. In real simulations it can be calculated as

$$F_G(s(x),t) = w\sum_{t'<t} e^{-\frac{(s(x)-s(x(t')))^2}{2\delta s^2}}. \tag{2.10}$$

In the Monte Carlo or molecular dynamics simulation this bias is added to the normal potential of the system. The force generated by this biasing potential will discourage the system from revisiting the same spot and encourage an efficient exploration of the free energy surface (FES). Since the history-dependent potential iteratively compensates the underlying free energy, a system evolving with metadynamics will tend to escape from any free energy minimum through the lowest free energy saddle point. As the system diffuses on the FES, the Gaussian potentials accumulate and fill the FES wells, which permits the system to migrate, in a short time, from well to well. An example of a free energy profile filled by this biasing potential is shown in Figure 2.2.

After a while, the sum of the Gaussian terms will almost exactly compensate the underlying FES. So, for long t,

$$lim_{t\to\infty}F_G(s(x),t) \approx -F(s), \tag{2.11}$$

this property does not derive from any ordinary thermodynamic identity, since the metadynamics is a non-equilibrium process. The problem of working with history-dependent dynamics is that the forces (or the transition probabilities) on the system depend explicitly on its history. Hence it is not a priori clear if, and in which sense, the system can reach a stationary state under the action of this dynamics. In ref. [73], the validity of metadynamics was demonstrated rigorously by introducing a mapping of the history-dependent evolution into a Markovian process, in the original variable and in an auxiliary field that keeps track

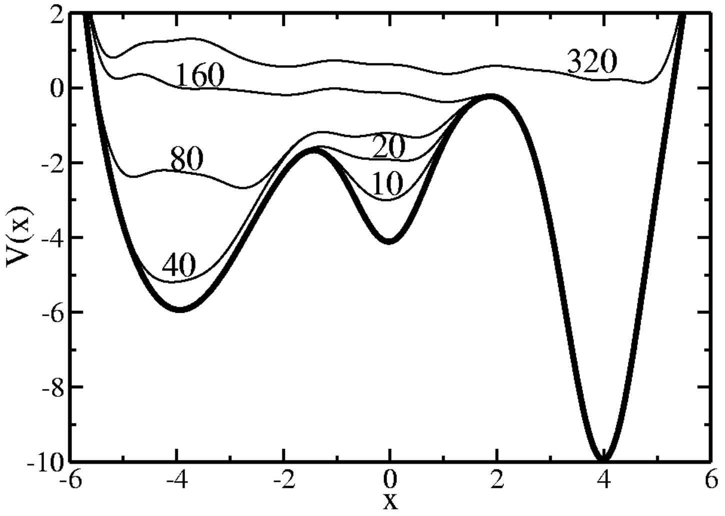

Figure 2.2: Free Energy profile filled by the biasing metadynamics potential.

of the configurations visited. For Langevin dynamics, it was shown that the average over several independent simulations of the metadynamics biasing potential is exactly equal to the negative of the free energy (Eq. 2.11), and an explicit expression for the standard deviation was found.

What makes metadynamics a flexible tool is that it can be used not only for efficiently computing the free energy but also for exploring new reaction pathways and accelerating rare events. Even though its efficacy has been proven in very different areas like condensed matter physics, chemistry, and biophysics [60], the method has some problems: *i*) Its efficiency scales badly with the dimensionality, since filling the free energy wells in high dimensions can be very expensive; *ii*) If a relevant variable is forgotten the algorithm is inaccurate. If the system performs a transition in the hidden degrees of freedom, the thermodynamic forces become inconsistent with the Gaussian potential and hysteresis will be present. To resolve the first issue, a new method that combines two different techniques, replica exchange and metadynamics, was proposed [61]. It is called bias exchange metadynamics, and it will be explained in the next section. To address the second issue a clear systematic strategy has not been yet proposed.

2.4 Bias Exchange Metadynamics

As it has been shown, ordinary metadynamics is an algorithm that can be exploited for both efficiently computing the free energy and exploring new reaction pathways, *i.e.*, for accelerating rare events. It is performed in the space defined by a few collective variables $s(x)$, where the dynamics is biased by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory of the collective variables. Qualitatively, as long as the CVs are uncorrelated, the time required to reconstruct a free energy surface scales exponentially with the number of CVs [60]. Therefore, the performance of the algorithm rapidly deteriorates as the dimensionality of the CV space increases. This makes an accurate calculation of the free energy prohibitive when the dimensionality of the space is larger than three. This is often the case for protein related problems, where it is very difficult to select *a priori* a limited number of variables that describe the process.

To overcome these difficulties a new method called bias exchange metadynamics (BE) was proposed by S. Piana and A. Laio [61]. BE is a combination of replica exchange (REMD) [54] and metadynamics [59], in which 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. 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 viceversa. In this manner, a large number of different variables can be biased, and a high-dimensional space can be explored after a sufficient number of exchanges. The result of the simulation is not however a free-energy hypersurface in several dimensions, but several (less informative) low-dimensional projections of the free energy surface along each of the CVs. The high-dimensional hypersurface can still be reconstructed using a weighted histogram approach described in Ref. [74].

In more details, let us consider N_R non correlated replicas of the system, all at the same temperature T, and each biased along a different collective variable, $s^\alpha(x)$ with $\alpha = 1, ..., N_R$.

Each replica accumulates a history-dependent metadynamics potential as

$$V_G^{\alpha}(x,t) = V_G(s^{\alpha}(x),t). \tag{2.12}$$

The replicas are allowed to exchange their configurations, like in conventional REMD and in the approach introduced in [73]. The exchange move consists on swapping the atomic coordinates x^a and x^b of two replicas a and b (evolved under the action of two different history-dependent potentials), selected at random among the N_R available. The move is accepted with a probability:

$$P_{ab} = min(1, e^{\beta\left(V_G^a(x^a,t)+V_G^b(x^b,t)-V_G^a(x^b,t)-V_G^b(x^a,t)\right)}) \ . \tag{2.13}$$

The normal potential energy of the system cancels out exactly for this kind of move. If the move is accepted, the CVs of replica a perform a jump from $s^a(x^a)$ to $s^a(x^b)$, and for replica b from $s^b(x^b)$ to $s^b(x^a)$. The exchange moves introduce a jump process on top of the ordinary metadynamics evolution. This greatly increases the diffusivity of each replica in the whole CV space. The working principle of the approach is better illustrated in a simple example. Consider a dynamics on a two-dimensional potential like the one in Figure 2.3. If one performs simple metadynamics biasing x, one obtains an estimate of the free energy affected by large errors: indeed, the system jumps between the two wells at the bottom and only rarely jumps to the two wells at the top (due to rare thermal fluctuations). Obtaining the correct free energy would require taking the average over several transitions along y. In practice, the free energy profile will not converge.

In Fig. 2.4 we show the result of a simulation consisting of two metadynamics on two replicas, one biasing x, the other y. From time to time, we allow the two replicas to exchange configurations, accepting the exchange according to Eq. 2.13. Even if the computational cost has only doubled with respect to the simulation above, one observes a very significant reduction of the hysteresis: now the metadynamics potential almost exactly compensates the free energy, both as a function of x and y, indeed, the profiles are practically flat lines. This shows that the hysteresis is much reduced, and that, like in ordinary metadynamics, in BE

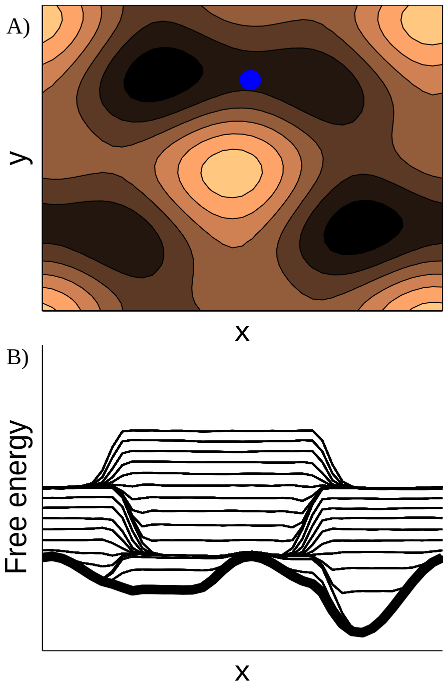

Figure 2.3: A) Example of an intrinsically two dimensional free energy surface. B) Not converged free energy profile, if only one collective variable is biased.

the Gaussian potential converges to the negative of the free energy.

2.4.1 Choice of the collective variables

Similar to the other methods that reconstruct the free energy as a function of a set of generalized coordinates, in BE the choice of the CVs, $s(x)$, 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. Unfortunately, there is no *a priori* recipe for finding the correct set of CVs, and in many cases it is necessary to proceed by trial and error. To choose the proper set, one needs to exploit some basic knowledge on the topological, chemical, and physical properties of the system. In the case of proteins, that are chains of amino acids with well-defined topological features, the commonly used CVs are:

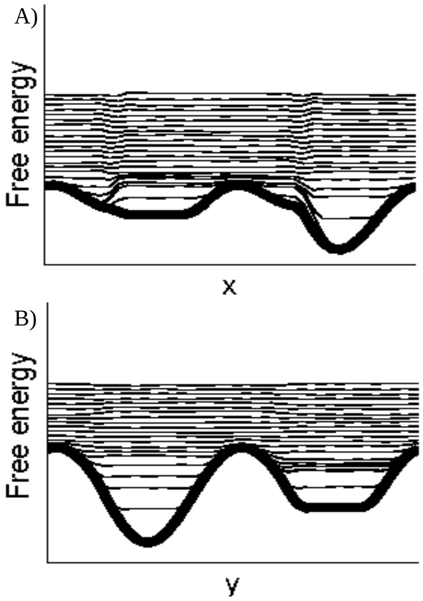

Figure 2.4: Converged free energy profiles, of the system shown in Fig. 2.3-A, when the two collective variables (x, y) are biased with BE.

- *The coordination number*: this is probably the most used CV. It is defined as:

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

  with

  $$C_{ij} = \frac{1 - (\frac{r_{ij}}{r_0})^n}{1 - (\frac{r_{ij}}{r_0})^m} \tag{2.14}$$

  (or an analogous smooth function of r_{ij}) where r_{ij} is the distance between atoms or groups i and j, and m, n are exponents that allow to tune the smoothness of the function. This CV can be used to count the number of contacts, *e.g.*, chemical bonds, hydrogen bonds, or hydrophobic contacts, between selected sets of atoms or groups in the protein.

- *Dihedral distance*: it measures the number of dihedrals ϕ_i (involving the backbone atoms C-N-C$_\alpha$-C) that are similar to a reference ϕ_0:

$$N_D = \sum_i \frac{1}{2}[1 + cos(\phi_i - \phi_0)]. \tag{2.15}$$

It can be efficiently used to count how many residues belong to an α-helix. In principle, it can also count the number of residues belonging to a β-sheet. We use only the diheral ϕ because it provides a clear discrimination between structures in α-helix ($\phi_0 = -45°$) with those in β-sheet conformations ($\phi_0 = 135°$). However, in this case, controlling the dihedrals is normally not sufficient to drive the system towards the correct configuration, a β-bridge, which requires the formation of the correct hydrogen bonds between residues that are at a large sequence separation [75].

- *Dihedral correlation*: it measures the similarity between adjacent dihedral angles of the protein backbone:

$$D_C = \sum_i \frac{1}{2}[1 + cos(\phi_i - \phi_{i-1})]. \tag{2.16}$$

Since secondary structure elements α-helices or β-sheets have a correlation between successive dihedrals, also this CV can be used to detect the presence of secondary structure elements.

- *Beta-rmsd*: In Ref. [75] it was noticed that despite the impressive variety of beta-structures observed in proteins, they are composed of 3+3 residue blocks which are remarkably similar. Regardless of the amino acid composition, all 3+3 beta blocks are within 0.08nm of RMSD. This fact allowed the definition of an ideal (*i.e.*, average) ‘beta block’. With this, it is possible to define a CV which counts how many fragments of 3+3 residues in the protein chain belong to a β-sheet secondary structure, by computing their RMSD with respect to the ideal beta conformation [75]:

$$S = \sum_\alpha n \left[\mathrm{RMSD}\left(\{\mathbf{R}_i\}_{i\in\Omega_\alpha},\ \{\mathbf{R}^0\}\right)\right], \tag{2.17}$$

$$n\,(\mathrm{RMSD}) = \frac{1 - (\mathrm{RMSD}/0.1)^8}{1 - (\mathrm{RMSD}/0.1)^{12}}, \tag{2.18}$$

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 $\left\{\mathbf{R}^0\right\}$ are the corresponding atomic positions of the ideal beta conformation. In the case of antiparallel beta, all sets Ω_α of residues of the form (i, i+1, i+2; i+h+2, i+h+1, i+h) are summed over in Eq. 2.17. For parallel beta, sets (i, i+1, i+2; i+h, i+h+1, i+h+2) are instead considered. For each residue, only backbone N, C_α, C, O and C_β atoms are included in the RMSD calculation.

The same procedure can be also applied to define the ideal alpha helix block formed by six consecutive residues, in order to define a CV measuring the amount of alpha secondary structure. In this case the sum in Eq. 2.17 runs over all possible sets Ω_α of six consecutive protein residues (i, i+1, i+2, i+3, i+4, i+5), and $\left\{\mathbf{R}^0\right\}$ are the atomic positions of the ideal alpha block.

The first three types of CV were successfully used to fold with BE some small proteins such as Advillin [76], Tryptophan Cage [74], Insulin [77]. But in the case of larger and more complex proteins the alpha/betarmsd CVs may prove more efficient, as it has been shown by simulations on GB1 and SH3, which are about 60 amino acids long and have a large content of secondary structure [75]. In the following chapter, we will use alpha/betarmsd CVs in order to enhance the exploration of the possible protein-like conformations that a peptide can take.

2.4.2 Choice of the BE parameters

In this section we discuss how to choose the parameters that are specific to BE simulations with protein related problems. Part of the analysis presented here is based on Ref. [78], where extensive benchmarks were performed using a coarse-grained force field, UNRES [33], and comparing BE with standard REMD. The authors considered 1E0G, a 48-residue-long α/β protein that folds with UNRES into the experimental native state.

Frequency of exchange moves

In BE exchanges between the bias potentials in the different variables are attempted every exchange time (τ_{exch}). Normal metadynamics is recovered taking $\tau_{exch} = \infty$, and the capability of the method of exploring a high dimensional CV space derives from choosing an appropriate exchange frequency (namely, from choosing a finite τ_{exch}). In REMD it is known that it is marginally better to exchange as often as possible, as exchanges always enhance decorrelation. The exact role of τ_{exch} in BE is less clear: in order to be effective and lead to genuine transitions from different free energy minima, a bias along a specific CV has to be active for the time required to diffuse across the free energy well. In other words, if the direction of the bias changes very frequently (small τ_{exch}), the system could have the tendency to remain close to the free energy minima and not to overcome the barriers. Extensive folding simulations of 1E0G indicate that very frequent or very rare exchanges ($\tau_{exch} \to \infty$) make BE marginally less efficient (see Fig. 2.5). For this system, it was found that the optimal exchange time is $\tau_{exch} = 120$ ps, which gives the system enough time to explore the local free-energy wells before a change in the direction of bias. If the exchanges are not performed at all ($\tau_{exch} = \infty$), folding is observed within an average time that is ~ 5 times larger. This shows that using BE (with practically any finite $\tau_{exch} = 120$) leads to a significant improvement of the capability of the system in exploring the configuration space.

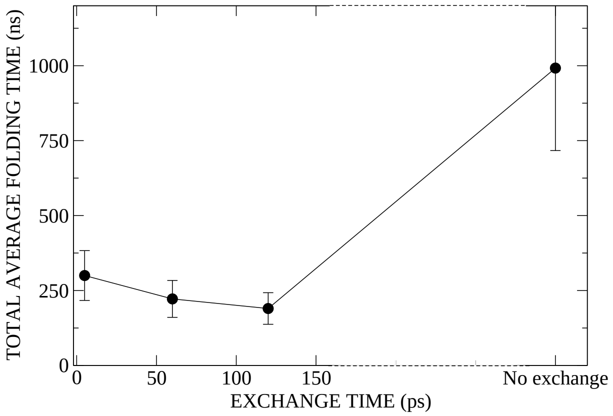

Figure 2.5: BE folding simulations on 1E0G at $T = 280$ K with different τ_{exch}. The Gaussian height is held fixed $w = 0.012 kcal/mol$, and a set of 8 CVs has been used (see Ref. [78] for details). The average folding times are shown. Error bars are obtained by averaging over 20 independent simulations.

Effect of the temperature

In standard REMD the replicas span a wide temperature range and their number has to be increased if the system is large [79]. Instead, in BE all the replicas can have the same T, because the enhanced sampling does not rely on temperature jumps but on the effect of the bias. This is an advantage, if the protein is modeled with an explicit solvent Hamiltonian, since the large increase in the number of degrees of freedom, due to the addition of the solvent molecules, does not require an increase in the number of replicas. As shown for 1E0G in Fig. 2.6, with increasing temperature the system is able to find faster the folded state of the protein, with best performance between $T = 315$ and $350K$. Comparing this with the results reported by Ref. [80] for the specific heat of the system, the temperature which optimizes the performance of BE is close to the peak of the specific heat. This is an expected result due to the temperature dependence of the Boltzmann probability of conformational transitions $e^{-\Delta E/k_\beta T}$: sampling is enhanced by a high T, but if T is higher than the critical temperature, the system will mostly explore unfolded structures reducing the efficiency in localizing the folded state.

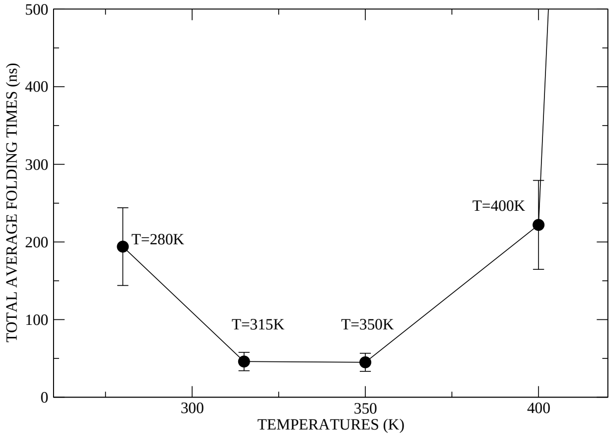

Figure 2.6: Effect of temperature on BE folding simulations of 1E0G. $\tau_{exch} = 20$ ps, $w = 0.012$ kcal/mol, and a set of 4 CVs has been used (see Ref. [78] for details). The average folding times are shown. Error bars are obtained by averaging over 20 independent simulations.

Dimensionality of the bias potentials

The efficiency of standard metadynamics degrades with increasing dimensionality of the Gaussian hills, as it takes more time to fill with Gaussians a high dimensional space than a low-dimensional one. On the other hand, if the free-energy landscape of the system is intrinsically high dimensional and one uses only a one-dimensional bias, the metadynamics estimate of the free energy is affected by large errors. BE provides, at least in principle, a solution to this problem, as one can explore high dimensional free energy landscape by using several low-dimensional biases. Benchmarks indicate that using one-dimensional Gaussians increases the efficiency of the method by a factor of ≈ 2 compared with two-dimensional Gaussians [78] (see Fig. 2.7).

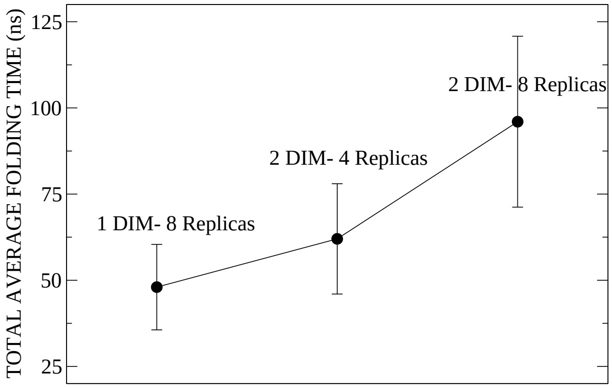

Figure 2.7: Effect of the dimensionality of the bias potential on BE folding simulations of 1E0G. Average folding times for the simulations that use one or two dimensional hills with different number of replicas. (see Ref. [78] for details). Error bars are obtained by averaging over 20 independent simulations.

Comparison with replica exchange

In Ref. [78], the performance of BE was compared to that of REMD by calculating the folding times of protein 1E0G with diverse simulation setups. When using the optimal BE setup for this system: six one-dimensional CVs (see Ref. [78] for details), $T = 315 - 370K$, Gaussian height $> 0.02kcal/mol$, and $\tau_{exch} = 20 - 120ps$, 1E0G was found to fold systematically within $\sim$ 20 ns of total simulation time. It was also found that the time required to fold the same system with replica exchange was more than an order of magnitude higher ($\sim$ 390 ns). Remarkably, even if the BE setup is not optimal, the performance of the method remains adequate (~ 100 ns for most of the setups). This makes the approach a viable candidate for attempting to fold average size proteins by an accurate all-atom explicit solvent force field.

All this theoretical background will be used in the following chapter as necessary tools to explore extensively the conformational space of a protein, and clarify issues related to how big is the universe of protein structures, and how many different protein-like conformations can a single peptide take.

Chapter 3

Exploring the conformational space of a protein

3.1 Introduction

The total number of different protein folds (or topologies) that exist in the protein data bank (PDB) is much less than the number of protein sequences that have been solved (see Fig. 3.1). Indeed, folds are evolutionarily more conserved than sequences and the same fold can house proteins performing different biological functions [81, 82]. Thus a fundamental question concerns the extension of the library of protein folds: are the observed structures a small fraction of the whole fold universe? If so, then is it because evolution has not yet run enough to explore it or rather a selection principle is what has slowed down/stopped the search for alternatives?

