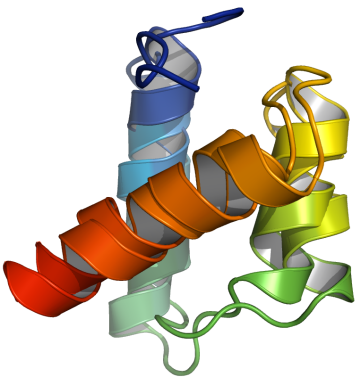
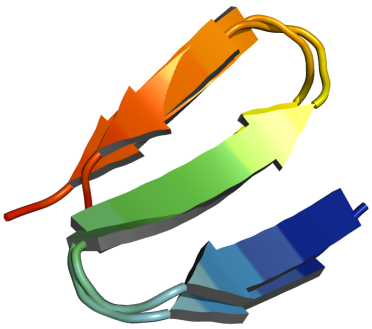
Development and Application of a Free Energy Force Field for All Atom Protein Folding.

DISSERTATION

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
der Fachbereich Physik
der Universität Dortmund





vorgelegt
von

ABHINAV VERMA
aus
Chandigarh, Indien

2007

To My Parents and Teachers

Tag der mündlichen Prüfung	:	19. Jan 2007
Vorsitzender	:	Prof. Dr. Bernhard Spaan
1. Gutachter:	:	Priv-Doz. Dr. Wolfgang Wenzel
2. Gutachter:	:	Prof. Dr. Alfons Geiger
Vertreter der promovierten wissenschaftlichen Mitarbeiter:	:	Dr. Carsten Raas

Contents

List of Figures

List of Tables

Preamble

> "If you want to understand function, study structure." - Francis Crick, Nobel Prize in Medicine, 1962.

Proteins are the workhorses of all cellular life. They constitute the building blocks and the machinery of all cells. DNA carries the genetic information which encodes the production of protein molecules. To produce a protein, the corresponding gene is first transcribed into mRNA and then translated into a polypeptide chain of amino acids in the ribosome.

Proteins perform a variety of roles in the cell: structural proteins constitute the building blocks for cells and tissues, enzymes, like pepsin, catalyze complex reactions, signaling proteins, like insulin, transfer signals between or within the cells. Transport proteins, like hemoglobin, carry small molecules or ions, while receptor proteins like rhodopsin generate response to stimuli. The mechanisms of all these biophysical processes depend on the precise folding of their respective polypeptide chains.

From the work of C.B. Anfinsen and co-workers in the 1960s we now know that the amino acid sequence of a polypeptide chain in the appropriate physiological environment can fully determine its folding into a so-called native conformation. Unlike man-made polymers of similar length, functional proteins assume unique three-dimensional structures under physiological conditions and there must be rules governing this sequence-to-structure transition. Protein structures can be determined experimentally, by X-ray crystallography or NMR methods, but these experiments are still challenging and do not work for all proteins. From the theoretical standpoint it is still not possible to reliably predict the native three-dimensional conformation of most proteins given their amino acid sequence alone.

The triplet genetic code by which the DNA sequence determines the amino acid sequence of polypeptide chains is well understood. However, unfolded polypeptide chains lack most of the properties needed for their biological function. The chain must fold into its native three dimensional conformation in order to perform its function. Despite much research in this direction and the emergence of novel folding paradigms during the last decade, much of the mechanism by which the protein performs this auto-induced folding reaction is still unclear.

To perform their biological function polypeptide chains interact with their aqueous or lipid environment to fold into discrete, highly organized three-dimensional structures. Because of great advancement in sequencing techniques for proteins and nucleotides compared to structure determination methods, the number of known protein structures lags far behind the number of known sequences. Various genome projects have rapidly increased the number of known sequences. Entire genomes are

reported for the human, the mouse, the chicken, the fruit fly and many fungi. Currently over one million protein sequences are known, compared to about 40,000 structures deposited in the Protein Data Bank (the world-wide database of protein structures). Reliable theoretical methods for protein structure prediction could help to reduce this gap between sequence and structural databases and elucidate the biological information in structurally unresolved sequences.

Therefore it would be very helpful to develop methods for protein structure prediction on the basis of the amino acid sequence alone. Even if this goal it is not fully realized, methods that can complete partially resolved experimental protein structures would be very helpful to determine the structure of proteins where neither theoretical methods nor experimental techniques alone can succeed. For the trans-membrane family of proteins, present day experimental methods fail, which is responsible for the entire communication of the cell with its environment. Theoretical methods would be very helpful to investigate these proteins.

There are large number of related questions, for instance regarding the interactions of a given protein with a large variety of other proteins, where theoretical methods could also contribute to our understanding of biological function. Protein-protein interactions govern the cell signaling processes and are very important for the assembly of large protein structures in the cell. Because it is known that proteins change their shape upon binding to other proteins, the structure of the isolated constituents is only an approximation to the structure found in the complex in which the proteins ultimately function. In order to address these questions it is important to develop accurate atomistic models for protein structure prediction. To use a protein structure for emerging methods of computer aided drug design the resolution of the protein structure must be below 1 Å. In order to predict the binding sites or interacting complex of two proteins a resolution between 3-5 Å is desirable.