Addressing these issues on the basis of the principles of physics and chemistry is currently at the center of intense investigation. For a few proteins, native backbone geometries were shown to be closely mimicked by local energy minima of poly-alanine chains [18]. More recently, a unified approach to the origin of protein folds was proposed in which the inherent anisotropy of a chain molecule, the geometrical and energetic constraints placed by hydrogen bonds, steric hindrance and hydrophobicity yield a free energy landscape with minima

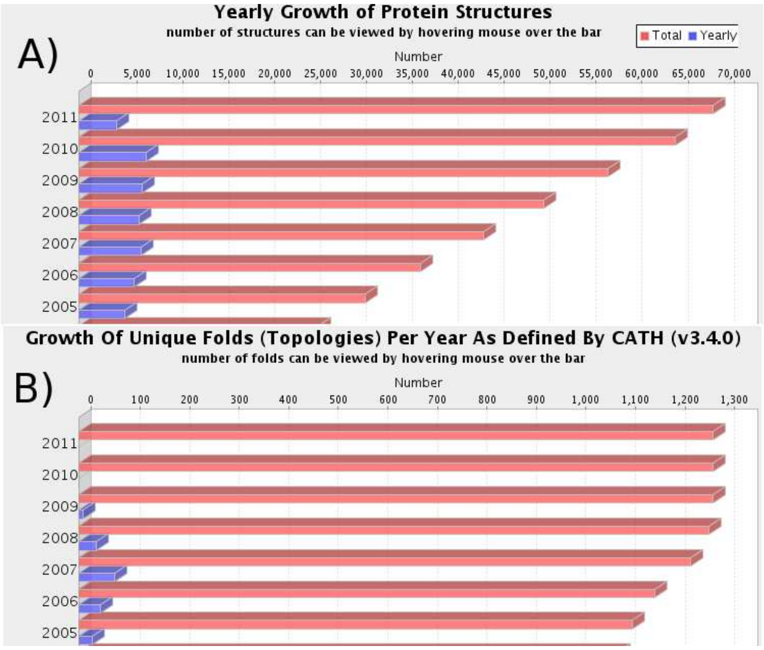

Figure 3.1: A) Number of protein sequences (Blue: per year, Pink: Total) solved experimentally in the PDB. B) Number of different protein folds (or topologies) in the PDB.

resembling protein structures [19, 83, 84]. One of the predictions is that a limited library of folds exists. Likewise based on a coarse grained model, Zhang et al proposed [20] that there is a one-to-one correspondence between the Protein Data Bank (PDB) library and the structures that one can obtain with a homopolymer from the requirement of "having compact arrangements of hydrogen-bonded, secondary structure elements and nothing more" [20].

Recent advanced sampling methods, described in Chapter 2, allow us to address these issues by accurate atomistic simulations. In this chapter we describe the results of a 50 μs molecular dynamics simulation of a 60-residue polypeptide chain performed with an accurate all-atom interaction potential and a setup specifically designed in order to extensively explore the configuration space. In the simulation we visit practically all the $\sim$ 300 folds observed in nature for proteins of comparable length. At variance with what found in [20], we find that natural folds are only a small fraction of the structures that are explored. Many of the structures found in our simulation resemble real proteins (in terms of secondary content, stability and compactness) but have not been observed in nature. This finding immediately rises a question on the nature and meaning of these novel folds: why are they not exploited in real proteins? Do natural folds have something ‘special’ or have they simply been selected at random?

3.2 Simulation setup

By using bias exchange metadynamics (BE) [61] an enhanced sampling technique described in Chapter 2, we simulate a 60 amino acid polyvaline (VAL60) described by an all-atom potential energy function. MD simulations are performed using the AMBER03 [85] force field and the molecular dynamics package GROMACS [86]. Simulations are mainly performed in vacuum, but tests have been performed also in water solution (see below). The temperature is controlled by the Nose-Hoover thermostat, and the integration time step is 2 fs.

In order to explore the conformational space we use BE with 6 replicas. The CVs are described in detail in Section 2.4.1 [75] and are designed in order to evaluate by a differentiable function of the coordinates the fraction of a secondary structure element (α helix, parallel β sheet and antiparallel β sheet). For instance, for the antiparallel β sheet the variable counts how many pairs of 3-residue fragments in a given protein structure adopt the correct beta conformation, measured by the RMSD from an ideal block of antiparallel beta formed by a pair of three residues. We here use six CVs, defined with:

$$S = \sum_{\alpha} n\left[\mathrm{RMSD}\left(\{\mathbf{R}_i\}_{i\in\Omega_\alpha},\ \{\mathbf{R}^0\}\right)\right], \tag{3.1}$$

$$n\left(\mathrm{RMSD}\right) = \frac{1-\left(\mathrm{RMSD}/0.1\right)^8}{1-\left(\mathrm{RMSD}/0.1\right)^{12}}, \tag{3.2}$$

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 the ideal alpha or beta conformation. Three CVs use an alpha $\{\mathbf{R}^0\}$ template, where their sum in Eq. 3.1 runs over one third of the protein, respectively. One CV uses an anti-parallel beta template, and two CVs use a parallel beta template, with the sum in Eq. 3.1 running over the whole protein. The Gaussians entering in the metadynamics potential are added every 10ps. Their height and width are 5kcal/mol and 0.3. Exchanges between the biasing potentials are allowed every 25ps. The exchanges greatly enhance the capability of the dynamics of exploring new structures [61, 87]. These parameters have been optimized according to the criteria described in Section 2.4.2.

The main scope of this work is exploring exhaustively the conformational space of an average length polypeptide described by a realistic potential energy function. The final choice of simulating VAL60 in vacuum with $\varepsilon_r = 80$ at 400 K, and then optimizing the configurations with $\varepsilon_r = 1$ was taken after considering several alternatives. We first considered performing the simulation on a 60-alanine in vacuum (ALA60), as alanine is used in Ref. [18]. This system was evolved using the BE setup described above for 1.5μs generating ∼1200 structures with a high secondary content.

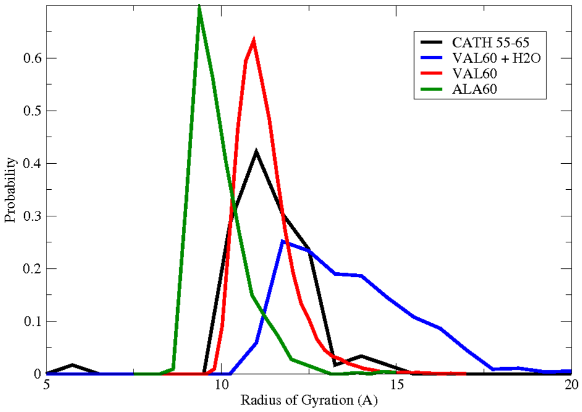

Figure 3.2: **Distribution of the radius of gyration for the VAL60, VAL60+H$_2$0, ALA60 and CATH 55-65 sets of structures.**

However, the structures generated in this manner are too compact to be comparable with experimental structures. Indeed, the histogram of the radius of gyration for ALA60 is peaked approximately 1 Å too low with respect to what observed for real proteins of similar length (see Fig. 3.2). This is due to the relatively low steric hindrance of the side chain of ALA. The same histogram computed for VAL60 is instead fully consistent with the distribution observed in real proteins. We also performed test simulations of VAL60 solvated in TIP3P water at 350K. This system was evolved for 0,8μs with the same BE setup. In this case ∼ 1400 structures with a high secondary content are found, but most of these structures are not independent, as the correlation time in water is much larger than in vacuum. More importantly, the structures generated in water have on average a large radius of gyration (see Fig. 3.2). This is an indication that at 350 K the system explores mainly non-compact structures. Of course, one could perform the simulation at lower temperature, but this would lead to an even larger correlation time, making an exhaustive exploration of the configuration space too time consuming with existing computational resources. Performing the exploration with $\varepsilon_r = 80$ is not strictly necessary, as test simulations performed with $\varepsilon_r = 1$ are also able to explore structures with a high secondary content. However, VAL60 with $\varepsilon_r = 1$ has a relatively high preference for β structures (see Fig. 3.11). With $\varepsilon_r = 80$, α and β structures become approximately isoenergectic for VAL60, removing a possible bias in the exploration.

3.3 Results

3.3.1 A library of 30,000 folds

With the optimal setup for BE, we simulate VAL60, in vacuum, at 400K, and $\varepsilon_r = 80$, for 50 μs. This allowed creating $\sim$ 30,000 structures characterized by a significant secondary content and a small radius of gyration. From the trajectory, one sees that exploration proceeds mostly by local reorganization of secondary structure elements. From time to time the system unfolds completely, exploring a totally independent topology. A selection of the 30,000 structures is represented in Figure 3.3-A and a repository, with their all-atom configuration, is available in http://www.ploscompbiol.org/. This is our first major result: *it is possible to generate by molecular dynamics at an all-atom level a library of tens of thousands of folds.*

One wonders if the structures that are explored in this manner have protein-like topologies only because of the bias, and would fall apart in normal conditions. In order to address this issue, for all the structures generated by molecular dynamics we performed a steepest decent (SD) simulation with ε_r =1, aimed at localizing the closest potential energy surface minimum. For the last configuration the C$_\alpha$ RMSD was calculated with respect to the initial structure. The distribution of this quantity is shown in Figure 3.4. Most of the structures do not drift significantly apart from the initial configuration, and the C$_\alpha$ RMSD remains relatively

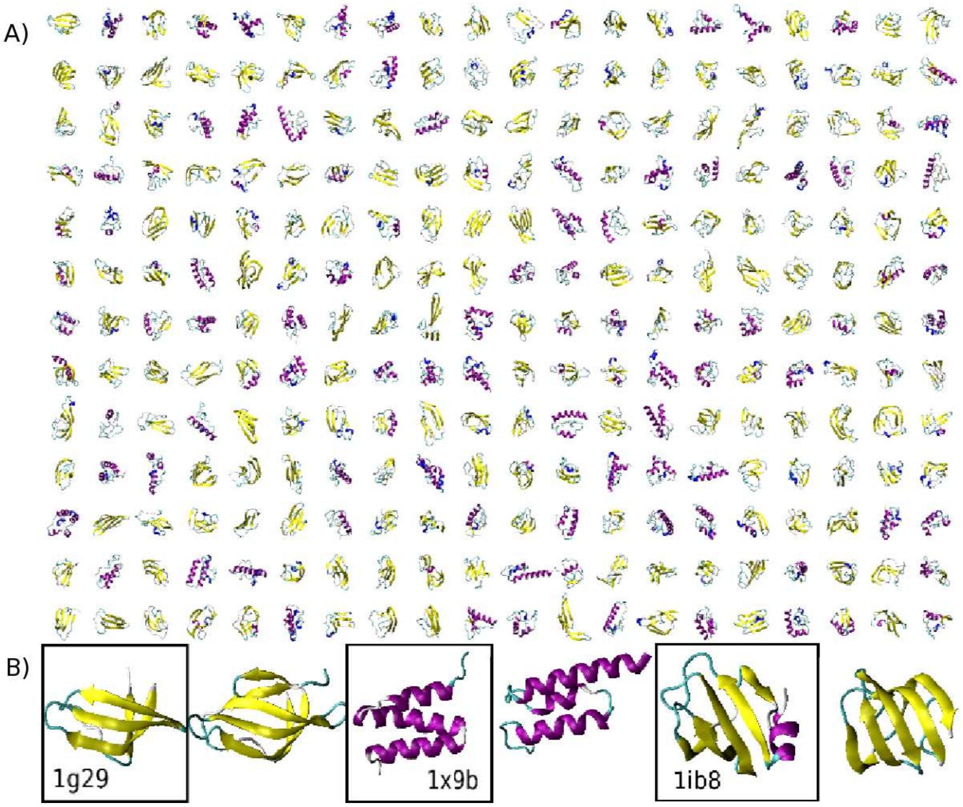

Figure 3.3: **Gallery of representative VAL60 structures generated by molecular dynamics**(A): A selection of 260 out of the 30,000 structures generated by MD, visualized by VMD [88]. The structures were selected from the 50 μs molecular dynamics trajectory if they satisfied the following conditions: (i) have more than 30% of secondary content according to DSSP [89] (ii) have a gyration radius smaller than 15 Å; (iii) be separated more than 50 ps in simulation time. The structures obtained in this manner are further optimized by steepest decent with $\epsilon_r = 1$ until a local potential energy minimum is reached (see Methods). (B): Examples of successful alignments. The CATH structure is represented together with its VAL60 equivalent for three cases.

small, within 2 Å in most cases. Thus, we conclude that the VAL60 structures generated by molecular dynamics are close to local energy minima. The set of structures generated in this manner form the database on which we perform the analysis.

We also checked if the structures that are generated in this manner are stable if the homopolymer chain is formed by another amino acid. At this purpose, $\sim$ 1500 VAL60 structures were chosen randomly. For each of these structure the valines were replaced by alanines (ALA60). Following the same procedure described above, a SD simulation was run until the closest local minimum is reached. The C_α RMSD from the initial ALA60 configuration was calculated. The distribution of this quantity is shown in Figure 3.4. Quite remarkably, even if one changes the amino acid sequence from VAL60 to ALA60 the structures do not change significantly, remaining within $2-3$ Å of C_α RMSD from the initial structure. This confirms the prediction of Ref. [18]. These checks demonstrate that in normal conditions the VAL60 structures are at l

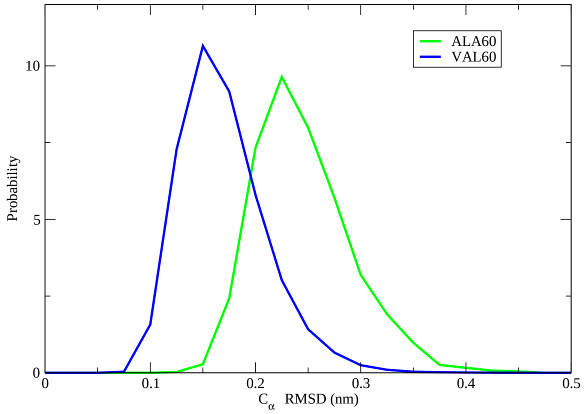

Figure 3.4: **C_α RMSD distributions for the 30,000 VAL60 and the ~1500 ALA60 minimized through SD.** The RMSD is measured with respect to the initial configuration.

Even though these structures correspond to local minima, one still wonders if their structural quality is good and if they resemble real proteins. In order to address this issue, we monitored several structural quantities on our dataset. In Figure 3.5-A we show the Ramachandran plot of the VAL60 structures. One can see that the dihedrals populate the allowed regions. The relative height of the various peaks is determined by the probability to observe the different secondary structural elements and the random coil in the full dataset. The 'stereochemical quality' of the VAL60 set was also assessed using PROCHECK [90]. This program provides an overall quality measure, called G-factor, which takes into account dihedrals, bond lengths and angles, as compared with stereochemical parameters derived from well-refined, high-resolution structures. If the G-factor is higher than -1.0 the structure is considered to be 'normal'. In Figure 3.5-B the G-factor distribution is shown for the VAL60. For a comparison, we computed the same distribution also for the structures of length smaller than 75 amino acids belonging to the CATH database [91].

We also used PROCHECK to estimate the average hydrogen bond energy. The distribu-

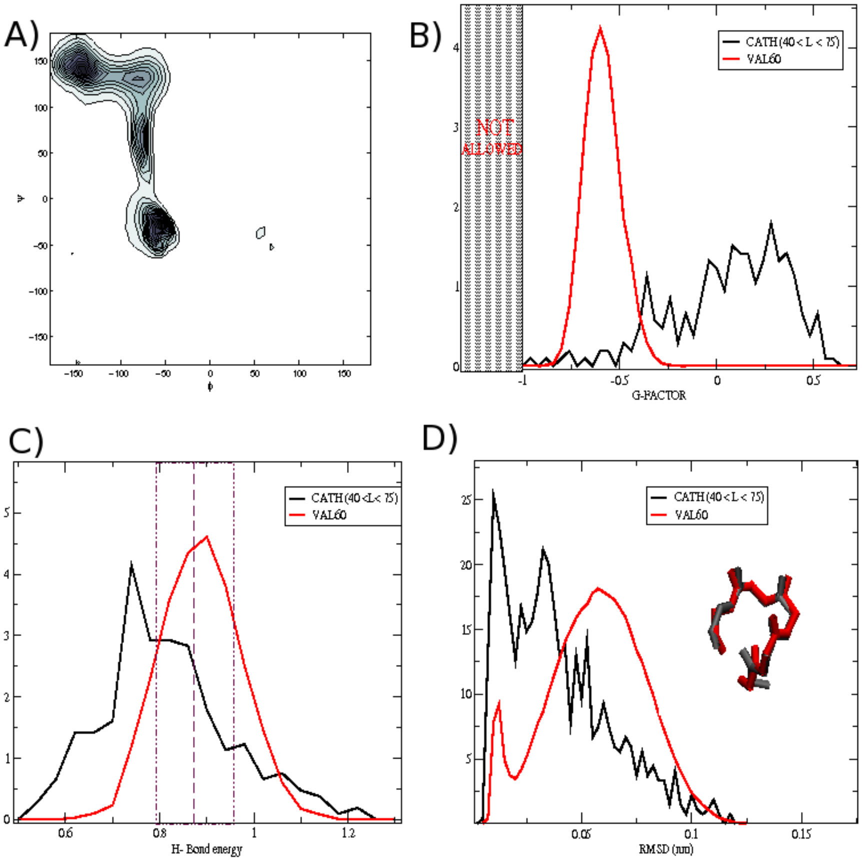

Figure 3.5: **Structural quality assessment for the VAL60 set.** (A) Ramachandran plot for the VAL60 structures. (B) G-factor [90] distribution and (C) H-bond energy distributions [90] for the VAL60 and CATH ($40 < L < 75$) sets. (D) Minimum RMSD distribution for a set of 150000 5 amino acids long fragments of the VAL60 set, and 1000 fragments of the CATH set. Inset: an example of an alignment between two fragments with a RMSD of 0.7Å.

tions of this quantity for VAL60 and CATH is shown in Figure 3.5-C and compared (dash line) with its ideal mean and standard deviation [90]. For the VAL60 set the G-factor and the H-bond energy, though not as good as for CATH, are in accordance with what is expected for realistic proteins. Lastly, in order to check if medium size structures generated by our sampling procedure are representative of the PDB, the VAL60 structures were fragmented in small 5 amino acids long structures and were compared by backbone RMSD [92] to all the fragments of the same length found in CATH. The minimum RMSD value was obtained for each small fragment. The distribution of this quantity is shown in Figure 3.5-D. It is found that the VAL60 fragments have on average at least one CATH structure within 0.6Å of RMSD. All the structural descriptors we considered demonstrate that our structures are protein-like. The distributions are similar but not identical to the one of real proteins, due to the fact that in our simulation we considered an homopolymer formed by only one amino acid, valine.

3.3.2 How many independent structures?

In order to understand how many independent structures are actually explored, and if the set contains all the known folds, a measure of the degree of similarity between two protein structures is needed.

We used the TM-align approach [93]. This method, regardless of the primary sequence of the two proteins, attempts to align their secondary structure elements allowing insertions and deletions of residues. The fraction of aligned residues is called *coverage*, and is the first measure of similarity. Afterward, the algorithm finds the rotation and translation that minimizes the relative distance between pairs of aligned residues (*RMSD*). The optimal coverage and RMSD are then combined into a single similarity measure, the *TM-score*. It is defined as

$$TM - score = max\left[\frac{1}{L_{target}}\sum_{i}^{L_{ali}}\frac{1}{1+(d_i/d_0(L_{target}))^2}\right], \tag{3.3}$$

where L_{target} is the length of target protein that other PDB structures are aligned to; L_{ali} is the number of aligned residues; d_i is the distance between the ith pair of aligned residues, and $d_0(L_{target}) = 1.24(L_{target} - 15)^{1/3} - 1.8$ is a distance parameter that normalizes the distance so that the average TM-score is not dependent on the protein size for random structure pairs. The original version of the TM-align algorithm has been modified in order to assign the secondary structure elements with more accuracy. Instead of considering only the C_α coordinates as in Ref. [93], our modified version reads for each protein the secondary structure assignment given by DSSP [89]. When the proteins have different lengths, the length of the target protein is used in the TM-score definition [93]. The TM-score is equal to one for two identical structures. Two structures are considered to represent the same fold if their TM-score is greater than 0.45, while for two randomly chosen structures the TM-score is approximately equal to 0.25.

In order to find the independent structures we proceeded as follows: first we selected the structure with the largest number of neighbors, namely with the largest number of structures at a TM-score larger than 0.45. We assign it as the first independent structure and remove it, together with all its neighbors, from the list of structures. We iterate this procedure until the list is empty. In Figure 3.6 we plot the number of independent structures found as a function of the number of structures explored by MD. This data can be accurately reproduced with a double exponential fit ($RMS = 0.0128$), which allows estimating as $\sim$ 10,000 the number of independent structures that would be explored in an infinitely long MD run.

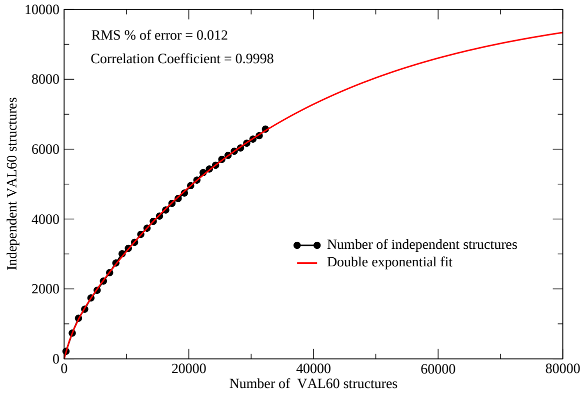

Figure 3.6: **Number of independent VAL60 structures as a function of the number of structures obtained in the MD trajectory.** Data fitted with a double exponential function of the form $9630(1 - e^{\frac{-x}{33948}}) + 619(1 - e^{\frac{-x}{1387}})$.

3.3.3 The majority of natural folds between $40 - 75$ a.a. are reproduced

By using the tests previously reported, we find, that at first sight, the VAL60 structures cannot be distinguished from folds adopted by proteins. Following Ref. [20], we checked if the set of structures generated by molecular dynamics reproduces all the known folds in the PDB.

For comparing the VAL60 structures with the existing folds in nature four different databases were considered: PDBSelect [94], TOP500 [95], CATH [91] and all the independent single domain proteins from the PDB (SD-PDB), as obtained searching the keywords: *protein*, *one chain*, and *no ligands* in the pdb repository. Each set was filtered by selecting only the proteins that had length L between 40 and 75 amino acids, had more than 30% of secondary structure, had no gaps and a gyration radius smaller than 15 Å. In order to find the independent structures, each set was further screened using the procedure described in the previous section. After applying all these screens, the number of independent folds in PDBSelect, TOP500 and SD-PDB is smaller than 100. Instead the CATH set still contains 265 folds. For this reason, we choose CATH as our reference library of folds. The names of the CATH structures used in this work are given in the Appendix.

For each structure in the CATH database, we searched, in the set of the 30,000 structures of VAL60 generated by molecular dynamics, for its most similar structure as quantified by the TM-score. In Fig. 3.3-B, three CATH structures with their respective VAL60 equivalent are shown. As shown in Fig. 3.7, for almost every CATH structure it is possible to find a VAL60 structure that is very similar. For CATH structures of length between 55 and 65 amino acids the average coverage is 75%, and the average RMSD is of only 2.8 Å. The VAL60 set reproduces, with even greater success, CATH structures of shorter length. Instead, structures of 65 or more amino acids are reproduced less accurately, as the maximum coverage that can be attained is, by definition, smaller than their length. However, even in these cases, the RMSD restricted to the aligned residues is small, of 3 Å or less. Comparison of the VAL60 set with even longer chains is not considered here: the long chains can contain extra secondary structure elements that do not significantly affect the quality of the alignment but change the topological details of the fold.

The excellent capability of the VAL60 set of reproducing the known folds is confirmed by monitoring the progress of exploration as a function of the number of structures found during the simulation. At this purpose, we assumed that a CATH structure is ‘found’ when molecular dynamics explores a VAL60 structure whose TM-score (with respect to the CATH structure) is higher than 0.45. Visual inspection reveals that two structures of similar length and of relative TM-score larger than 0.45 are structurally and topologically similar. In Figure

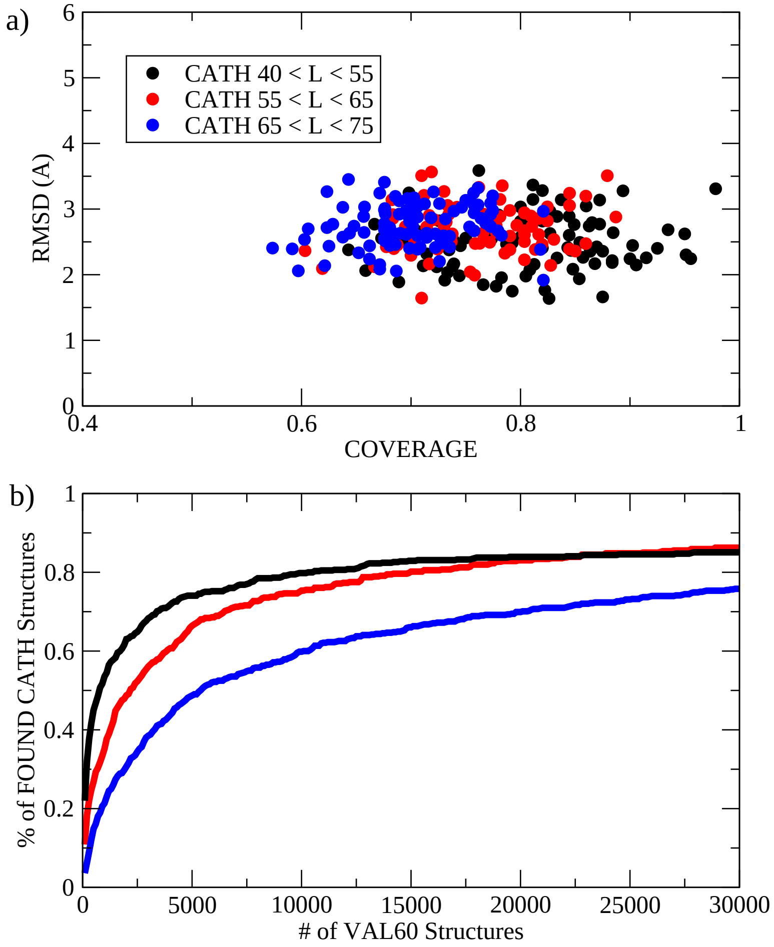

Figure 3.7: **Similarity between the VAL60 set and the PDB structures from the CATH database.** (a) Coverage vs RMSD for the CATH proteins divided in different length classes with respect to their most similar VAL60 structure. (b) Percentage of CATH structures that are reproduced by a structure in the VAL60 set (TM-score $>$ 0.45) as a function of the number of the VAL60 structures obtained in the simulation.

3.7-b we plot, for different length classes, the fraction of CATH structures that are found as a function of the number of VAL60 structures (which is approximately proportional to simulation time). At the end of the simulation, for length L=55-65 the fraction of found structures is 86% (85% for L=40-55 and 78% for L=65-75). 100 % of the structures of length L=40-65 are reproduced within a TM-score of 0.4. This shows that the computational setup used in this work allows us to explore the majority of the folds in nature, at least within the limited range of lengths considered.

3.3.4 The universe of *possible* folds is much larger than the PDB

The exploration of VAL60 structures by molecular dynamics proceeds in an almost random manner, with no obvious preference for a specific class of folds or secondary structure element. Indeed we checked that it is, on average, equally likely to find a specific CATH structure as finding a VAL60 structure for the second time. For this, we consider a small fraction of the MD trajectory used for generating the VAL60 dataset. In this fraction of the trajectory $\sim$ 2000 independent structures are generated. Using the rest of the trajectory, we compute the number of times n that each of these structures is observed (namely, the number of times a structure with relative TM-score larger than 0.45 is visited). The histogram of n is calculated for 20 different sets, each including 100 VAL60 structures.

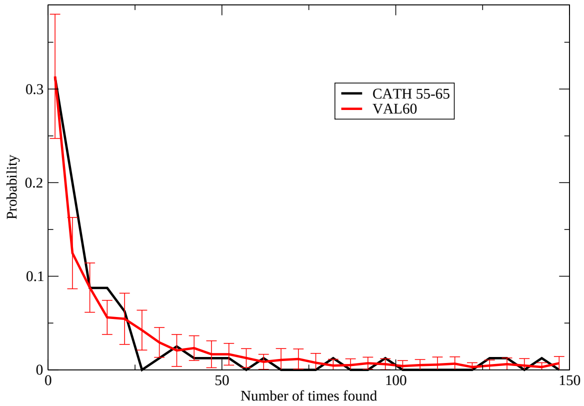

Figure 3.8: **Probability of finding n times a CATH structure and a VAL60 structure.**

Its average and standard deviation (error bars) are plotted in Fig. 3.8. This is compared to the same histogram computed for the CATH set with $55 < L < 65$ ($\sim$80 structures). Strikingly, the two histograms are very similar, indicating that the probability of finding a CATH structure in this length range is similar to the probability of finding a VAL60 structure a second time. In other words, in our sampling strategy there is no particular bias for generating a structure observed in nature. However, one realizes that the two sets of structures, CATH and VAL60, cannot be fully equivalent. Indeed, according to a clustering procedure done in Section 3.3.2, in 50 μs the simulation explores $\sim$7,000 independent structures, much more than the structures in CATH ($\sim$ 300 in a length range between 40 and 75).

One could argue that finding or not a one-to-one correspondence might just depend on the chosen similarity threshold [96]. In order to quantitatively investigate this issue, we addressed the following question: Do structural descriptors exist whose distributions are different between the two sets CATH and VAL60? If the answer is yes, a biased search mechanism reflecting an evolutionary pressure may be envisaged. Otherwise a random search mechanism in a continuous structure space may be enough to account for the choice of the observed folds out of all possible structures. While at first sight structures belonging to the VAL60 and CATH sets look indistinguishable, a more detailed analysis reveals that several VAL60 structures include a large fraction of parallel β-sheets. This secondary structure element is much less common in the CATH set restricted to L<75. We quantify this observation by looking at the distributions of normalized contact order and the contact locality. Two residues are considered to be in contact, if at least one pair of their heavy atoms is found at a distance smaller than $4.5\mathring{A}$ and they are separated by at least three residues in sequence. Then, the contact order (CO) [63] is defined as the average sequence separation between contacting residues divided by the chain length. The contact locality (CL), is a structural descriptor that counts the fraction of contacting residue pairs which are formed within the same half of the chain [97]. The total number n of pairwise contacts is $n = n_N + n_C + n_{NC}$, where n_N and n_C are the contacts between residues both belonging to the half of the chain towards the N-terminus and the C-terminus, respectively, and n_{NC} are the contacts between residues belonging to different halves of the chain. CL is then defined as $CL = (n_N + n_C)/n$.

In Figure 3.9-a the CO vs CL is plotted for the CATH set 40 <L< 75, and the VAL60 set. The distribution of CATH is significantly restricted towards lower CO and higher CL values with respect to VAL60, consistent with the observation that parallel β-sheets are found less frequently in CATH.

We also checked that the CO distribution computed for the subset of VAL60 that are recognized to be similar to CATH is largely overlapping with the CO distribution for the CATH set (see Fig. 3.9-b). This demonstrates the consistency of the similarity measure provided by the TM-score.

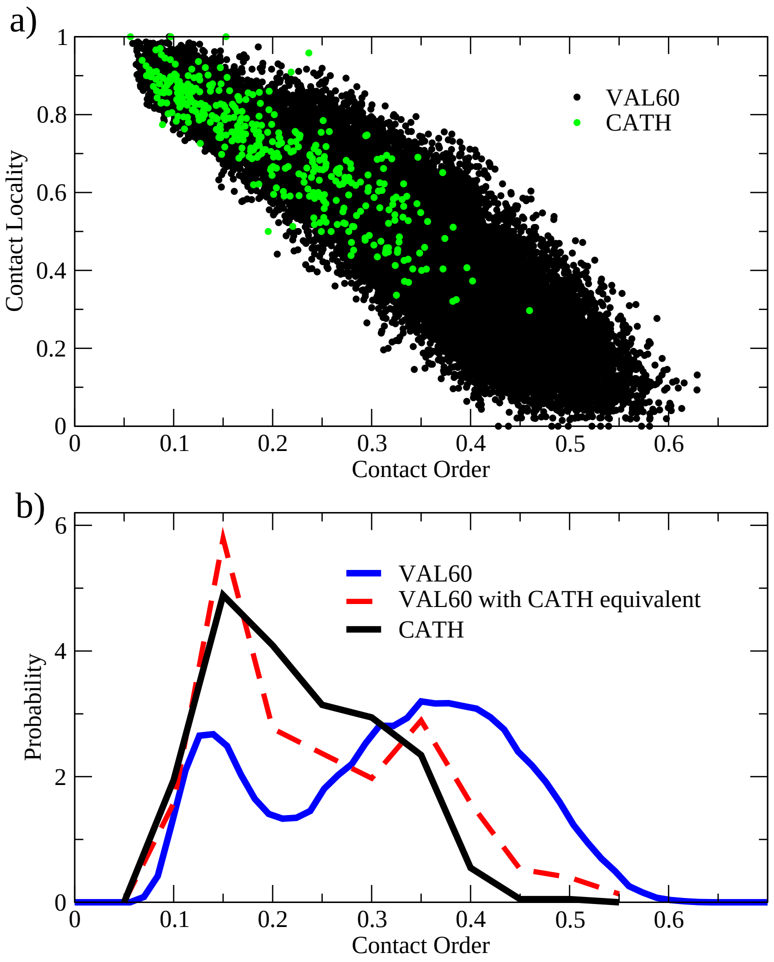

Figure 3.9: **Contact order and Contact Locality distributions of CATH and VAL60.** (a) CO vs CL represented for the CATH set of length 40 <L< 75, and the VAL60 set. (b) CO distributions for the CATH set of length 55 <L< 65, VAL60 set and for the subset of independent VAL60 structures that have TM-score > 0.45 with a structure in the same CATH set. Independent structures are obtained as described in Methods.

Finally, we also analyzed the distribution of the CO restricted to the different structural classes. The CO distributions was calculated for all-α structures and all-β structures of CATH and VAL60. The results are shown in Figure 3.10. While the bias towards low CO is present for all-β structures, for all-α structures it is not effective. It is also remarkable that the CO distribution for β structures in the VAL60 set that are similar to a CATH structure is very similar to the probability distribution for the all-β CATH structures.

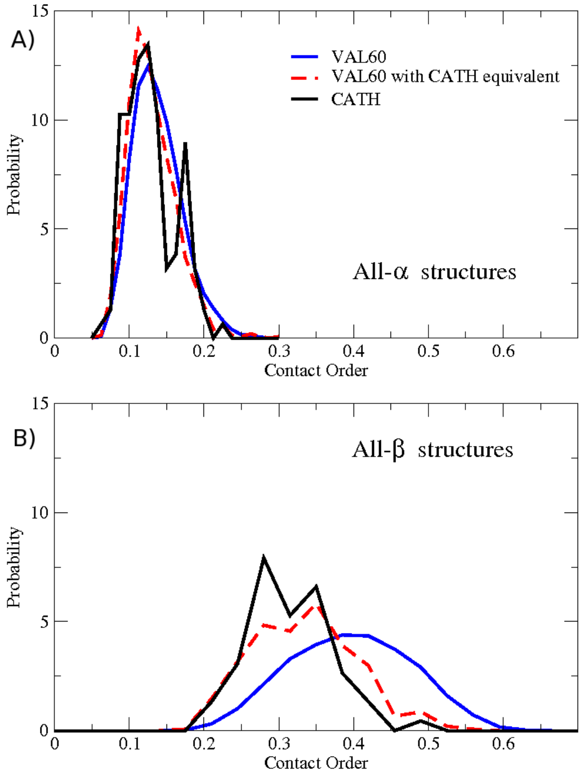

Figure 3.10: **Contact Order for different structural classes** The CATH and VAL60 sets divided in two structural classes: all-α structures, or all-β structures.

3.4 Discussion

By using atomistic simulations and bias exchange metadynamics [61] we have generated a database of $\sim$ 30000 structures corresponding to energy minima of a 60 amino acids polypeptide. Clearly, the length of 60 amino acids used in the simulation does not provide a complete representation of the full protein universe, which includes a very large amount of much longer proteins. However, our results indicate that, within the limited length range we considered, the VAL60 set is indeed representative of the space inhabited by real proteins. In fact, this set includes all the folds existing in nature for proteins of similar size, confirming that the observed protein folds are selected based on geometry and not on the chemistry of the aminoacid sequence [98, 99, 100, 18, 19, 83, 84, 20]. However, we find that the known folds form only a small fraction of the full database. Natural folds are indistinguishable in terms of secondary content and compactness from non-natural folds, but are characterized by a relatively small contact order and a relatively high contact locality. Why has nature made this choice? One can argue that, due to a higher beta content, large CO structure could have a higher tendency to aggregate. Another possible explanation relies on kinetic accessibility, as the contact order is known to correlate with the folding time of two-state globular proteins [63]. Evolution might have selected the folds under the guidance of a simple principle: reducing the entanglement in the bundle formed by the protein in its folded state. Bundles with shorter loops might be preferable, as they are explored more easily starting from a random coil.

How has nature been able to select low contact order structures? In order to address this issue, we investigated the role of specific amino acids in selecting a fold among the possible structures. At this scope, we compared the correlation between potential energy and CO of the structures obtained by energy minimization of VAL60 and ALA60 (see Section 3.3.1).

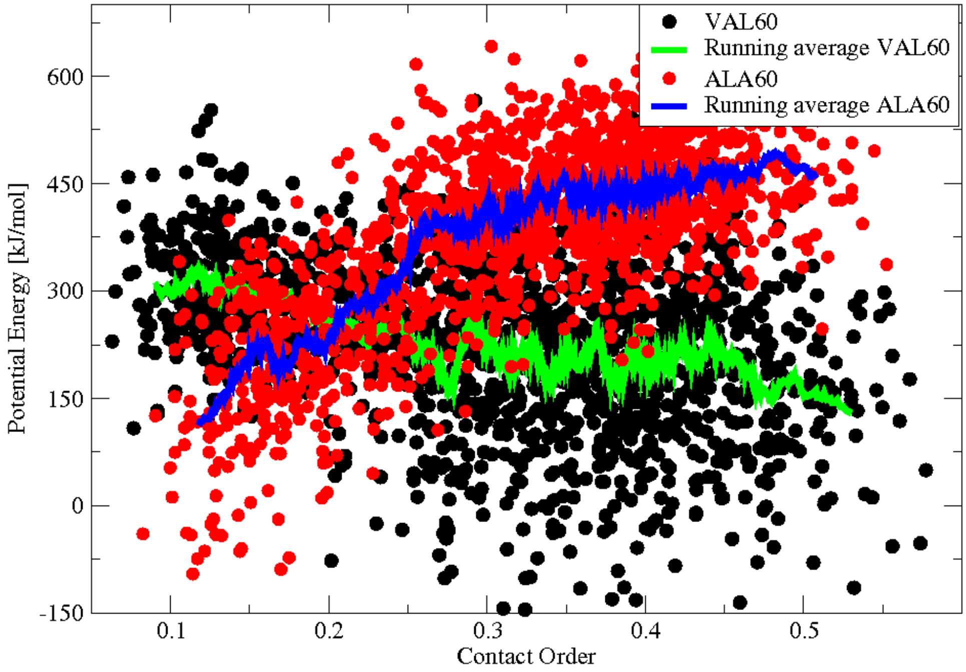

Figure 3.11: **Correlation between potential energy and contact order for VAL60 and ALA60 structures.** For a subset of ∼1500 structures from the VAL60 set we generated a corresponding set of ALA60 structures by finding the local potential energy minima after conversion of valine into alanine residues (see Methods). We then sorted all the structures according to their CO. Each point in Figure 4 corresponds to a structure. We also represent the running average of the energy over a window of 50 structures.

Figure 3.11 vividly demonstrates that different low energy structures may be discriminated when different sequences are mounted on all the possible “presculpted” structures [19]. Whereas energetically VAL60 prefers structures with high CO and a large content of strands, ALA60 promotes conformations with low CO and which are rich in helices. Evolution, possibly also guided by the kinetic bias hypothesized above, can then proceed by using a repertoire of 20 types of amino acids, to select and design the sequences which minimize the free energy of a desired structure against other competing structures.

We have shown that generating a huge set of possible protein-like structures is feasible with a computational analysis based only on physico-chemical information, with no need of modeling or pre-determined protein structural knowledge. The protein-like character of these structures is assessed via Ramachandran and G-factors, coordination number and contact order which are, mostly by construction, well satisfied. However, these features might be too general and perhaps there could be other structural descriptors that distinguish in a more rigorous manner native-like folds from those that are not. As a final remark, we believe that the VAL60 structures and the computational procedure to generate them, also with different types of amino acids and with different lengths, may play a role in future developments. The availability of a rich library of possible folds and realistic decoys could allow for advances in the two main applicative challenges in protein physics: the prediction of the native state of any given sequence and the design of the sequence folding into a desired fold. Moreover, it provides an important test set on which to define native-likeness beyond most commonly used features. The VAL60 structures might be also used to check predictions in synthetic biology [101]. Furthermore the library could be exploited to obtain models of misfolded protein structures related to neurodegenerative diseases [1].

Motivated with the huge number of new protein folds found in this work, our further goal is to design a new non-existing protein topology. Due to the complexity of this problem (see section 1.2.3), we find ourselves in the necessity of constructing an efficient and accurate knowledge based potential that gives an energy score to a tridimensional structure with a certain amino acid composition. In the next chapter, by using Bayesian Analysis we will construct this potential and compare its performance with other state-of-the-art scoring functions.

Chapter 4

BACH: Bayesian Analysis Conformational Hunt

4.1 Introduction

Knowledge based potentials are energy functions derived from databases of known proteins that empirically capture aspects of the physical chemistry of protein structure and function. These potentials are widely used in numerous applications because of their relative simplicity, accuracy and computational efficiency. They have been used in protein design, in 'ab-initio' protein structure prediction, in the assessment of the stability of mutant proteins, in deciphering protein docking and protein-DNA interactions (see Section 1.3.1). They are also the fundamental tools for protein structure analysis and in model quality assessment, namely in recognizing the native state of a protein sequence within an ensemble of putative alternative structures (decoys). Fold recognition is a problem of growing complexity since very accurate sets of decoys, at atomistic level, can nowadays be rapidly and efficiently obtained through very sophisticated templated based methods like ROSETTA [102] or TASSER [103], or even through molecular dynamics simulations [104], as seen from the previous chapter. Being able to recognize the folded state in these large sets is still considered extremely challenging.