Related to the question of protein structure prediction is the question of how the proteins attain their final conformation - the so called protein folding problem. It remains one of the astonishing mysteries responsible for the evolution of life how these complex molecules can attain a unique native conformation with such precision. No man-made polymer of similar size is able to assemble into a predetermined structure with the precision encountered in the proteins that have evolved in nature.

Given its complexity it is not surprising that the protein folding process occasionally fails, and many of such failures are related to cellular disfunction or disease. Therefore it is important not only to be able to predict the final structure of proteins but also very desirable to understand the mechanisms by which proteins fold.

Experiments of C.B. Anfinsen and co-workers showed convincingly that many proteins can indeed adopt their native conformation spontaneously, *i.e.* sequence determines structure. This led to the "thermodynamic hypothesis" which states that the native three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, etc.) minimizes the Gibbs free energy of the whole system. The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence in a given environment. This led to the "Levinthal Paradox" which suggested that there must be pathways for protein folding, as a simple protein with a 100 amino acids is estimated to have a vast configurational space of the order of $2^{100}(\sim 10^{30})$ possible conformations. Unless there is a specific mechanism, such a protein will need more than the age of the universe to

locate its global free energy minimum in an exhaustive search of this configurational space. For this reason, Levinthal stipulated that there must be a specific, multi-step folding reaction that leads to the native conformation. Unfortunately only very few of the proposed multitude of intermediates in this folding path were ever found. Instead a family of small proteins was detected that folds in a two-state fashion, in which no discernable folding intermediates exist at all. Protein folding in this scenario appears as a single reaction between folded and unfolded state of the protein with only one intervening energy barrier. For some proteins even barrierless folding was observed.

Levinthal's paradox can be resolved by the funnel paradigm of protein folding. In this paradigm proteins are generally thought to have globally funneled energy landscapes with a small gradient directed towards the native state. This "folding funnel" landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather than being restricted to a single mechanism. Single molecule experiments, such as atomic force microscopy and optical tweezers have confirmed the existence of such funnels in protein folding.

For protein structure prediction, a bi-annual contest (Critical Assessment of Techniques for Protein Structure Prediction, CASP) attempts to measure progress in the field of protein structure prediction. In this contest, sequences of recently resolved proteins are published, but the experimental structures are withheld until the end of the contest. In these contests methods that copy structure from proteins of very similar amino-acid sequence are successful if there is a sufficient degree of sequence-identity. For proteins with less than thirty percent sequence identity to a known protein, however, the resulting structures often do not agree well with the experimental results. It is presently debated whether the last four to six years have seen any progress for low homology proteins at all.

The ultimate, very long-range goal of protein structure theory would be the development of methods to design proteins for a specific function. This would be very helpful for medical purposes and technological applications in nanobiology, but will require an understanding of various factors that influence the folding of the polypetide and their sequence determinants. It is currently possible to modify existing proteins and also to generate a variety of hybrids. But the ability to design completely new proteins to carry out novel functions requires a much more profound understanding of how sequences determine folding.

Many theories and increasingly computational methods have been developed to understand the folding process. Simplified models have been applied to understand its physical principles. Lattice based methods were among the first models that allowed efficient sampling of conformational space. The lattice models, either 2D square or 3D cubic, were used to study protein folding and unfolding, but they were too simplified for protein structure prediction. Subsequently "Gō-Models" were developed, where only native contacts interact favorably, and were useful to characterize some aspects of the folding of small proteins. The success of these models is limited by the fact that all residues interact in the same way. Further development led to statistically obtained knowledge based potentials. These potentials were obtained and parameterized on the structures available from the Protein Data Bank. The knowledge based potentials are mostly used for fold recognition or protein structure prediction.

With the increase in computational resources and speed, all-atom molecular dynamics simulations of protein folding have been undertaken. For most proteins, it is still not feasible to determine the protein structure from extended conformations using a single molecular dynamics simulation. This is

due to the fact that at the all-atom level, the typical time step in a molecular dynamics simulation is about 1 femtosecond while the protein folding occurs at at millisecond timescale. A single simulation would need years to complete. Replica exchange MD simulations have been successful in folding proteins from extended conformations, but are still limited to the size of 20-30 amino acids.