In this chapter, we will introduce a statistical potential achieving an unprecedented level of accuracy in scoring the folded state in various decoy sets. This performance is not obtained by training the potential on the decoy sets themselves. Instead, the parameters are learned on a relatively small set of experimental folded structures. Moreover and possibly more importantly, the potential depends on a much smaller number of parameters than other potentials, making its evaluation extremely computationally efficient and its definition robust.

Knowledge based potentials rely on the thermodynamic hypothesis that the native structure has the lowest free energy of all states under native conditions. Consequently the aim is to construct scoring functions whose global minimum corresponds to the known native structures of different sequences. In principle [105] it is possible to apply directly this requirement and to extract, via learning algorithms, the free parameters by solving a set of inequalities which impose that the native state of a protein is recognized as having an energy below other conformations of the same length, that should represent excited states of the system. The alternative conformations are either chosen by threading [106] or by using existing sets of decoys. A major advantage of such a scheme is the possibility of verifying directly whether the chosen parametrization of the energy is appropriate. In fact, if it is not possible to adjust the parameters to satisfy all the inequalities, the scoring function will be inadequate, this is typically the case [107]. However this approach depends crucially on the quality of the chosen alternative structures and a single unphysical decoy may invalidate the full procedure.

Most often, statistical potentials are derived in order to reproduce the distributions of different structural features obtained from a sample of native structures. The majority of knowledge based potentials employ Boltzmann law [42] to convert into potentials the relative observed frequencies of a structural event, as compared to the frequencies of the same event on a reference state in absence of interactions. The theoretical basis of Boltzmann inversion have been questioned [108, 109] and other ideas such as Bayesian analysis [110], linear and quadratic programming [111] and information theory [112] have been invoked to justify the approach.

Several forms of knowledge based potentials have been developed and explored in the last decades. They can be categorized depending on how proteins are modeled, on how interactions are defined, and on how the reference state is chosen. Both residue and atomic level potentials have been proposed. Interactions are weighted with either distance dependent [113] or independent pairwise potentials [114, 115] or even keeping into account the chemical distance between residues [116]. Knowledge based potentials have also been derived for complex structural features such as many body interactions, torsion angles and solvent accessibility. The reference state typically results from an hypothetical system and it has been estimated for example by using quasi-chemical approximations [114], isotropic reference states [116, 117] and on the basis of reshuffled systems [118, 119]. Recently, composite scoring functions, obtained with the combination of different potentials, have been introduced to boost the performances of the single contributions. One of the most successful methods of this type, QMEAN6 [113] uses four statistical potentials terms covering the major structural features of proteins and two additional terms describing the agreement of predicted and calculated secondary structure and solvent accessibility. The relative weight of different contributions are chosen by optimizing the performance on a set of decoys.

To benchmark the potentials several decoy sets have been released, generated by various methods of de novo prediction, by using comparative modeling molecular dynamic simulations and loop modeling [120, 121, 122, 123]. Very recently [124] it has been argued that some of the most popular decoy sets are deficient, since obvious differences between the native state and the decoys are present which make the discrimination trivial. It is now believed that decoy sets composed by models submitted during the biennal CASP (Critical Assessment in Structure Prediction) [64] competition are not affected by such a problem. Since CASP7 a separate fold recognition section (model quality assessment program) has been introduced, so that the blind comparison of methods on CASP models is now part of the competition. Given the set of decoys, several criteria have been proposed to assess the performances of various potentials, such as the ability to distinguish the native structure or near native structures, the correlation coefficient between the energy of the models and a similarity coordinate (such as RMSD or GDT [125]) and the Z-score. It emerges that consensus based methods perform significantly better than those accepting single models. However, these approaches take the consensus of different quality assessment programs based on the idea that, if they are independent of each other, the probability of a correct prediction is higher than for the best single program. Thus, they are of little use in practical applications. Among the knowledge based potentials, all atom distance dependent statistical potentials like QMEAN6 [113], ROSETTA [102], and RF_CB_SRS_OD [119] produced the best performances (see Ref. [119] for a detailed description of the different performances). Available potentials require a number of parameters that may reach the order of hundred of thousands. This large number of parameters is unpleasant, since one expects that the general physical chemical laws that rule the folding process should be captured in a relatively simpler way. Moreover, the necessity of using and optimizing too many details may affect the robustness of the potentials and the possibility to use them for different purposes than fold recognition: indeed, the most efficient statistical potentials are normally not used for protein-protein interaction [126].

We developed a new potential aimed at reproducing the average propensity for pair residues of forming contacts, or forming secondary structure elements, or the propensity of a residue to be exposed to the solvent. The ideas at the basis of its construction are the following:

- In order to decide if two residues are in contact or not, we consider the full atomic configuration of the system, assigning a contact only if two of their side chain atoms form a true physical contact, *e.g.* a hydrophobic contact or a hydrogen bond.
- We say that two residues form a contact involved in a secondary structure element if they form the correct backbone-backbone hydrogen bonds, according to DSSP [89].
- In order to decide if a residue is solvent exposed or not, we compute explicitly its solvent accessible surface, that, in turn, depends on the full atomic configuration of the system.
- Residues that are not seen in the PDB structure, even if they are part of the sequence expressed by the crystallographer, are considered as disordered residues and are included in the countings.
- Finally, we take into account both the possibility of having or not having (negative event) a structural feature (contact with another residue or with the solvent or hydrogen bonds). This turned out to be one of the key ideas for boosting the performance of the potential.

This manner of constructing the potential makes it a rather complex function of the coordinates, in which, however, the specific chemical nature of the different residues enters only via a relatively small number of parameters, of the order of 1000. When tested on various set of decoys our potentials performs clearly better than all the others in discriminating the native state and slightly better as regarding correlation coefficients and Z-score. The results are particularly good for the recently released models of CASP9. We term our statistical energy function BACH (Bayesian Analysis Conformation Hunt) since it is simple, elegant and complete.

4.2 Methods

In the following, we will describe in detail the development of BACH, and we will present the procedure we have done in order to check the robustness of its parameters calculated over different training sets. We will also describe the various decoy sets that we have used for assessing the quality of BACH.

4.2.1 Development of BACH

The BACH energy function is based on two terms

$$E_{\mathrm{Bach}} = pE_{\mathrm{pair}} + E_{\mathrm{sol}}, \tag{4.1}$$

where E_{pair} and E_{sol} are statistical potentials learned from a training set of native PDB structures (see Section 4.2.2). The two terms take respectively into account the effective pairwise residue-residue interactions and the single residue solvation terms. p is a parameter that fixes the relative units of the two energy terms. It is chosen in such a way that the energy per residue has approximately the same standard deviation, over the training set. For all what follows, we use $p = 0.3$.

Pairwise Term

The pairwise statistical potential E_{pair} is based on classifying all residue pairs within a protein structure in five different structural classes, that we label by means of index x: the two residues may form a α-helical hydrogen bond / bridge ($x = 1$), or may form an anti-parallel β-bridge ($x = 2$), or may form a parallel β-bridge ($x = 3$), or may be in contact with each other through side chain atoms ($x = 4$), or may not realize any of the previous four conditions ($x = 5$). First, α- and β-bridges are assigned by using a slightly modified version of the DSSP algorithm [89] that employs a more stringent threshold (-1 Kcal/mol in place of the original -0.5 Kcal/mol, as already done in [127]) of the partial electrostatic energy used in DSSP to check the formation of each of the two hydrogen bonds that compose a bridge. A residue pair is then assigned to the side chain - side chain contact class ($x = 4$) if it has not already assigned to any of the hydrogen bonded classes ($x < 4$) and if an inter-residue pair of the side chain heavy atoms is found at a distance lower than 4.5Å. If none of the above conditions are verified, the pair of residues is assigned to the no-interaction class ($x = 5$). The pairwise statistical potential E_{pair} then requires five distinct 20×20 symmetric interaction matrices ϵ^x_{ab}, one for each of the classes defined above, where a and b vary among the 20 amino acid types. Overall $5 \times 210 = 1050$ independent parameters:

$$E_{\text{pair}} = \sum_{i<j} \epsilon^{x_{ij}}_{a_i a_j}, \tag{4.2}$$

where i, j are indexes labeling residue positions along the chain, a_i is the amino acid type of residue at position i, and x_{ij} is the structural class to which the residue pair at positions i and j is assigned.

The five interaction matrices ϵ^x_{ab} are determined from a training set of native protein structures (see Section 4.2.2) employing the ensemble of all residue pairs from the training set as the reference state [110]:

$$\epsilon^x_{ab} = -\ln\left[\frac{\frac{n^x_{ab}}{\sum_x n^x_{ab}}}{\frac{\sum_{ab} n^x_{ab}}{\sum_x \sum_{ab} n^x_{ab}}}\right], \tag{4.3}$$

where n^x_{ab} is the total number of residue pairs of type a and b found in the structural class

x within the training set. Residues that are not seen in the PDB structure, even if they are part of the sequence expressed by the crystallographer, are considered as disordered ones and included in the countings, so that any pair involving a disordered residue is classified in the 'no-interaction' class ($x = 5$).

Solvation Term

The solvation statistical potential E_{sol} is based on classifying all residues in two different environmental classes, either buried (b) or solvent exposed (s). The environmental class is defined based on the evaluation of the solvent accessible surface area (SASA) according to Connolly [**?**]: the surface that can be accessed by the center of a solvent probe sphere. Its calculation is performed by the SURF [128] tool of VMD graphic software [88]. The SASA is computed by SURF for all heavy atoms of the protein chain by trying to roll a probe sphere (representing a water molecule) on the surface of the set of spheres centered at heavy atom coordinates. We input to SURF the same value (1.8Å) for both the radii of all atom types and the radius of the probe sphere. The latter is higher than what employed in VMD (1.4Å) because we want to avoid considering internal cavities as exposing surface area to the solvent. The output of SURF is the number of triangle vertices, associated to each atom of the protein chain, that are used in the triangulated representation of the protein surface employed by VMD. By summing over all atoms of a given residue, we finally obtain the number of vertices t associated to that residue, which is proportional to its SASA. The area associated with each vertex is $\sim 0.15\mathring{A}^2$. The distributions of the values of t observed in the training set for alanine, valine and arginine residues are shown in Fig. 4.1, the observed behavior is typical of all residue types. The presence of a sharp peak at $t = 0$, well separated by a broader peak at larger values of t, allows a clearcut definition of residue environment as either buried ($t \leq t^*$) or solvent exposed ($t > t^*$), using the same threshold $t^* = 10$ for all residues.

The single residue statistical potential E_{sol} requires two separate sets of 20 parameters λ_a^e, for each of the environment classes defined above, where a varies among the 20 amino acid

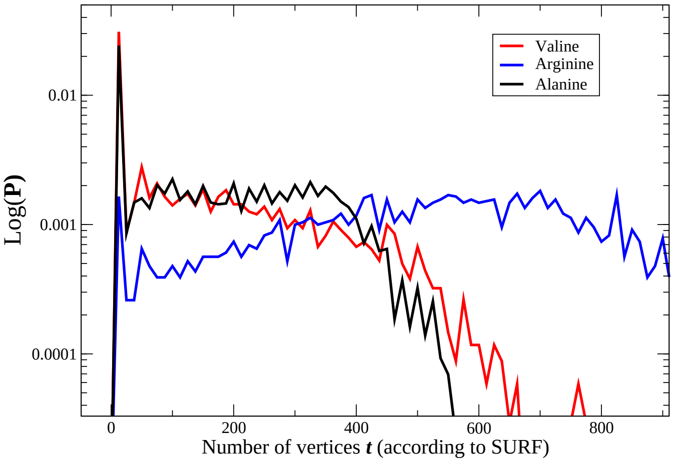

Figure 4.1: Logarithm of the probability distribution for the values of the number of vertices t observed, according to SURF [128], in the training set Top500 [95] for the alanine, valine and arginine residues.

types.

$$E_{\text{sol}} = \sum_i \lambda_{a_i}^{e_i}, \tag{4.4}$$

where the index i labels residue position along the chain, a_i is the amino acid type of residue at position i, and e_i is the environmental class to which the residue at position i is assigned.

The two parameter sets λ_a^e (for overall 40 parameters) are determined from the training set of native protein structures (see Section 4.2.2) employing the ensemble of all residues from the training set as the reference state [110]:

$$\lambda_a^e = -\ln\left[\frac{\frac{m_a^e}{\sum_e m_a^e}}{\frac{\sum_a m_a^e}{\sum_e \sum_a m_a^e}}\right], \tag{4.5}$$

where m_a^e is the total number of residues of type a found in the environment class e within the training set. Residues that are not seen in the PDB structure, even if they are part of the sequence expressed by the crystallographer, are included in the counting as solvent exposed ($e = s$).

4.2.2 Training set

The BACH parameters have been learned using the TOP500 database as the training set [95]. This set includes 500 non redundant single domain protein conformations, extracted from both monomeric and multimeric PDB protein structures, between 30-840 amino acids long, which have been solved with resolution better than $1.8\dot{A}$ by X-ray crystallography (no NMR). The structures in the set include disordered and not resolved regions. We count the contributions of these amino acids as non-interacting ones (see above). The BACH parameters learned over the TOP500 dataset are presented in the Appendix.

We have checked that the choice of the fold library does not affect the BACH parameters. In Fig. 4.2 the correlation between the BACH pairwise parameters calculated for 8000 structures of the CATH [91] and for the Top500 databases, is presented. Correlation is excellent especially for parameters corresponding to favourable interactions, that are by definition highly represented in the datasets. A similar correlation is obtained for the solvation parameters. These results are also consistent when learning the parameters on the PDBselect [94] library of folds.

We have also checked that the BACH energies, calculated for a set of structures using the parameters obtained with two different training sets have an excellent correlation. In Fig. 4.3 we show the correlation between the BACH energies for the structures in the CASP8 - T0397 decoy set, when the parameters are learned over the top500 and CATH databases.

4.2.3 Decoy sets

An ensemble of structures with the same primary sequence are defined as decoys. They are important because they provide a benchmark for checking the quality of different scoring functions. In this section we will explain the decoy sets that we have used (or generated) in order to further test BACH's performance.

CASP decoy sets

The performance of BACH has been assessed on selected decoy sets from CASP 8/9 [129]. We have chosen these decoys because recently [124] it has been argued that they are the

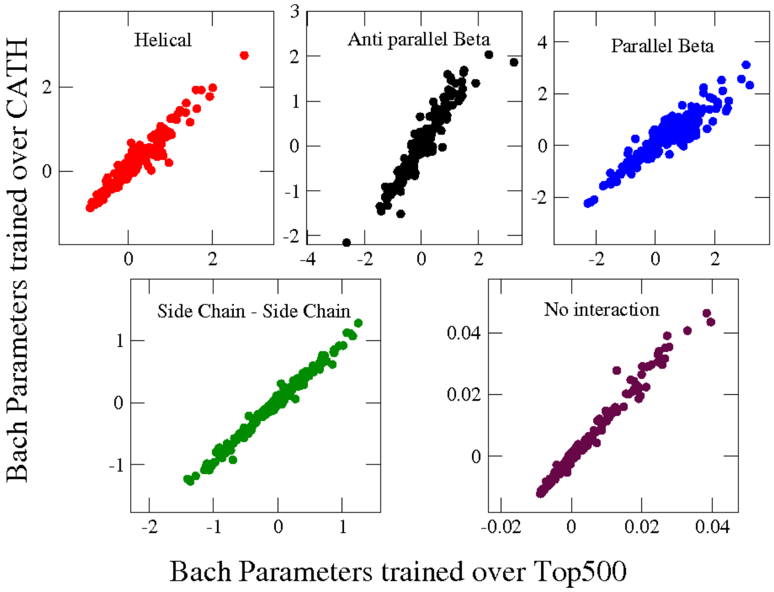

Figure 4.2: Correlation between the BACH pair-wise parameters learned with 8000 CATH structures and with the Top500 dataset. For each class of BACH pairwise interactions we plot in the *x-axis* the parameters learned over the top500 dataset, and in the *y-axis* the parameters learned over the CATH dataset.

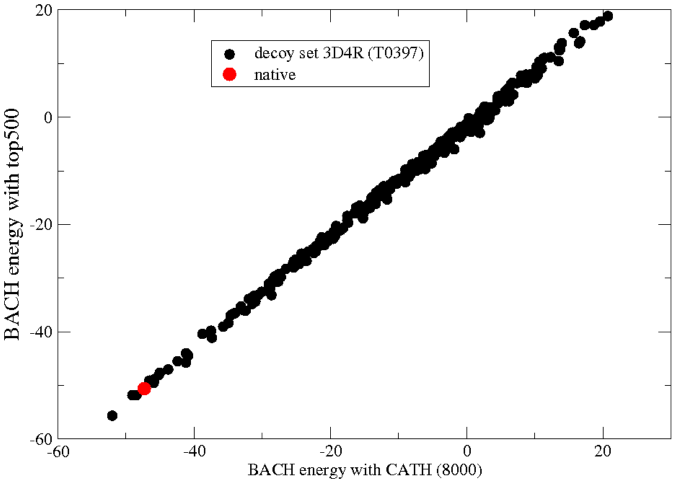

Figure 4.3: Correlation between the BACH energies for the structures in the T0397 (CASP) decoy set, when learning the parameters over two different training sets (CATH and top500 datasets; *x-axis*, and *y-axis*, respectively).

most difficult ones for discriminating the native structure in the set. A decoy set was used from CASP8/9 if the sequence of the crystalographic structure is the same as the predefined sequence given in the CASP competition, *i.e.* the decoy structures have the same sequence as the native structure. The list of the decoy sets used is presented in the Appendix. A total 33 decoys sets were selected. The structures in each decoy set were used if they had the same length and sequence as the native structure, and had all the side-chain and backbone atoms. If for any reason a structure could not be analyzed by one of the scoring functions with whom we compare the performance of BACH, the structure was taken out of the set.

Since the folded structure of proteins solved by NMR is given in the form of 20 different models we defined as native, the pdb model that had the lowest energy (for each potential). So, for two different scoring functions, the native structure could be a different model out of the 20 presented in the NMR pdb file.

Decoy set generated by Molecular Dynamics

We also generated a decoy set using molecular dynamics in combination with bias-exchange metadynamics [61] to produce realistic structures of the same protein (GB3, pdb code 2OED). Following a procedure similar to the one described in Chapter 3. Three different simulations were performed: *i)* a simple MD of 10 ns in explicit solvent at 330 K, *ii)* a simple MD of 10 ns in implicit solvent at 330 K, and *iii)* a bias exchange metadynamics of 20 ns in implicit solvent at 400 K, to enhance the conformational searching. This was done by using the GROMACS package [86], and employing the AMBER99-ILDN force field [130] for the protein. For the explicit solvent simulation the system has been solvated by 6524 water molecules in a 212 nm^3 cubic periodic box, using TIP3P water model [131]. For the implicit solvent the OBC model [132] was used. The particle-mesh Ewald method was used for long-range electrostatic with a short-range cutoff of 1.2 nm. All bond lengths were constrained to their equilibrium length with the LINCS algorithm [133]. The time step for the MD simulation has been set at 2.0 fs and the Nosé-Hoover thermostat [69] with a relaxation time of 1ps was used. The atomic coordinates and the energy were saved every 10 ps.

For the BE simulation, the PLUMED plugin [134] for Gromacs were used together with a similar setup as that of the simulation in Chapter 3. Four replicas, each biased over a different collective variable (described in detail in Section 2.4.1) have been chosen: *Coordina-*

tion Number to change the number of contacts inside the protein, *Alpharmsd*, *Parabetarmsd*, and *Antibetarmsd* which explore respectively the content in α-helix, parallel and anti-parallel β-sheets. One-dimensional hills with a height of 0.2 kJ/mol were introduced every 2 ps, and exchanges of the bias potentials between the replica were attempted every 20 ps. Analogously as in Chapter 3, the structures generated in this manner are highly different in secondary structure and tertiary contacts content.

4.3 Results

4.3.1 Discriminating the native conformation

The main goal of this work was to develop an efficient and simple knowledge based potential that is able to discriminate the native conformation from a set of different structures with the same primary sequence. We have calculated the BACH energy over various decoy sets, and checked its ability of assigning the lowest energy value to the native conformation. As a similarity measure between the decoys and the native, we use the GDT ('Global distance test') [125], which is defined as

$$GDT = \frac{F_{\leq 1} + F_{\leq 2} + F_{\leq 4} + F_{\leq 8}}{4}, \tag{4.6}$$

where $F_{\leq X}$ denotes the percent of residues under distance cutoff of $X\AA$ after an optimal superposition has been applied to the two proteins. This metric is considered to be a more accurate measurement than the RMSD metric [92], which is sensitive to outlier regions created by poor modelling of individual loop regions in a structure that is otherwise resonably accurate. GDT measurements are used as a major assessment criteria in the analysis of results in CASP.

In figure 4.4 we show the BACH energy versus the distance in GDT to the native structure for four decoy sets of CASP8/9. One can see that, in these four cases, BACH is able to discriminate the native structure from the other decoys, by assigning to it the lowest energy value. Moreover, in almost all cases, the closer a decoy structure is to the native state (as

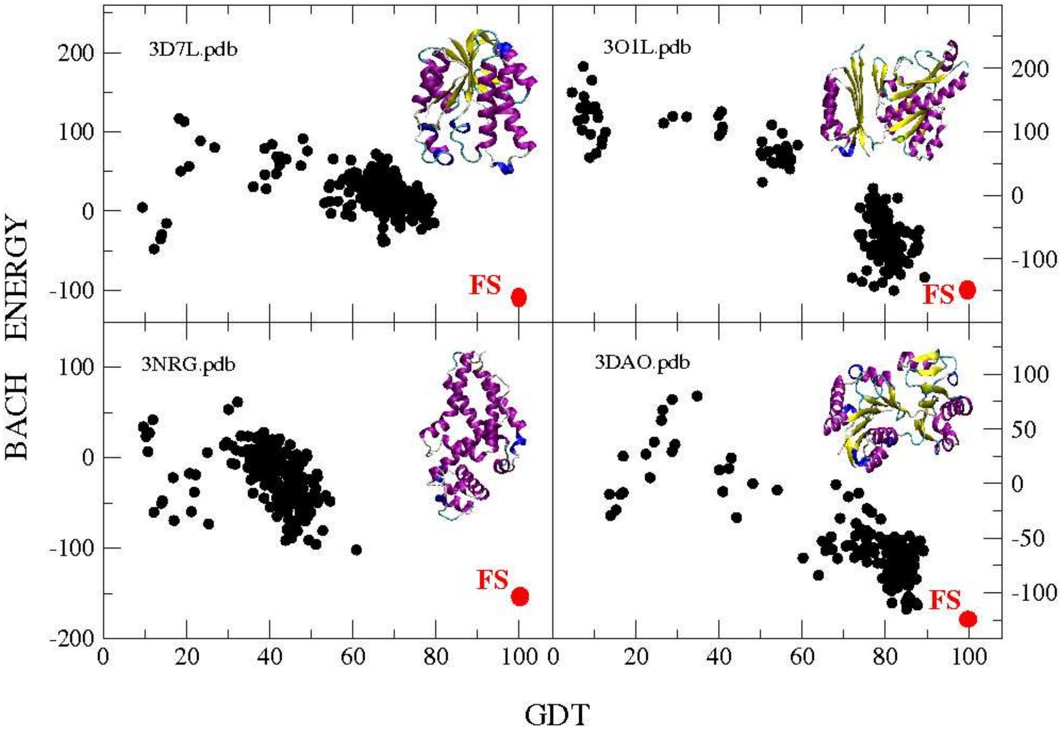

Figure 4.4: BACH energy as a function of the GDT with respect to the native structure, for four decoy sets of CASP8/9. FS: Folded state.

quantified by a large GDT), the lowest its BACH energy on average is. The BACH energy versus GDT distribution show for the four cases a typical funneled-like distribution.

We have also tested BACH's ability to discriminate the crystalographic conformation of proten GB3, on a set of thousands of structures of generated by molecular dynamics and bias exchange metadynamics. The details of the simulation are presented in Section 4.2.3. In figure 4.5 we show the BACH energy as a function of the GDT for the decoy structures generated with BE.

The structures generated in the BE simulation are very diverse in secondary and tertiary content, and span a wide range of GDT distances with respect to the native. They can be considered as analogous to the structures generated in Chapter 3. In this case, the native conformation ranked #5 out of 13565 structures. A running average of the BACH energy of all the structures is also shown in figure (red line in Fig. 4.5), suggesting also in this case the presence of a funnel on the average BACH energy for large GDT.

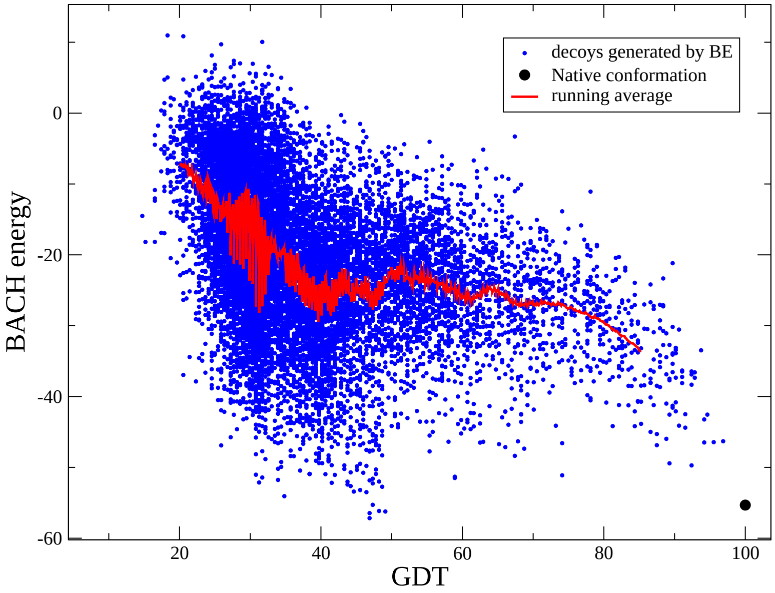

Figure 4.5: BACH as a function of the GDT to the native structure for different sets of structures generated by bias exchange metadynamics. Red line: running average.

We have checked that a similar performance is observed for several different decoy sets, including the sets, semfold [135], 4state [122], fisa [121], RosettAll [123], and the other decoy sets of CASP. A quantitative measure of this ability is demonstrated in Section 4.3.2. For the great majority of cases BACH is able to discriminate the native conformation. We find that BACH works equally well for all protein classes: all-α, all-β or mixed α/β proteins, and for proteins of different length (between 80-500 amino acids long). This is a first qualitative indication that BACH is rather powerful for protein structure discrimination. In the following, we will quantify this results and compare the performance of BACH with other knowledge based potentials.

4.3.2 Comparison with other knowledge based potentials

We compare the performance of BACH with other three knowledge based potentials: QMEAN6 [113], ROSETTA [102] and RF_CB_SRS_OD [119]. These potentials have been shown by Fiser *et al* [119], to perform at the top level in decoy discrimination. The comparison is performed on a subset of CASP8/9 targets (see Section 4.2.3), as it has been previously shown that these decoys are the most challenging [124].

Normalized Rank

The first quality assessment measurement that we use is the *normalized rank*, defined as the rank of the native structure divided by the total number of structures in the decoy set. For example, if the native structure has the lowest (resp. highest) possible energy among a set of 100 structures, its rank will be 0.01 (resp. 1). In figure 4.6-A we show the normalized rank for the decoy sets in CASP 8/9 (see Section 4.2.3) for BACH, QMEAN6, RF_CB_SRS_OD and ROSETTA.

Strikingly, for almost all the decoy sets BACH has the lowest rank, namely the best performance in discriminating the native structure from the decoys. BACH ranks the native within the best 5% for 28 decoy sets, whilst QMEAN6 does it for 23, and ROSETTA for 19 out of a total of 33 sets. Moreover, for 19 sets BACH ranks the native structure as the first, whilst QMEAN6 does this for 14 cases, RF_CB_SRS_OD for 13 and ROSETTA only for 3. These results show that BACH is able to discriminate accurately the native structure, assigning to it in many cases the lowest energy value.

Z-score

Another standard measure to characterize the performance of a scoring function is the Z-score. It is defined as

$$Z - score = \frac{|E_{nat} - \mu|}{\sigma}, \tag{4.7}$$

where E_{nat} is the energy of the native structure, μ is the mean and σ is the standard deviation of the energy values. The Z-score measures how big the gap is between the energy value of the native and the mean. The largest the Z-score, the better the potential is for discriminating the native structure. In figure 4.6-B, the Z-scores for the decoy sets in CASP 8/9 are shown for each potential. As one can see, BACH has, in almost all cases, the largest Z-score values. This could be important when one wants to predict which is the native structure in a set, *i.e.* if there are a lot of decoy structures that are close in energy to the native, this task would be much more difficult.

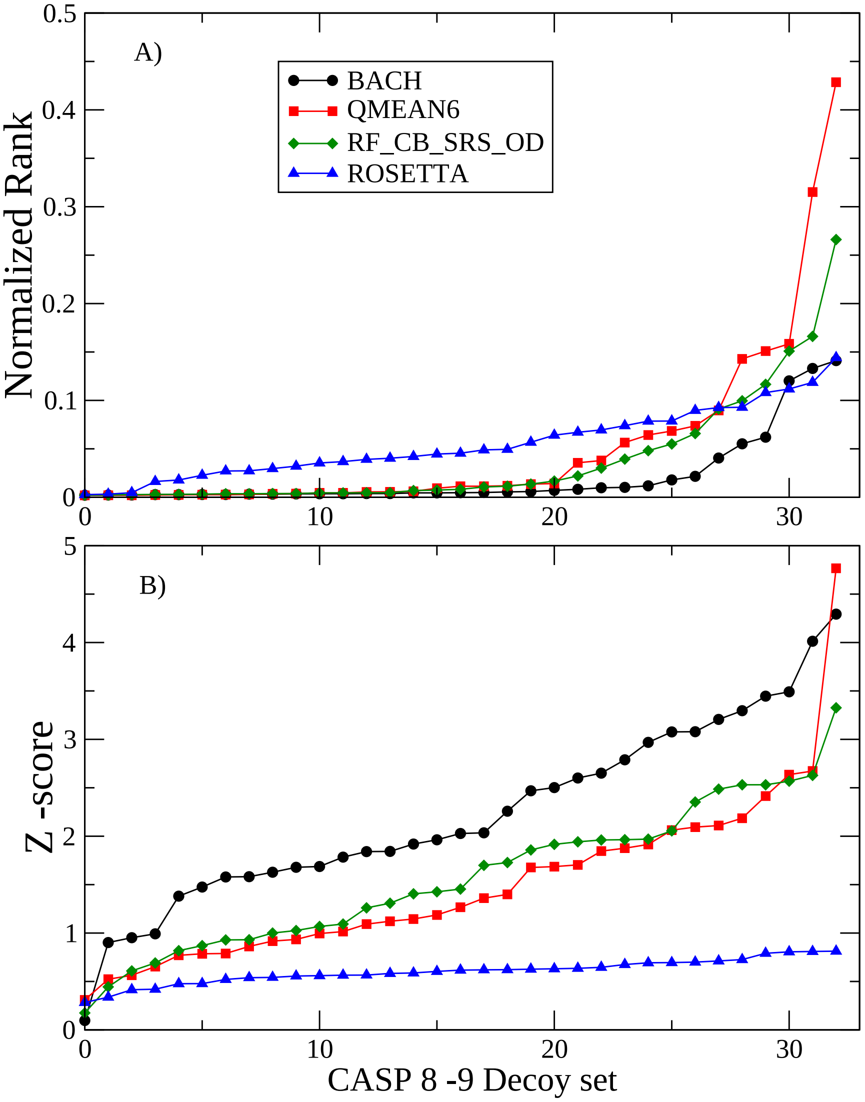

Figure 4.6: A) Normalized rank and B) Z-score sorted for the decoy sets in CASP 8/9, and calculated for the BACH, QMEAN6, RF_CB_SRS_OD and ROSETTA scoring functions.

It is to remark that even if BACH has a much more simple definition, and uses a lot less parameters still its performance is outstanding in ranking the native structure and with a large gap between its energy and the mean of the set. In both of the cases QMEAN6 has the second best performance, closely followed by RF_CB_SRS_OD, while ROSETTA performs more poorly.

Correlation Coefficient

Another standard quality performance measure is the Pearson correlation coefficient (ρ), a measure of the linear dependence between two variables x and y. It measures how big are the fluctuations of the points with respect to a linear fit of the data. It is defined as

$$\rho_{x,y} = \frac{\sum_i (x_i - \mu_x)(x_y - \mu_y)}{\sigma_x \sigma_y}, \tag{4.8}$$

where μ_x, μ_y and σ_x, σ_y are the mean and standard deviation over the x and y variables, respectively. This measure varies in the interval $[-1, 1]$. The closer it is to one, the more the data can be described with a linear fit. We calculated the Pearson coefficient between the score of each potential (y) and the GDT (x) with respect to the native structure, which has also been included in the computation. Its absolute value for the 33 decoy sets and for the four potentials we considered is shown in Fig. 4.7. For this specific quantity, QMEAN6 performs better than BACH, namely it produces on average a more linear dependence of the scoring function on GDT.

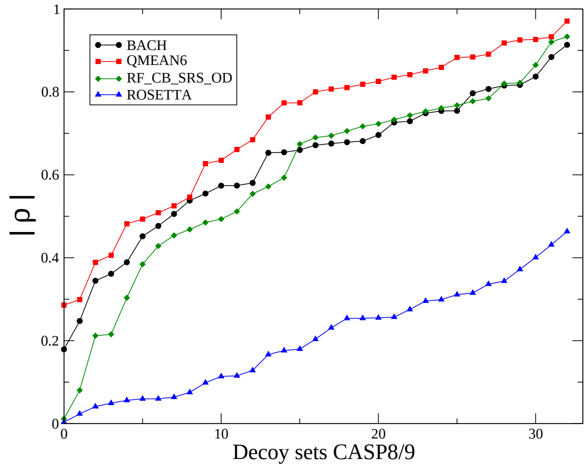

Figure 4.7: Absolute value of the Pearson correlation coefficient for BACH, QMEAN6, RF_CB_SRS_OD and ROSETTA, sorted for the decoys in CASP 8/9.

However, the highest Pearson coefficient is observed in the decoy sets in which the native state is poorly discriminated. This is shown in figure 4.8, where we plot the Pearson correlation coefficient versus the Z-score for BACH and QMEAN6 for the 33 decoy sets of CASP 8/9. In almost all the cases in which the Pearson coefficient is higher than 0.8 the Z-score is below 1.5, indicating that the folded state is rather poorly discriminated. Apparently, a scoring function producing a small standard deviation in the scoring versus GDT is normally less capable of distinguishing the native state from a set of structures. This could indicate that the profile of the energy score as a function of the GDT to the native, is not necessarily best fitted to a linear curve, and fluctuations cannot be avoided. This is confirmed from the tests performed on decoy sets generated by molecular dynamics (see below).

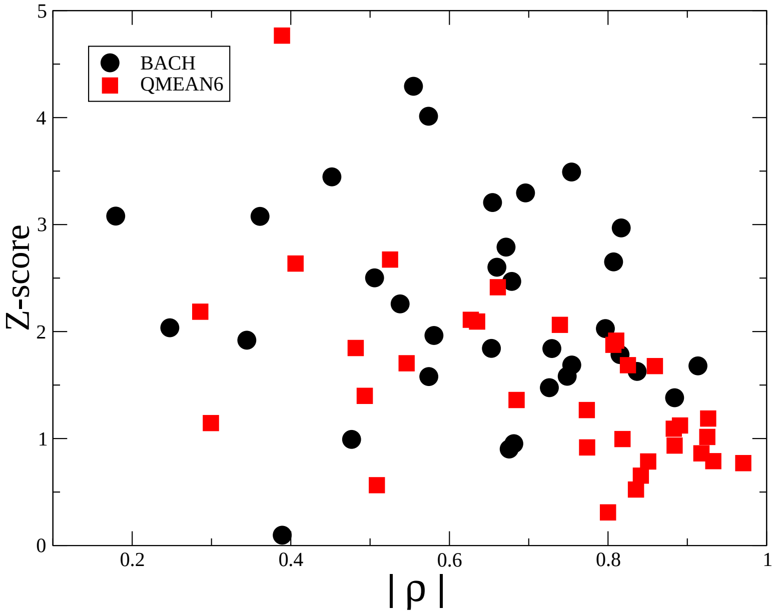

Figure 4.8: Absolute value of the Pearson correlation coefficient (ρ) versus the Z-score for BACH and QMEAN6 for the decoys in CASP 8/9.

Discriminating the most native-like conformation in a decoy set

One wonders how would BACH perform if it would be used in an iterative procedure to select low energy conformations without knowing the folded state. Normally, global optimization algorithms (for example genetic algorithms) schematically work in the following manner: i) first, an ensemble of structures is generated; ii) then, one selects by a scoring function a small subset of N low energy structures, iii) from this subset one creates a new ensemble of structures. This procedure is iterated until no new low-energy conformation can be found. As any KBP is unavoidably affected by fluctuations (see Section 4.3.4), this procedure is not statistically reliable if N=1, as a single small value can be the effect of a fluctuation. In other words, if one would only take the best structure, there would be high chances of missing a relevant "branch" of good structures due to fluctuations. In the Appendix, we show the value of the GDT of the closest structure to the native for each decoy set, we find that this value changes significantly from set to set ranging from 30 to 90 in GDT. If the native conformation is not present in the set, ideally the best potential should be able to provide the smallest difference in GDT between the first ranked structure in energy and the closest structure to the native.

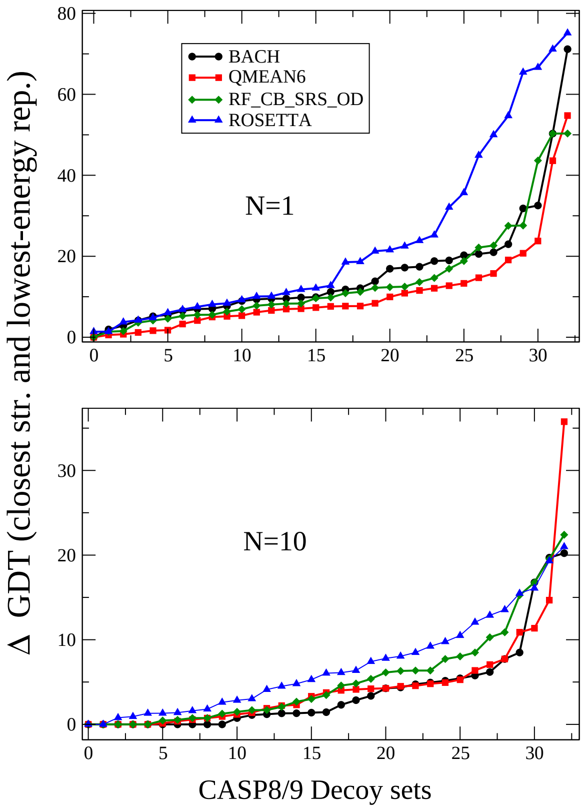

Figure 4.9: Difference in GDT between the closest structure to the native in the set and the structure ranked within A) N=1, and B) N=10 with highest GDT for BACH, QMEAN6, RF_CB_SRS_OD and ROSETTA, sorted for the decoy sets of CASP8/9.

In Fig. 4.9-A we show, for each potential, the difference in GDT of the lowest-energy structure (N=1) and the structure in each set which is closest to the folded state. In order to avoid effects of fluctuations, we take as representative of the 'lowest-energy structure' the decoy that is ranked within the first ten (N=10), according to each potential, and that has the highest GDT. In Fig. 4.9-B we show, for each potential, the difference in GDT between this representative (lowest-energy) structure and the structure in each set which is closest to the folded state. It is clear that while for N=1 QMEAN and RF_CB_SRS_OD perform better than BACH, the situation is reversed for N=10. The performance of ROSETTA marginally poorer performance in both cases.

In the Appendix, we present a table with the different quality assessments that measure the performance of BACH over the decoy sets of CASP8/9.

4.3.3 The performance of BACH on traditional decoy sets

We also benchmarked the performance of BACH over decoy sets that are considered the standard ones to test the scoring functions: semfold [135], 4state [122], fisa [121] and RosettAll [123]. In figure 4.10-B, we have selected, from these sets, only the decoys that have a monomeric single domain native structures with no ligands. In figure 4.10-A, we plot the normalized ranking of each potential over all the sets, also those including target structures that are polymeric, have ligands or are membrane proteins (at variance with the subset of CASP8/9 targets we considered above).

Even if these decoy sets include proteins in which the native state is not monomeric, or is a membrane protein, or is co-crystallized with large ligands or DNA, BACH still performs rather well in discriminating the folded state. We also see that all the scoring functions and, in particular ROSETTA, perform much better with these decoys than with the CASP decoy sets. However, the performance of BACH remains marginally better than the one of the competitors, also on these decoy sets. In particular, the normalized rank of BACH is smaller than 0.2 for all the monomeric targets.

In the Appendix, the Z-score, Pearson correlation coeffiecient and GDT of the closest structure ranked within the first ten (similarly as described above) are presented for BACH, RF_CB_SRS_OD and ROSETTA for all the decoy sets in these traditional decoy sets that

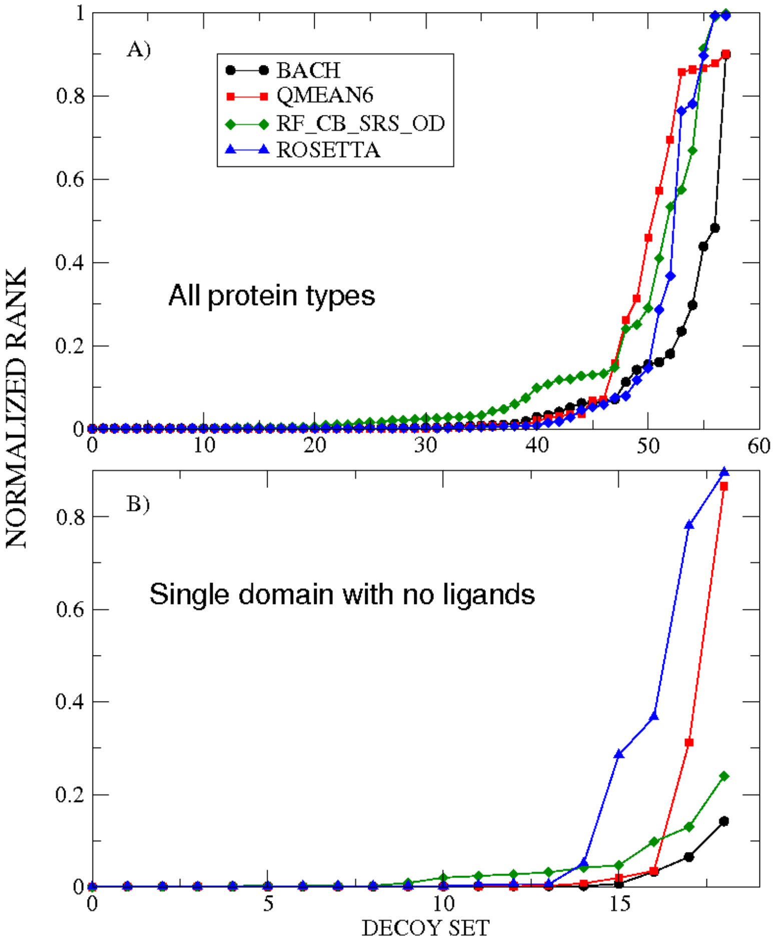

Figure 4.10: Sorted normalized rank for the native structure, for A) all the decoy sets, B) only single domain proteins in semfold, 4state,fisa, and RosettAll.

have a monomeric single domain native structures with no ligands. These quantities have also a similar behavior as those found in Section 4.3.2 for the decoy sets of CASP8/9.

4.3.4 Fluctuations are essential for scoring a structure

As discussed in Section 4.3.1, some of the structures generated by using bias exchange metadynamics (see Fig. 4.5) have a low BACH energy (lower than that of the crystallographic structure), even if they are different in tertiary structure from the folded state. One wonders, if these structures are signal of a flaw in the BACH potential, that is not able to recognize them as misfolded states. In order to investigate this, we selected five of these structures (the five of lowest BACH energy and GDT smaller than 50) and we performed, for each of these, 5ns of unbiased molecular dynamics in implicit solvent. For comparing with the native conformation, we also performed 10 ns of molecular dynamics simulations for the folded state in explicit solvent and in implicit solvent (see Section 4.2.3 for details).

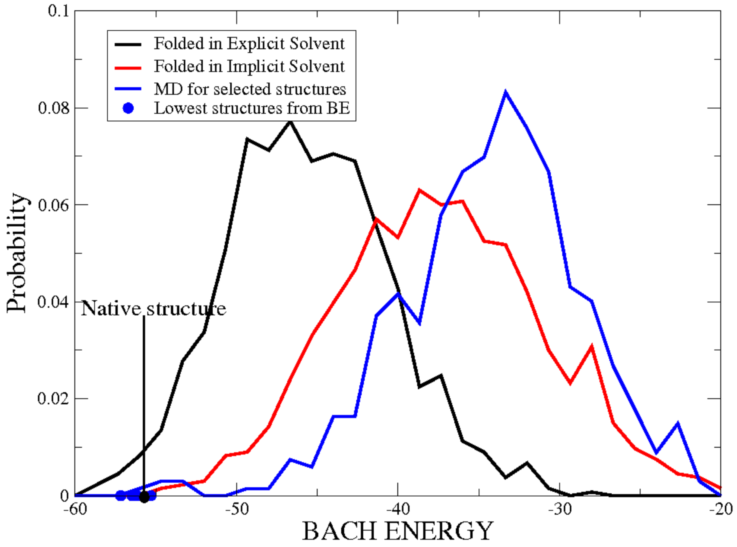

Figure 4.11: BACH energy distributions for the structures in the molecular dynimcs simulations of the folded state in explicit solvent (black line), the folded state in implicit solvent (red line) and the BE structures with low BACH energy (blue line). Black point: Native conformation; blue points: structures with lowest BACH energy obtained from BE.

In figure 4.11, we present the distributions of the BACH energies for the three sets of structures: MD of *i*) the folded state in explicit solvent, *ii*) the folded state in implicit solvent and *iii*) the structures generated by BE with lowest BACH energy. As shown in Figure, the MD simulation in explicit solvent starting from the folded structure has high fluctuations in BACH energy, of the order of 15. These fluctuations are due to the atomic motion at finite temperature, as occurs for any observable that depends on the coordinates. Nevertheless, these structures show the lowest BACH energies among all the sets, both in the average value and in the minimum. This demonstrates that the low BACH energy structures found with BE where just fluctuations in the dynamics, and do not correspond to meaningful low energy conformations. Remarkably, the PDB structure is not the structure of lowest BACH energy, possibly reflecting the presence of unavoidable (non systematic) errors in the experimental coordinates, which are also affected by fluctuations.

It is also interesting to observe that the MD simulation in implicit solvent produces structures of comparably high BACH energy (red line in Fig. 4.11). We have checked that this is due to a slightly wrong packing of the hydrophobic core.

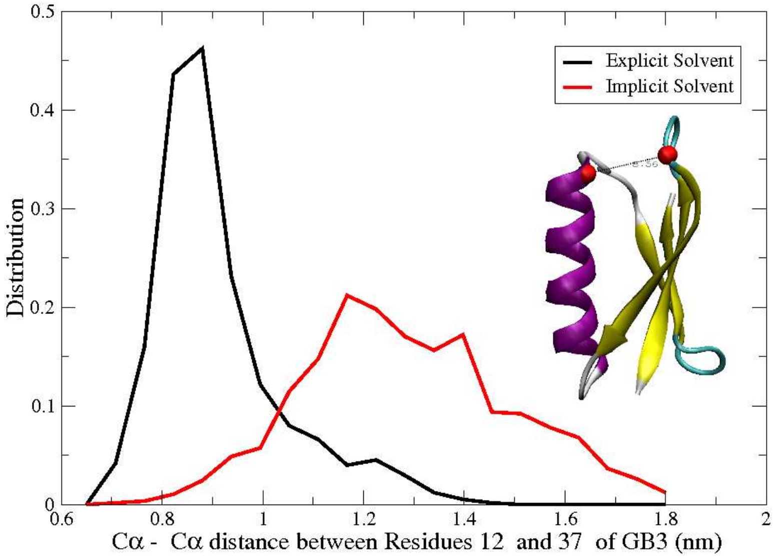

Figure 4.12: A) Distribution of the C_α distance between residues LEU 12 and ASN 37 of protein GB3, in explicit (black line) and implicit (red line) solvent MD simulations. *Subfigure:* Native conformation of protein GB3, red points: C_α atoms of residues LEU 12 and ASN 37.

This is illustrated in Fig. 4.12, where we show the distribution of the distance between the C_α atoms of two residues: LEU 12 and ASN 37 (red atoms in the protein shown in Fig. 4.12) for the structures in the explicit and implicit solvent simulations of the native conformation. Whilst the simulation in explicit solvent remains compact with a distance distribution close to the value of the crystallographic structure ($8.5\mathring{A}$), the structures in the implicit solvent simulation span a much large range, up to $18\mathring{A}$. This is a clear indication that there are structural differences between the sets, and that the implicit solvent model brings structural defects when simulating the folded state.

As we have seen (see Fig. 4.11) the BACH energy is able to capture these tiny structural differences. We checked if other scoring functions are also able to do the same. In Fig. 4.13-A

(B) we have plotted, respectively, the RF_CB_SRS_OD and ROSETTA energy distributions for the three MD ensembles.

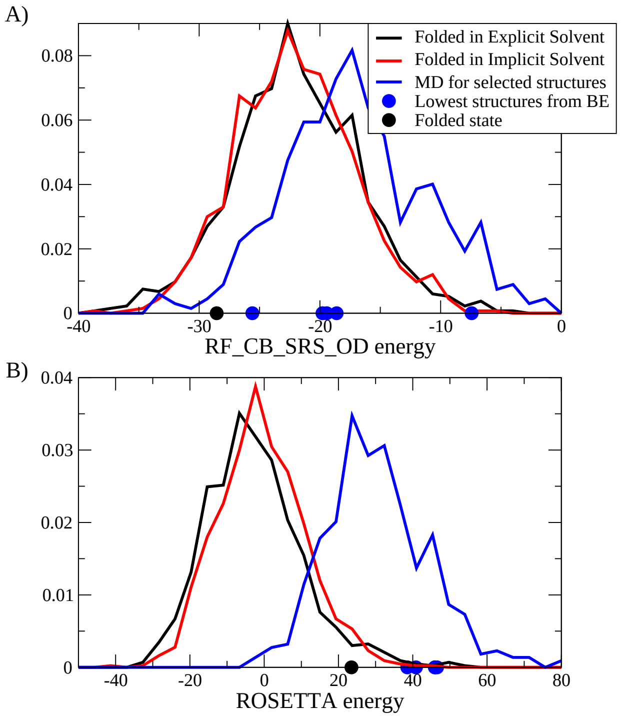

Figure 4.13: A) RF_CB_SRS_OD and B) ROSETTA energy distributions for sets of structures generated with molecular dynamics simulations of the folded state in explicit solvent (black line), the folded state in implicit solvent (red line) and BE structures with low BACH energy (blue line). Black point: Native conformation; blue points: structures with lowest BACH energy obtained from BE.

First it is to note that also these scoring functions produce large fluctuations when evaluated over structures simulated at finite temperature (and basically corresponding to the same structure). However, one can see that the distributions between the explicit and implicit solvent simulations of the folded state are practically indistinguishable for both scoring functions. The set from the MD simulations started from BE structures with lowest BACH energy are sharply discriminated by ROSETTA, and less clearly by RF_CB_SRS_OD. When comparing the initial starting points of the simulations (black and blue points in Fig. 4.11 and Fig. 4.13) to the final distributions, we find that BACH creates a larger gap between the native and the low energy structures selected from BE. ROSETTA also slightly increases the gap, while RF_CB_SRS_OD reduces it substantially. There is also a different behavior in how the native conformation is found within the distribution of the explicit solvent simulation: BACH has the native in the low energy tail quite far from the distribution center, RF_CB_SRS_OD has it closer to the center, while ROSETTA has it in the high energy tail. This mirrors the hierarchy that we find in the ranking and Z-score results found on different decoys sets (see Fig. 4.6 and Fig. 4.10).

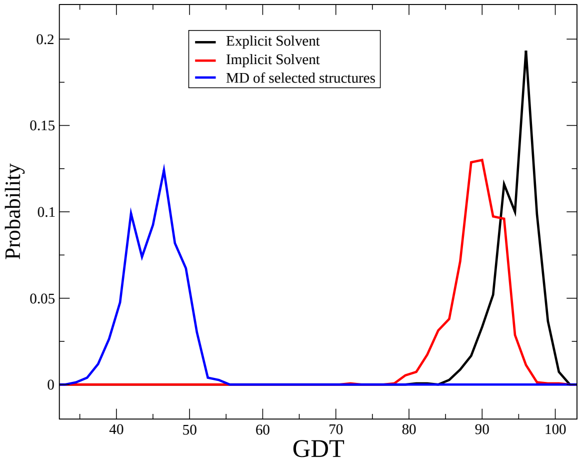

Figure 4.14: GDT distributions for sets of structures generated with molecular dynamics simulations of the folded state in explicit solvent (black line), the folded state in implicit solvent (red line) and BE structures with low BACH energy (blue line).

Lastly, in Fig.4.14 we show the GDT distributions for the structures of the different MD simulations. This figure shows that the explicit and implicit solvent ensembles are very similar when they are compared using a metric like the GDT. Nevertheless, there are still structural differences between the sets as those shown in Fig. 4.12. The results presented in Fig. 4.11 show that by computing the *probability distribution* of the BACH energy on a finite temperature run it is even possible to discriminate sets with small structural differences, like the ones described previously.

4.4 Discussion

We have developed a simple but robust knowledge based potential, called BACH, for discriminating the native structure within a decoy set. BACH's performance has been compared with other state-of-the-art potentials. When tested over a lot of different decoy sets it achieves the best results both on traditional decoy sets, and on the CASP8-9 decoy sets. Not only it is the best in assigning to the native structure the lowest energy value, but it also gives the largest gap between the energy of the native and the mean of the set. We find that BACH is not the potential with best correlation between the energy value and the GDT. However, we have also shown that a high linear correlation of the energy score with the GDT normally corresponds to a relatively poor capability of discriminating the folded state. The quality of BACH has also been assessed by analyzing tens of thousands of structures generated by molecular dynamics. We have found that BACH is able to discriminate the native structure ensemble, computed for a finite temperature run with an explicit solvent Hamiltonian, from other sets of structures generated with implicit solvent or enhanced sampling techniques. Moreover, we have found that the single BACH energy value is not very meaningful, as it is affected by thermal fluctuations. These fluctuations are present also in experimental data, as shown by the spread in BACH energy we observe in NMR models of the same protein. This is an indication that a more reliable quality measure of a structure is the *probability distribution* or the global minimum of the scoring function computed in a finite temperature run.

We for see that some symphonies in protein physics could be composed with the elegance and completeness of BACH. Due to BACH's simplicity, efficiency, and optimal performance it could be useful in a broad range of applications. Some examples are protein-protein interaction discrimination, docking, protein design, peptide-ligand design, generation of decoys. It can also be used in simulations for distinguishing the meta-stable conformations of the system, and for checking the correct native conformations.

Chapter 5

Conclusions

Proteins are the essential machinery on which cells are based. Discovering the tertiary structure of a protein, or the quaternary structure of the complexes they form, can provide important knowledge about how the protein perform its function, and information about the biological mechanism in which it is involved. This is not only important in biology but can be of great help in medicine by designing cures for pathologies. Much progress in understanding the role and function of different proteins has been made with experimental techniques, but nowadays, it is also possible to study proteins with the aid of computer simulations. Accurate computational models and simulation methodologies are currently used to study phenomena like protein folding [74, 16, 12], protein substrate binding [136], transitions of ions through channels [31], enzymatic reactions [32] and many others. Nevertheless, there are still some limitations in simulations. The major problem relies on the balance between the accuracy of the physical interactions (resolution level), and the computational cost of exploring vastly the system's conformational space. In this thesis, we have tried to address these issues in two manners: *i*) by showing that by using a suitable enhanced sampling technique one can explore exhaustively the conformational space of a peptide, and *ii*) by developing an accurate and efficient knowledge based potential for fold discrimination.

We have used bias exchange metadynamics [61], a powerful enhanced sampling technique, to explore through molecular dynamics the conformational space of a 60 amino acids polyvaline chain. With a relatively low-computational cost, we have generated a database of over 30,000 structures with high secondary content and small radius of gyration. The structures in the set correspond to metastable states of the system (local energy minima), and we have checked that they are stable even if the homopolymer chain is formed by another amino acid. We checked that these structures resemble real proteins by measuring the quality of several structural descriptors (Ramachandran values, G-factor [90], H-bond energies, etc.). We have found that the structures explored in this single simulation, can reproduce the great majority of known folds in the PDB. For CATH structures of length between 55 and 65 amino acids, the average coverage (aliened residues) is 75% and the average RMSD of the aligned residues is only 2.8 Å. At the end of the simulation, the fraction of CATH structures that we find in our dataset is 86%. The fact that a single peptide, which is biased only through secondary structure conformations, can reproduce all the existing folds, confirms that the observed protein topologies are selected based on geometry and not on the chemistry of the aminoacid sequence. However, we discover that the folds observed in nature represent only a *tiny* fraction of all the possible structures that a polypeptide can take, this fraction cannot be reproduced by choosing random structures from the database. Indeed, we find that *natural* folds are characterized by a small contact order, namely short loops in the bundle. This is consistent with the observation that parallel β-sheets are found less frequently in the PDB than in the dataset generated by us. One could argue that, due to a higher beta content, large contact order structures could have a higher tendency to aggregate and are therefor avoided by evolution. Another explanation of this effect can be that the contact order is known to correlate with the folding time of two-state globular proteins [63]. Thus, evolution might have selected the folds under the guidance of a simple principle: reducing the entanglement in the bundle formed by the protein in its folded state. Bundles with shorter loops might be preferable, as they are explored more easily starting from a random coil. Possibly, nature has been able to select low contact order structures by using a repertoire of 20 types of amino acids to select and design the sequences which minimizes the free energy of a desired (low CO) structure against other competing structures.

The availability of the new structures generated by us, calls for an application in an important challenge in protein physics: the design of the sequence that stabilizes a desired fold. Motivated by this, in the second part of the thesis, we have developed a simple but robust statistical potential achieving an unprecedented capability of recognizing the folded state in a decoy set, namely set of alternative conformations. We term our energy function BACH: Bayesian Analysis Conformation Hunt, since it is developed using Bayasian inference. Like other potentials, BACH aims at reproducing the average propensity for pair residues of forming contacts, or forming secondary structure elements, or the propensity of a residue to be exposed to the solvent. However, it depends only on ~ 1000 parameters, making its evaluation extremely efficient and its definition robust. The parameters are learned using Bayesian analysis on a relatively small set of experimental folded structures. We compared the performance of BACH with other knowledge based potentials: QMEAN6 [113], ROSETTA [102] and RF_CB_SRS_OD [119], which have been shown to perform extremely well in decoy discrimination [119]. For the decoy sets of CASP8/9, BACH ranks the native within the best 5% for 28 decoy sets, whilst QMEAN6 does it for 23, and ROSETTA for 19 out of a total of 33 sets. When tested over several different decoy sets, not only BACH is the best in assigning to the native structure the lowest energy value, but it also finds the largest gap between the energy of the native and the mean of the set (highest Z-score). If we measure the Pearson correlation coefficient, QMEAN6 performs marginally better than BACH, namely it produces on average a more linear dependence of the scoring function on the GDT similarity measure. However, the highest Pearson coefficient is observed in the decoy sets in which the native state is poorly discriminated. The quality of BACH was also assessed by analyzing structures of protein GB3 generated by an all-atom molecular dynamics simulation biased by bias exchange metadynamics. Also in this case, BACH is able to discriminate the native structure in a very large decoy set. However, we find that the BACH energy evaluated on a single configuration is not very meaningful, as in a finite temperature molecular dynamics simulation the structures are affected by significant fluctuations. Structures almost identical to the folded state can have by chance a BACH energy higher than that of a completely different structure. These fluctuations are not artifacts of the simulation, as fluctuations of similar amplitude are also present among different models of the same protein obtained by NMR. Based on these results, we propose that a more reliable quality measure of a structure is the *probability distribution* of the scoring function (for example BACH) computed in a finite temperature run, rather than a single value. Because of its accuracy and computational efficiency, BACH could be applied for protein design, protein structure prediction, in assessing the stability of mutant proteins, in studying protein-protein interaction and much more.

Appendix

CATH Structures

1aipC03	1ayjA00	1b9wA01	1ck7A02	1djxB01	1e8gA04	1e8rA00
1ektA00	1el6A01	1extA02	1fbnA01	1fuqA03	1gaxA04	1ha8A00
1hicA00	1hw7A02	1jeqA05	1jroA03	1k8bA00	1lm8V01	1mpgA03
1poiA02	1tvkB03	1uglA00	1x9bA00	1xrsB01	1zwvA00	2bm0A03
2hbaA00	2jrrA01	2oyoA01	1c0mA02	1deeG00	1ptqA00	1gjzA00
1nh2D01	1amlA00	1atxA00	1bazA00	1bbgA00	1bgwA01	1bhpA00
1biaA03	1bzkA00	1c55A00	1ck7A03	1cvuA01	1e0eA00	1e0gA00
1e3oC02	1e8pA00	1ed7A00	1ehsA00	1ekeB02	1ep3B03	1eptA00
1erdA00	1f5tA02	1fbrA01	1fd3A00	1fs1A00	1g29102	1go3F02
1gp8A00	1h1jS00	1h59B00	1h5wB03	1h6wA01	1inpA01	1ji8A01
1k1vA00	1lkoA02	1m1eB01	1mbmB03	1ncsA00	1nd9A00	1olgA00
1pnkA02	1qhkA00	1qo0D02	1twfI01	1uhaA02	1vpuA00	1w4eA00
1y1bA00	2cxnA03	2hjqA01	2otkE00	5reqB03	1a5tA02	1aapA00
1aipE03	1au7A02	1b4aA01	1ci3M02	1d2dA00	1dd9A03	1dxsA00
1e4eA02	1efaA01	1gcyA02	1h3nA03	1h9eA00	1hywA00	1k3rA02
1ly2A02	1pg5B02	1qxfA00	1qypA00	1rk6A03	1sqgA03	1t50A00
1tkeA03	1u94A02	1vq0A02	1vq8W02	2bayE00	2gycX00	2j8gA03
1a76A02	1aiwA00	1b04A03	1b3qA04	1bl0A02	1bunB00	1bxyA00
1cseI00	1dkgA02	1dqaD01	1dtdB00	1eakA01	1eejA01	1eh9A02
1ex7A01	1f94A00	1fjrA01	1g19A02	1gccA00	1go3F01	1hz6B00
1i2tA00	1i9gA01	1inlC02	1j7mA00	1ji8A02	1jlcB04	1kfwA02
1koyA00	1kvdA00	1kxpD05	1on2A02	1pceA00	1r69A00	1rq6A00
1syxB00	1tkeA01	1ucsA00	1umqA00	1wqjI00	1xjhA00	1yuaA01
2cc6A00	2ecsA00	2fj8A01	2gpfA01	2hg7A00	2jr6A01	2jrmA00
2nn4A00	3bulA04	1b3aA00	1b69A00	1b8tA01	1bbyA00	1bcoA02
1brwA03	1c7sA04	1c7vA00	1ccwB02	1d2nA02	1dvpA02	1ehiA03
1elvA03	1hp8A00	1j2zA02	1jajA02	1k0rA04	1khcA02	1ky9B04
1l6hA00	1lq7A00	1mhyG01	1mntA00	1mpxA02	1musA01	1qzpA00
1rrzA00	1tvfA02	1uxyA02	1v0eA01	1xakA00	1xccA02	2derA02
2hjjA00	2jn4A00	2nllA00	2proC01	1nh8A03	1ib8A02	1c4qA00
1cqqA01	1i6uA01	1qyrA02	1a79A02	1a9xA03	1apjA00	1au7A01
1b22A00	1b6rA01	1cfaA00	1cktA00	1dzfA02	1e8oA00	1eiaA02
1eijA00	1f3mA00	1fjgR00	1g8lA04	1gh9A00	1hc7B03	1ic8A02
1iq8A02	1iq8A03	1je3A01	1jw2A00	1kgqA01	1ku1A01	1mmsB00
1o54A01	1os6A00	1pgxA00	1pkpA01	1qsaA02	1r8eA02	1tolA02
1uj8A00	1vajA02	1vqqA03	1wj2A00	1zjaA02	2g2uB01	2jovA01
2nocA01	1aw0A00	1cidA02	1fx0A01	2px6A02	1dj7B00	2a3dA00
1a62A02	1axnA01	1b0xA00	1b24A02	1cpyA02	1fjgM01	

Table 1: List of the names for CATH folds used in Chapter 3.

BACH parameters

Residue	CYS	PHE	LEU	TRP	VAL	ILE	MET	HIS	TYR	ALA	GLY	PRO	ASN	THR	SER	ARG	GLN	ASP	LYS	GLU
CYS	-0.211	-0.137	-0.361	0.9653	-0.008	-0.474	-0.161	0.4756	-0.021	-0.352	0.8958	1.4693	0.2773	0.8283	0.5880	0.5509	-0.022	0.0953	0.0152	-0.240
PHE		-0.212	-0.353	-0.215	-0.014	-0.094	-0.499	0.0642	-0.096	-0.489	0.5817	1.3664	0.4203	0.3099	0.6839	-0.175	-0.248	0.5362	-0.221	-0.215
LEU			-0.785	-0.377	-0.374	-0.653	-0.853	-0.005	-0.261	-0.690	0.3836	0.8334	0.0894	0.0590	-0.030	-0.494	-0.507	0.1905	-0.232	-0.390
TRP				-0.517	-0.211	-0.344	-0.472	0.1454	-0.120	-0.456	0.7918	1.1273	0.1465	0.0134	0.2743	-0.201	-0.185	0.1895	-0.182	-0.479
VAL					-0.084	-0.133	-0.474	-0.052	0.1423	-0.579	0.7519	0.7511	0.4796	0.2198	0.1857	-0.138	-0.033	0.4863	-0.110	-0.110
ILE						-0.524	-0.576	0.0518	-0.134	-0.599	0.3400	0.8024	0.3714	0.2255	0.4348	-0.214	-0.506	0.1320	-0.135	-0.378
MET							-0.677	-0.120	-0.323	-0.637	0.4061	1.0009	0.0851	-0.166	0.2274	-0.258	-0.250	-0.081	-0.153	-0.279
HIS								-0.231	0.2498	-0.185	0.8595	1.9361	0.7015	0.2416	0.2972	0.0921	-0.080	0.8015	0.0089	-0.185
TYR									-0.155	-0.219	1.0360	1.2343	0.2285	0.3337	0.2275	-0.089	-0.318	0.4476	0.1886	0.1291
ALA										-0.785	0.4782	0.9565	0.0928	-0.042	0.0633	-0.627	-0.736	-0.033	-0.639	-0.705
GLY											1.6287	2.7539	1.1453	0.7776	1.0341	0.7897	0.5566	1.1858	0.6288	0.7273
PRO												6.9101	1.7354	1.3784	2.0046	0.9204	0.8964	1.6140	1.0029	1.0616
ASN													0.0763	0.4003	0.7676	0.0760	-0.109	0.5017	-0.014	0.0371
THR														0.4295	0.6391	-0.009	0.0413	0.5083	0.0844	0.1056
SER															0.7661	0.1889	0.1501	0.6987	0.1053	-0.042
ARG																-0.221	-0.680	-0.278	-0.229	-0.897
GLN																	-0.726	-0.139	-0.642	-0.591
ASP																		0.4482	-0.110	-0.127
LYS																			-0.034	-0.888
GLU																				-0.580

Table 2: Symetric matrix of the α-helical hydrogen bond/bridge ($x = 1$) BACH parameters.

Residue	CYS	PHE	LEU	TRP	VAL	ILE	MET	HIS	TYR	ALA	GLY	PRO	ASN	THR	SER	ARG	GLN	ASP	LYS	GLU
CYS	-2.633	-1.091	-0.837	-0.723	-0.902	-0.644	-0.229	-0.483	-1.268	-0.079	-0.190	0.3405	0.7711	-0.081	0.2064	0.4028	0.7460	0.6755	-0.349	0.6359
PHE		-1.276	-0.754	-0.726	-1.218	-1.189	-0.745	-0.274	-0.979	-0.336	-0.204	0.3117	0.7162	-0.224	-0.012	-0.039	-0.003	0.8264	0.1608	0.0259
LEU			-0.553	-0.601	-0.883	-0.804	-0.325	0.1947	-0.684	0.1497	0.3668	0.8623	0.7910	0.0848	0.1773	0.3579	0.2744	1.0540	0.3536	0.6163
TRP				-0.770	-0.684	-0.688	-0.263	-0.528	-1.063	-0.210	0.0904	0.3915	-0.037	0.3852	-0.343	-0.320	-0.515	0.5115	-0.593	0.2194
VAL					-1.416	-1.457	-0.642	-0.474	-1.074	-0.501	-0.223	0.9063	0.5019	-0.657	-0.090	-0.331	-0.283	0.7719	-0.307	-0.118
ILE						-1.189	-0.582	-0.403	-1.057	-0.525	0.1375	1.3264	0.6426	-0.308	-0.000	-0.244	-0.195	0.9002	0.1215	0.2034
MET							-0.653	-0.128	-0.347	0.1584	0.5383	0.4699	0.4295	-0.223	0.1847	0.3056	0.7004	1.3286	0.0033	0.2943
HIS								-0.785	-0.437	0.2665	-0.040	1.0949	0.4664	-0.398	0.1727	0.1553	-0.228	0.1651	0.0212	0.1452
TYR									-0.967	-0.225	0.0440	0.0703	0.0477	-0.597	-0.303	-0.484	-0.345	0.4592	-0.548	-0.280
ALA										0.4967	0.6568	1.4868	1.3718	0.0511	0.8864	0.6333	0.8410	1.9158	0.8113	1.0997
GLY											0.6493	1.5072	1.0381	0.4770	0.7358	0.8210	0.8426	1.1395	1.4615	1.4415
PRO												2.3801	1.0278	0.7265	1.3710	1.2052	0.9242	3.2577	1.0527	1.4243
ASN													0.7611	0.0656	-0.037	0.3461	0.0548	1.3119	0.6349	0.7493
THR														-1.125	-0.392	-0.336	-0.257	0.3056	-0.493	-0.257
SER															-0.100	0.0519	0.0142	0.7673	0.1169	0.2799
ARG																0.3588	-0.024	0.0477	0.2391	-0.214
GLN																	0.1516	0.6191	0.2591	0.4422
ASP																		0.9541	0.0454	1.3823
LYS																			0.1706	-0.607

Table 3: Symetric matrix of the anti-parallel β-bridge ($x = 2$) BACH parameters.

Residue	CYS	PHE	LEU	TRP	VAL	ILE	MET	HIS	TYR	ALA	GLY	PRO	ASN	THR	SER	ARG	GLN	ASP	LYS	GLU
CYS	-0.564	-0.921	-0.237	-0.927	-1.298	-1.364	-0.454	0.9496	-0.240	0.0034	1.0821	5.3445	5.3052	0.4756	0.1353	-0.024	1.3681	1.1640	5.4659	0.0258
PHE		-1.514	-0.947	-0.146	-1.499	-1.513	-0.784	-0.100	-0.949	0.0688	-0.217	2.1124	0.4826	0.0107	0.0092	1.3656	0.5496	1.4484	1.4478	0.8021
LEU			-1.041	-0.217	-1.732	-1.763	-0.916	-0.086	-0.648	-0.469	0.3079	2.4334	1.2635	-0.231	0.6303	0.9029	0.3211	1.0475	0.8336	0.7235
TRP				-0.618	-0.861	-1.116	-0.468	1.1050	-0.185	0.0902	0.7124	1.8245	0.4506	-0.060	0.4607	1.7481	0.0009	2.0498	0.6478	1.9401
VAL					-2.277	-2.061	-1.208	-0.586	-1.160	-0.843	-0.085	1.2406	0.3485	-0.749	-0.005	-0.110	0.2458	0.3643	0.5644	0.4226
ILE						-2.167	-0.807	-0.611	-1.167	-0.774	0.1129	1.2553	0.6585	-0.431	0.3775	0.0459	-0.089	0.6750	0.8067	-0.060
MET							-0.436	0.0235	-0.908	0.8546	0.4101	2.1035	0.2714	0.1913	0.1707	-0.090	0.6293	0.7466	1.3673	0.5715
HIS								-0.674	-0.421	1.0427	0.6711	2.2765	0.3953	-0.470	-0.149	0.6148	0.2110	0.8824	0.1412	0.4571
TYR									-0.603	-0.105	0.1007	1.2870	0.3195	0.4023	0.0774	0.6768	0.0476	0.2450	0.4479	0.5649
ALA										0.3610	0.5857	1.9463	1.1954	0.3400	0.7653	0.8676	0.6958	1.2257	1.0553	1.2518
GLY											1.5273	1.8416	1.1493	0.7126	1.1678	0.9041	0.5748	1.3096	1.3904	1.4245
PRO												5.9978	2.9026	1.5308	2.4786	2.2327	1.6263	3.1866	3.0611	6.8339
ASN													-0.303	-0.298	0.7203	1.0482	0.9000	1.0466	1.2177	2.3829
THR														0.0562	-0.083	-0.237	0.5388	0.7174	-0.002	0.7310
SER															0.0209	0.5294	1.0176	0.3926	1.2890	0.9244
ARG																1.1350	0.5369	1.0541	1.6150	0.9347
GLN																	0.7736	1.0741	0.6298	1.0251
ASP																		1.3685	0.7560	1.4447
LYS																			1.1604	0.0323
GLU																				1.6279

Table 4: Symetric matrix of the parallel β-bridge ($x = 3$) BACH parameters

Residue	CYS	PHE	LEU	TRP	VAL	ILE	MET	HIS	TYR	ALA	GLY	PRO	ASN	THR	SER	ARG	GLN	ASP	LYS	GLU
CYS	-2.277	-0.962	-0.456	-0.701	-0.409	-0.561	-0.549	-0.433	-0.678	0.3406	0.2723	-0.308	0.0511	0.0180	0.3085	0.1194	0.2484	0.5211	0.8494	0.6759
PHE		-1.362	-1.143	-1.279	-0.903	-1.099	-1.112	-0.523	-0.973	-0.241	0.1876	-0.390	-0.017	-0.194	0.0914	-0.228	-0.088	0.2454	0.1856	0.1552
LEU			-1.046	-1.023	-0.817	-1.031	-0.840	-0.140	-0.783	0.0439	0.6535	-0.022	0.3851	-0.046	0.4580	-0.047	0.1002	0.6746	0.4761	0.4500
TRP				-1.403	-0.875	-0.984	-1.103	-0.791	-1.069	-0.269	-0.099	-0.688	-0.325	-0.330	-0.054	-0.615	-0.511	-0.150	-0.189	-0.295
VAL					-0.731	-0.866	-0.639	-0.039	-0.529	0.0954	0.6518	0.0381	0.2675	0.0125	0.4440	0.1564	0.1985	0.6175	0.5462	0.5381
ILE						-1.068	-0.875	-0.091	-0.795	-0.058	0.4841	-0.084	0.3514	-0.091	0.4071	-0.067	0.0772	0.4470	0.4366	0.4017
MET							-0.901	-0.453	-0.921	0.0070	0.3841	-0.090	0.0894	0.0675	0.2391	0.0593	0.0786	0.5267	0.4271	0.3990
HIS								-0.906	-0.647	0.5387	0.3969	-0.134	-0.114	-0.108	0.0103	-0.203	-0.071	-0.365	0.3051	-0.187
TYR									-0.855	-0.089	0.0315	-0.653	-0.238	-0.167	-0.014	-0.591	-0.376	-0.187	-0.305	-0.219
ALA										0.8769	1.1266	0.5470	0.6923	0.6290	1.0166	0.7233	0.8765	0.9511	1.2496	1.0729
GLY											0.3092	0.5672	0.3900	0.3130	0.5543	0.3864	0.5461	0.4208	0.8880	0.8662
PRO												0.2738	0.0433	0.1811	0.3302	0.0364	0.0493	0.1955	0.4662	0.2091
ASN													-0.071	0.0624	0.2858	-0.080	-0.061	-0.047	0.1968	0.1432
THR														0.0794	0.3184	0.0773	0.0614	0.0509	0.3771	0.1304
SER															0.4837	0.2357	0.2623	0.0632	0.5830	0.2578
ARG																-0.006	-0.097	-0.607	0.6422	-0.548
GLN																	0.0915	0.1430	0.2170	0.2196
ASP																		0.4044	-0.338	0.5431
LYS																			1.1657	-0.377
GLU																				0.7590

Table 5: Symetric matrix of the side chain interaction ($x = 4$) BACH parameters.

Residue	CYS	PHE	LEU	TRP	VAL	ILE	MET	HIS	TYR	ALA	GLY	PRO	ASN	THR	SER	ARG	GLN	ASP	LYS	GLU
CYS	0.1183	0.0211	0.0079	0.0128	0.0074	0.0108	0.0093	0.0064	0.0128	-0.003	-0.003	0.0035	-0.001	-0.000	-0.003	-0.001	-0.003	-0.005	-0.007	-0.005
PHE		0.0382	0.0276	0.0328	0.0199	0.0266	0.0264	0.0085	0.0214	0.0038	-0.002	0.0050	-0.000	0.0025	-0.001	0.0030	0.0012	-0.003	-0.002	-0.001
LEU			0.0245	0.0227	0.0177	0.0248	0.0177	0.0018	0.0154	0.0002	-0.006	-0.000	-0.004	0.0005	-0.004	0.0007	-0.000	-0.006	-0.004	-0.004
TRP				0.0394	0.0181	0.0219	0.0257	0.0148	0.0244	0.0042	0.0007	0.0114	0.0045	0.0047	0.0005	0.0105	0.0085	0.0015	0.0028	0.0043
VAL					0.0168	0.0201	0.0124	0.0009	0.0097	-0.000	-0.006	-0.001	-0.003	0.0003	-0.004	-0.001	-0.002	-0.006	-0.005	-0.005
ILE						0.0271	0.0184	0.0015	0.0163	0.0018	-0.004	0.0002	-0.004	0.0012	-0.004	0.0010	-0.000	-0.004	-0.004	-0.003
MET							0.0192	0.0071	0.0195	0.0003	-0.004	0.0003	-0.001	-0.000	-0.002	-0.000	-0.001	-0.005	-0.004	-0.004
HIS								0.0190	0.0113	-0.005	-0.004	0.0007	0.0009	0.0015	-0.000	0.0025	0.0009	0.0048	-0.003	0.0025
TYR									0.0175	0.0014	-0.000	0.0106	0.0030	0.0021	0.0001	0.0100	0.0059	0.0021	0.0043	0.0029
ALA										-0.006	-0.008	-0.006	-0.006	-0.005	-0.008	-0.005	-0.006	-0.007	-0.008	-0.007
GLY											-0.004	-0.006	-0.004	-0.003	-0.005	-0.004	-0.005	-0.005	-0.007	-0.007
PRO												-0.004	-0.001	-0.002	-0.004	-0.001	-0.001	-0.003	-0.005	-0.003
ASN													0.0007	-0.000	-0.003	0.0007	0.0006	-0.000	-0.002	-0.002
THR														-0.000	-0.003	-0.000	-0.000	-0.001	-0.003	-0.001
SER															-0.005	-0.002	-0.003	-0.001	-0.005	-0.002
ARG																-0.000	0.0018	0.0103	-0.005	0.0099
GLN																	-0.000	-0.001	-0.001	-0.002
ASP																		-0.004	0.0048	-0.005
LYS																			-0.008	0.0068
GLU																				-0.006

Table 6: Symetric matrix of the no-interaction ($x = 5$) BACH parameters.

Residue	Exposed	Burried
CYS	0.1946	-0.556
PHE	0.1424	-0.444
LEU	0.2156	-0.597
TRP	-0.010	0.0445
VAL	0.2574	-0.670
ILE	0.2811	-0.709
MET	0.1643	-0.493
HIS	-0.102	0.6098
TYR	-0.058	0.2918
ALA	0.1454	-0.450
GLY	0.0469	-0.176
PRO	-0.105	0.6283
ASN	-0.124	0.8135
THR	-0.041	0.1959
SER	-0.048	0.2346
ARG	-0.185	1.9501
GLN	-0.152	1.1822
ASP	-0.154	1.2076
LYS	-0.204	3.1183
GLU	-0.181	1.8223

Table 7: Solvation BACH parameters

Table 8: **CASP decoy sets**: Decoy sets used in Chapter 4, PDB code (name in CASP).

3PNX (T0517)	3NRD (T0522)	3NRE (T0526)	2L0B (T0539)
2L0D (T0541)	2L3W (T0544)	2KRX (T0562)	3ON7 (T0563)
2KYW (T0569)	3NRQ (T0575)	2KY9 (T0579)	3NI8 (T0594)
3NKH (T0623)	3O1L (T0626)	3NUW (T0628)	3CYN (T0388)
3D4R (T0397)	3D6W (T0415)	3CZX (T0425)	3D3Y (T0427)
3DAI (T0432)	3D7L (T0433)	2K3I (T0437)	3DCP (T0440)
3DAO (T0445)	3DO6 (T0447)	3DMC (T0451)	2K5W (T0468)
2K49 (T0472)	3DLC (T0485)	2VWR (T0488)	3DLM (T0504)
3E03 (T0511)			

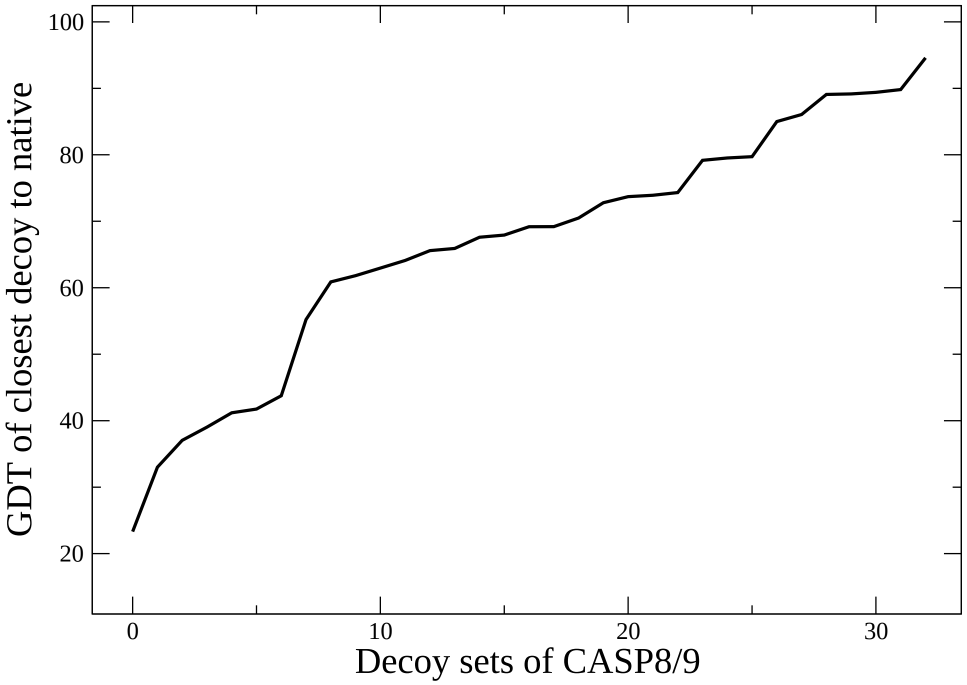

Figure 1: GDT of closest decoy structure to the native conformation sorted for the CASP8/9 sets.

Decoy	Normalized rank	GDT of closet	GDT of lowest rep.	Z-score	Pearson correlation	Fraction enrichment
T0517	0.0596	67.925	67.925	1.579254	0.574054	0.102564
T0522	0.0179	94.590	93.284	1.581800	0.748505	0.037037
T0526	0.0023	73.707	73.707	2.968533	0.816630	0.048785
T0539	0.0070	67.593	62.654	1.842799	0.653000	0.222222
T0541	0.0037	79.717	79.717	3.445862	0.451770	0.115385
T0544	0.0128	37.037	32.778	1.963298	0.580524	0.266667
T0562	0.0383	39.024	31.301	1.919497	0.344515	0.186047
T0563	0.0036	69.176	62.993	2.651130	0.807046	0.222222
T0569	0.0162	72.785	56.013	2.035285	0.247451	0.041667
T0575	0.0038	60.880	60.880	4.012596	0.573708	0.200000
T0579	0.0022	43.750	38.306	2.789333	0.671428	0.511111
T0594	0.0062	85.000	82.143	1.627530	0.836722	0.361702
T0623	0.0071	62.955	62.955	2.258477	0.537891	0.185185
T0626	0.0032	89.399	89.399	1.381472	0.883960	0.166667
T0628	0.0023	41.186	41.186	2.468571	0.678618	0.619048
T0388	0.0049	89.080	87.644	1.784483	0.815177	0.210526
T0397	0.0118	33.000	33.000	2.501684	0.505609	0.343750
T0415	0.0028	74.312	69.954	2.028250	0.796711	0.205882
T0425	0.0027	69.199	68.094	1.687010	0.754283	0.142857
T0427	0.0027	55.213	50.059	3.491227	0.753991	0.111111
T0432	0.0036	89.808	87.500	1.475357	0.726069	0.148148
T0433	0.0038	79.523	78.141	4.292572	0.554694	0.320000
T0437	0.1301	65.909	65.909	0.902215	0.675388	0.323529
T0440	0.0059	73.909	70.545	1.841253	0.729208	0.312500
T0445	0.0032	79.167	59.470	3.205908	0.654367	0.066667
T0447	0.0045	89.161	88.423	1.679495	0.913411	0.095231
T0451	0.0028	70.489	69.173	3.076912	0.361207	0.147059
T0468	0.1399	41.743	33.257	0.991921	0.476634	0.210526
T0472	0.0629	61.818	41.591	0.096790	0.389156	0.020430
T0485	0.0045	64.106	64.106	3.295974	0.695964	0.428571
T0488	0.1532	86.053	81.316	0.951732	0.681223	0.041667
T0504	0.0055	23.317	17.548	3.079031	0.179162	0.103760
T0511	0.0047	65.590	64.391	2.601460	0.659879	0.249700

Table 9: BACH's performance over the CASP8/9 decoy sets: decoy, Normalized rank, GDT of closest structure in the set, GDT of closest structure ranked within the best ten, Z-score, Pearson correlation coefficient, Fraction enrichment (the percentage of the top 10% lowest GDT structures that are found also in the top 10% best scoring ones).

Set	Decoy	BACH	RF_CB_SRS_OD	ROSETTA
semfold	1e68	3.612236	1.736280	1.245337
semfold	1khm	1.448607	3.298707	0.390097
semfold	1nkl	2.572452	3.253590	0.406699
semfold	1pgb	2.994818	2.465525	0.751449
4state	1r69	5.187321	2.530143	2.680363
4state	3icb	2.964012	1.763384	0.407856
4state	4pti	3.504422	2.089870	4.232069
RosettaAll	1aa2	6.333942	3.307967	2.111359
RosettaAll	1acf	6.225436	5.634524	1.772874
RosettaAll	1btb	5.705780	0.713018	1.649231
RosettaAll	1fbr	1.652924	1.132950	1.381048
RosettaAll	1gpt	2.030183	1.739186	1.382646
RosettaAll	1kte	5.977196	2.801269	2.591239
RosettaAll	1r69	3.828340	1.243447	1.984405
RosettaAll	2ezk	3.073327	2.378920	1.806570
RosettaAll	2ncm	2.857415	3.228812	1.892090
RosettaAll	5pti	1.046373	2.846518	1.643483
fisa	2cro	3.293369	1.185396	0.204062
fisa	4icb	6.656486	1.813432	1.293061

Table 10: **Z-score** for BACH, RF_CB_SRS_OD and ROSETTA calculated over standard decoy sets semfold, 4state, RosettaAll and fisa with single domain proteins.

Set	Decoy	BACH	RF_CB_SRS_OD	ROSETTA
semfold	1e68	0.275461	0.326351	0.007737
semfold	1khm	0.036879	0.079506	0.006546
semfold	1nkl	0.165764	0.228080	0.064933
semfold	1pgb	0.179924	0.188029	0.043189
4state	1r69	0.555645	0.622371	0.667629
4state	3icb	0.740209	0.759393	0.599443
4state	4pti	0.390731	0.487051	0.439148
RosettaAll	1aa2	0.231899	0.131486	0.025757
RosettaAll	1acf	0.175405	0.243301	0.012026
RosettaAll	1btb	0.195562	0.048517	0.030545
RosettaAll	1fbr	0.026616	0.052667	0.005185
RosettaAll	1gpt	0.100968	0.026099	0.030198
RosettaAll	1kte	0.261196	0.176756	0.028333
RosettaAll	1r69	0.237291	0.328273	0.041139
RosettaAll	2ezk	0.122587	0.198930	0.006688
RosettaAll	2ncm	0.090495	0.102257	0.059923
RosettaAll	5pti	0.090056	0.079878	0.009134
fisa	2cro	0.300151	0.360885	0.007331
fisa	4icb	0.362096	0.254032	0.234098

Table 11: **Pearson Correlation Coefficient** for BACH, RF_CB_SRS_OD and ROSETTA calculated over standard decoy sets semfold, 4state, RosettaAll and fisa with single domain proteins.

Set	Decoy	Closest	BACH	RF_CB_SRS_OD	ROSETTA
semfold	1e68	62.143	53.214	50.357	33.214
semfold	1khm	55.479	33.562	44.863	34.932
semfold	1nkl	52.885	26.923	38.462	25.962
semfold	1pgb	55.357	30.804	27.679	39.732
4state	1r69	93.254	87.698	82.937	93.254
4state	3icb	94.333	94.333	91.333	93.000
4state	4pti	83.190	71.552	81.897	79.310
RosettaAll	1aa2	30.952	28.571	20.238	17.619
RosettaAll	1acf	29.472	20.528	23.577	19.309
RosettaAll	1btb	32.303	25.843	23.596	25.843
RosettaAll	1fbr	31.989	18.817	24.731	31.989
RosettaAll	1gpt	52.660	36.702	33.511	32.447
RosettaAll	1kte	38.000	26.750	24.750	21.250
RosettaAll	1r69	69.262	53.689	67.213	44.262
RosettaAll	2ezk	48.118	30.376	37.366	29.301
RosettaAll	5pti	49.091	31.818	28.182	26.818
fisa	2cro	52.692	41.154	44.231	30.769
fisa	4icb	50.329	39.474	39.803	38.158

Table 12: GDT of closest structure in set. GDT of the closest structure ranked within the best ten for BACH, RF_CB_SRS_OD and ROSETTA over the standard decoy sets semfold, 4state, RosettaAll and fisa with single domain proteins.

Bibliography

[1] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75:333–366, 2006.

[2] H. Heise, W. Hoyer, S. Becker, O. C. Andronesi, D. Riedel, and M. Baldus. Molecular-level secondary structure, polymorphism, and dynamics of full-length alpha-synuclein fibrils studied by solid-state NMR. *Proc. Natl. Acad. Sci. USA.*, 102(44):15871–15876, 2005.

[3] M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia. Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445, 2011.

[4] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical research*, 33(12):889–897, 2000.

[5] W. L. Bragg. X-ray analysis of proteins. *Proceedings of the Physical Society of London Section A*, 65(395):963, 1952.

[6] J. Jeener, B. H. Meier, P. Bachmann, and R. R. Ernst. Investigation of exchange processes by 2-dimensional NMR-spectroscopy. *J. Comput. Phys.*, 71(11), 1979.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nuc. Ac. Res.*, 28:235–242, 2000.

[8] A. Fiser and A. Sali. MODELLER: Generation and refinement of homology-based protein structure models. In *Macromolecular Crystallography, PT D*, volume 374, pages 461+. 2003.

[9] J. Maupetit, P. Derreumaux, and P. Tuffery. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nuc. Ac. Res.*, 37:W498–W503, 2009.

[10] J. B. 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.

[11] 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.

[12] G. Jayachandran, V. Vishal, and V. S. Pande. Using massively parallel simulation and markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Comput. Phys.*, 124:164902, 2006.

[13] A. Rahman and F. H. Stilling. Molecular Dynamics study of liquid water. *J. Comput. Phys.*, 55(7):3336–&, 1971.

[14] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282(5389):740–744, 1998.

[15] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.*, 323:927– 937, 2002.

[16] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. Dror, M. P. Eastwood, J. Bank, J. M. Jumper, J. Salmon, Y. Shan, and W. Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(6002):341–346, 2010.

[17] J. M. Chandonia and S. E. Brenner. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins*, 58(1):166–179, 2005.

[18] T. Headgordon, F. H. Stillinger, M. H. Wright, and D. M. Gay. Poly(L-alanine) as a universal reference material for understanding protein energies and structures. *Proc. Natl. Acad. Sci. USA.*, 89(23):11513–11517, 1992.

[19] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. USA.*, 101(21):7960–7964, 2004.

[20] Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and J. Skolnick. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA.*, 103(8):2605–2610, 2006.

[21] C. Chothia and A. V. Finkelstein. The classification and origins of protein folding patterns. *Annual review of biochemistry*, 59:1007–1039, 1990.

[22] C. Chothia. Proteins - 1000 Families for the molecular biologist. *Nature*, 357(6379):543–544, 1992.

[23] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

[24] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[25] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324, 2002.

[26] A. Stein, R. Mosca, and P. Aloy. Three-dimensional modeling of protein interactions and complexes is going ‘omics. *Current opinion in structural biology*, 21(2):200–208, 2011.

[27] T. L. Shi, Y. X. Li, Y. D. Cai, and K. C. Chou. Computational methods for protein-protein interaction and their application. *Current Protein & Peptide Science*, 6(5):443–449, 2005.

[28] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998.

[29] M. L. Martins, S. C. Ferreira, Jr., and M. J. Vilela. Multiscale models for biological systems. *Current opinion in colloid & Interface Science*, 15(1-2):18–23, 2010.

[30] J. A. Mccammon, B. R. Gelin, and M. Karplus. Dynamics of Folded Proteins. *Nature*, 267(5612):585–590, 1977.

[31] M. Jensen, D. Borhani, K. Lindorff-Larsen, P. Maragakis, V. Jogini, M. Eastwood, R. Dror, and D. E. Shaw. Principles of conduction and hydrophobic gating in K(+) channels. *Proc. Natl. Acad. Sci. USA.*, 107(13):5833–5838, 2010.

[32] J. Villa and A. Warshel. Energetics and dynamics of enzymatic reactions. *J. Phys. Chem. B*, 105(33):7887–7907, 2001.

[33] S. Oldziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, and M. Makowski. Physics based protein-structure prediction using a hierarchical protocol based on the unres force field: assessment in two blind tests. *Proc. Natl. Acad. Sci. USA.*, 102:7547–7552, 2005.

[34] E. Paci, M. Vendruscolo, and M. Karplus. Validity of Go models: Comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.*, 83(6):3032–3038, 2002.

[35] V. Tozzini. Coarse-grained models for proteins. *Current opinion in structural Biology*, 15(2):144–150, 2005.

[36] T. Head-Gordon and S. Brown. Minimalist models for protein folding and design. *Current opinion in structural Biology*, 13(2):160–167, 2003.

[37] D Thirumalai and DK Klimov. Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Current opinion in structural Biology*, 9(2):197–207, 1999.

[38] R. DeVane, W. Shinoda, P. Moore, and M. L. Klein. Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids. *Journal of chemical Theory and Computation*, 5(8):2115–2124, 2009.

[39] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Oldziej, K. Wachucik, and H. A. J. Scheraga. Modification and optimization of the united-residue (unres) potential energy function for canonical simulations. i. temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B*, 111:260–285, 2007.

[40] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, 109(7):2469–2473, 2005.

[41] S. Kundu, D. C. Sorensen, and G. N. Phillips. Automatic domain decomposition of proteins by a Gaussian Network Model. *Proteins: Struct., Funct., Bioinf.*, 57(4):725–733, 2004.

[42] D. Reith, M. Putz, and F. Muller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24(13):1624–1636, 2003.

[43] I. Bahar and R. L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266(1):195–214, 1997.

[44] W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Comput. Phys.*, 128(24), 2008.

[45] C. D. Christ, A. E. Mark, and W. F. van Gunsteren. Feature Article Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.*, 31(8):1569–1582, 2010.

[46] E. A. Carter, G. Ciccitti, J. T. Hynes, and R. Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.*, 156(5):472–477, 1989.

[47] P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman. Free-energy Calculations by computer-simulation. *Science*, 236(4801):564–568, 1987.

[48] S. Kumar, P. W. Payne, and M. Vásquez. Method for free-energy calculations using iterative techinques. *J. Comput. Chem.*, 17:1269–1275, 1996.

[49] S. Kumar, D. Bouzida, R. Swendsen, P. A. Kollman, and J. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[50] E. Weinan, W. Q. Ren, and E. Vanden-Eijnden. Finite temperature string method for the study of rare events. *J. Phys. Chem. B*, 109(14):6688–6693, 2005.

[51] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemsitry*, 53, 2002.

[52] T. S. van Erp, D. Moroni, and P. G. Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Comput. Phys.*, 118(17), 2003.

[53] S. Trebst, M. Troyer, and U. H. E. Hansmann. Optimized parallel tempering simulations of proteins. *J. Comput. Phys.*, 124(17):174903, 2006.

[54] Y Sugita and Y Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141, 1999.

[55] H. Lei, C. Wu, H. Liu, and Y. Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA.*, 104(12):4925–4930, 2007.

[56] D. Paschek, H. Nymeyer, and A. E. Garcia. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent:

On the structure and possible role of internal water. *J. Struct. Biol.*, 157:542–533, 2006.

[57] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050, 2001.

[58] E Darve and A Pohorille. Calculating free energies using average force. *J. Comput. Phys.*, 115:9169–9183, 2001.

[59] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci USA*, 99(20):12562–12566, 2002.

[60] A. Laio and F. L. Gervasio. Metadynamics: a model to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008.

[61] S Piana and A Laio. A bias-exchange approach to protein folding. *J. Phys. Chem. B*, 111(17):4553–4559, 2007.

[62] F. Baftizadeh, P. Cossio, F. Pietrucci, and A. Laio. Protein folding and ligand-enzyme binding from bias-exchange metadynamics simulations. *Submitted*, 2011.

[63] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277(4):985–994, 1998.

[64] J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A Large-scale experiment to assess protein-structure prediction methods. *Proteins: Struct., Funct., Genet.*, 23(3):R2–R4, 1995.

[65] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollmann. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[66] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM - A Program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[67] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.*, 100(9):L47–L49, 2011.

[68] L. Verlet. Computer experiments on classical fluids .I. Thermodynamical properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–&, 1967.

[69] W. G. Hoover. Canonical dynamics - equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695, 1985.

[70] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gusteren, A. Di Nola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.

[71] M. Parrinello and A. Rahman. Crystal structure and pair potentials: a molecular-dynamics study. *Phys. Rev. Lett.*, 45:1196, 1980.

[72] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690, 1997.

[73] G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96(9), 2006.

[74] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput. Biol.*, 5:e100045, 2009.

[75] F. Pietrucci and A. Laio. A collective variable for the efficient exploration of protein beta-sheet structures: application to sh3 and gb1. *J. Chem. Theory Comput.*, 5:2197, 2009.

[76] S. Piana, A. Laio, F. Marinelli, M. Van Troys, D. Bourry, C. Ampe, and J. C. Martins. Predicting the effect of a point mutation on a protein fold: The villin and advillin headpieces and their Pro62Ala mutants. *J. Mol. Biol.*, 375(2):460–470, 2008.

[77] N. Todorova, F. Marinelli, S. Piana, and I. Yarovsky. Exploring the folding free energy landscape of insulin using bias exchange metadynamics. *J. Phys. Chem. B*, 113:3556–3564, 2009.

[78] 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. *J. Phys. Chem. B*, 114:3259, 2010.

[79] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1–2):141–151, 1999.

[80] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Oldziej, K. Wachucik, and H. A. Scheraga. Optimizing the performance of bias-exchange metadynamics: folding a 48-residue lysm domain using a coarse-grained model. *J. Phys. Chem. B*, 111:260–285, 2007.

[81] J. U. Bowie, J. F. Reidhaarolson, W. A. Lim, and R. T. Sauer. Deciphering the message in protein sequences - Tolerance to amino-acid substitutions. *Science*, 247(4948):1306–1310, 1990.

[82] B. W. Matthews. Structural and genetic-analysis of protein stability. *Annual review of biochemistry*, 62:139–160, 1993.

[83] T. X. Hoang, L. Marsella, A. Trovato, F. Seno, JR Banavar, and A Maritan. Common attributes of native-state structures of proteins, disordered proteins, and amyloid. *Proc. Natl. Acad. Sci. USA.*, 103(18):6883–6888, 2006.

[84] J. R. Banavar and A. Maritan. Physics of proteins. *Annual review of biophysics and biomolecular structure*, 36:261–280, 2007.

[85] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003.

[86] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7(8):306–317, 2001.

[87] F. Pietrucci, F. Marinelli, P. Carloni, and A. Laio. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.*, 131(33):11811–11818, 2009.

[88] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–&, 1996.

[89] W. Kabsch and C. Sander. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[90] 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.

[91] 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):1093–1108, 1997.

[92] W. Kabsch. Discussion of solution for best rotation to relate 2 sets of vectors. *Acta Crystallographica section A*, 34(Sep):827–828, 1978.

[93] Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nuc. Ac. Res.*, 33(7):2302–2309, 2005.

[94] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–417, 1992.

[95] S. C. Lovell, I. W. Davis, W. B. Adrendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by C alpha geometry: phi,psi and C beta deviation. *Proteins: Struct., Funct., Genet.*, 50(3):437–450, 2003.

[96] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski. The continuity of protein structure space is an intrinsic property of proteins. *Proc. Natl. Acad. Sci. USA.*, 106(37):15690–15695, 2009.

[97] G. D. Rose. Hierarchic organization of domains in globular proteins. *J. Mol. Biol.*, 134(3):447 – 470, 1979.

[98] D. Baker. A surprising simplicity to protein folding. *Nature*, 405(6782):39–42, 2000.

[99] B. Fain and M. Levitt. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proc. Natl. Acad. Sci. USA.*, 100(19):10700–10705, 2003.

[100] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37:289–316, 2008.

[101] P. L. Luisi, C. Chiarabelli, and P. Stano. From never born proteins to minimal living cells: Two projects in synthetic biology. *Origins of life and evolution of the biosphere*, 36(5-6):605–616, 2006.

[102] 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: Struct., Funct., Genet.*, 34(1):82–95, 1999.

[103] S. T. Wu, J. Skolnick, and Y. Zhang. Ab initio modeling of small proteins. *BMC Biol.*, 5:7, 2007.

[104] 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), 2010.

[105] G. M. Crippen. Easily searched protein folding potentials. *J. Mol. Biol.*, 260(3):467–475, 1996.

[106] I. Chang, M. Cieplak, R. I. Dima, A. Maritan, and J. R. Banavar. Protein threading by learning. *Proc. Natl. Acad. Sci. USA.*, 98(25):14350–14355, 2001.

[107] M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *J. Comput. Phys.*, 109(24):11101–11108, 1998.

[108] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257(2):457–469, 1996.

[109] F. Seno, A. Maritan, and J. R. Banavar. Interaction potentials for protein folding. *Proteins: Struct., Funct., Genet.*, 30(3):244–248, 1998.

[110] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275(5):895–916, 1998.

[111] J. Meller and R. Elber. Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins: Struct., Funct., Genet.*, 45(3):241–261, 2001.

[112] F. Seno, A. Trovato, J. R. Banavar, and A. Maritan. Maximum entropy approach for deducing amino acid interactions in proteins. *Phys. Rev. Lett.*, 100(7), 2008.

[113] P. Benkert, S. C. E. Tosatto, and D. Schomburg. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Struct., Funct., Bioinf.*, 71(1):261–277, 2008.

[114] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3):623–644, 1996.

[115] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein-sequence through the folding motif. *Proteins: Struct., Funct., Genet.*, 16(1):92–112, 1993.

[116] M. J. Sippl. Calculation of conformational ensambles from potentials of mean force - an approach to the knowledge-based prediction of local structures in globular-proteins. *J. Mol. Biol.*, 213(4):859–883, 1990.

[117] F. Melo and E. Feytmans. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.*, 277(5):1141–1152, 1998.

[118] M. Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, 2006.

[119] D. Rykunov and A. Fiser. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, 11, 2010.

[120] R. Samudrala and M. Levitt. Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9(7):1399–1401, 2000.

[121] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, 268(1):209–225, 1997.

[122] B. Park and M. Levitt. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258(2):367–392, 1996.

[123] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct., Funct., Genet.*, 53(1):76–87, 2003.

[124] J. Handl, J. Knowles, and S. C. Lovell. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, 25(10):1271–1279, 2009.

[125] A. Zemla. LGA: a method for finding 3D similarities in protein structures. *Nuc. Ac. Res.*, 31(13):3370–3374, 2003.

[126] M. F. Lensink, R. Mendez, and S. J. Wodak. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins: Struct., Funct., Genet.*, 69(4):704–718, 2007.

[127] A. Trovato, A. Maritan, and F. Seno. Aggregation of natively folded proteins: a theoretical approach. *Journal of Physics-Condensed Matter*, 19(28), 2007.

[128] A. Varshney, F. P. Brooks, and W. V. Wright. Computing smooth molecular-surfaces. *IEEE Computer Graphycs and applications*, 14(5):19–25, 1994.

[129] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano. Critical assessment of methods of protein structure prediction-Round VIII. *Proteins: Struct., Funct., Bioinf.*, 77:1–4, 2009.

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

[131] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[132] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.*, 55(2):383–394, 2004.

[133] B. Hess, H. Bekker, H. J. C. Berendsen, and G. E. M. J. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463, 1997.

[134] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello. Plumed: a portable plugin for free-energy calculations with molecular dynamics. *Comp. Phys. Comm.*, 180:1961, 2009.

[135] R. Samudrala and M. Levitt. A comprehensive analysis of 40 blind protein structure predictions . *BMC structural biology*, 2:3, 2002.

[136] F. Pietrucci, F. Marinelli, P. Carloni, and A. Laio. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. soc.*, 131(33):11811–11818, 2009.