In this thesis we explore an alternate approach for protein structure prediction and folding that is based on the Anfinsen's hypothesis that most proteins are in thermodynamic equilibrium with their environment in their native state. For proteins of this class the native conformation corresponds to the global optimum of the free energy of the protein. We know from many problems in physics and chemistry that the global optimum of a complex energy landscape can be obtained with high efficiency using stochastic optimization methods. These methods map the folding process found in nature onto a fictitious dynamical process that explores the free-energy surface of the protein. By construction these fictitious dynamical processes not only find the conformation of lowest energy, but typically characterize the entire low-energy ensemble of competing metastable states. Since the total free-energy change for protein folding under physiological conditions is small, often only a few kcal/mol, a characterization of the low-energy ensemble of thermodynamically accessible protein conformations may be sufficient not only to predict the structure of the protein, but also to characterize the folding process.

There are two important ingredients to this approach: first we need an accurate atomically resolved free energy force field for proteins. Second we need a set of simulation methods that can reliably explore and characterize the low-energy ensemble of protein conformations.

In this thesis, we will discuss

- the development of an an effective free energy force field to study protein folding for a wide range of proteins,
- the development and implementation of stochastic optimization methods that locate the global minimum, and
- all atom folding studies of various proteins.

The first two chapters in this thesis give an introduction to protein folding. The first chapter introduces protein composition and structure, their formation, properties and the problem of protein folding. The second chapter deals with the biomolecular simulations and current strategies used for computer simulations of biomolecules.

In the third chapter, we describe the protein force field PFF01 and the development of PFF02. PFF01 was a force field for helical proteins only. There is however, another class of protein secondary structure, so called β-sheets. PFF01 was not successful for folding proteins with this type of secondary structure. Two new terms were included in PFF01 to reliably fold three β-sheet proteins which misfold in in protein folding studies with PFF01. The new force field PFF02, however, still located the native conformation of various helical proteins at their global free energy minimum.

The validation of the force field can be obtained by carrying out all atom protein folding studies staring from extended conformations. Alternatively it can be done by ranking decoy sets of protein conformations in the force field. A decoy set is a large library of conformations of a protein generated

to approximately span all relevant low-energy regions of the free energy surface. If the energy of a near-native decoy is lower than the energy of all other decoys, then the force field is considered selective in locating native-like decoy at the energy minimum. The larger the energy difference is, the better is the selectivity of the force field. In the fourth chapter we study the validity and selectivity of the modified force field PFF02 for two decoy sets generated by PFF01 and Rosetta. PFF01 and Rosetta decoy set were available for 5 and 32 proteins respectively. We calculated the energies in PFF02 for various decoys (low energy conformations) for all of these proteins and compared them with the energies of near-native decoys generated from the native state of the protein. PFF02 emerged as highly selective with an average Z-score of -2.74 and -3.26 for PFF01 and Rosetta decoy sets respectively.

As mentioned earlier, the best test of the force field is to fold various proteins starting from extended conformations. As the low-energy region of the energy landscape of proteins are extremely rugged, efficient methods are needed to locate the global minimum of these protein energy landscapes. The fifth chapter describes the development and implementation of such methods that can be used for protein folding with our free energy force field PFF02. Two methods, namely the optimized basin hopping method and an evolutionary algorithm are described in this chapter. Both these methods were successfully used for protein folding. The basin hopping method is a simple and efficient method for protein folding and is serial in nature. The evolutionary algorithm, in contrast, is a parallel implementation, which scales near-perfectly with the number of processors. We tested the scaling up to 4096 processors. With the evolutionary algorithm it was possible to achieve all-atom folding of over 50 amino acid proteins in a single day.

In the sixth chapter, we study all atom folding of various proteins in PFF02 with basin hopping technique and evolutionary algorithm. We first report the folding of five helical proteins ranging from 20-56 amino acids in size. The proteins included a single helix, two-helical and three-helical bundle proteins. Next we studied the proteins with β-sheet elements from 12-20 amino acids.The proteins included six hairpins and a three-stranded β-sheet. Finally we study the folding of a 29 amino acid mixed protein with one helix and two strands. PFF02 could fold all of these proteins into their respective free energy minimum. The average root mean square deviation for the lowest energy conformations achieved by all atom folding of these 13 proteins is only 2.87 Å.

We have thus developed an universal free energy force field that locates the native state of a protein at its free energy minimum. We have also developed and implemented stochastic methods which can be used to fold proteins following the thermodynamic hypothesis and locate the global minimum. Using PFF02 with these stochastic methods, we could reliably and reproducibly predict the native state of various proteins including pure helical, pure beta and mixed systems.

1

Introduction

The proteins we observe in nature have evolved to perform specific functions, such as catalyzing various reactions and carrying ions or other small molecules to various parts of the body. The functional property of a protein depends upon its three dimensional structure. Under physiological conditions, a particular sequence of amino acids in a polypeptide chain folds into a compact three-dimensional structure. This three dimensional structure, due to the specific properties, makes a protein perform a specific biological function. These single chains, which are folded into a respective three dimensional structure, can still assemble together to form more complex functional units.

To understand the biological function of a protein, one needs to measure or predict its three dimensional structure from its amino-acid sequence. This prediction problem is still unsolved and remains one of the most basic challenges in biophysical chemistry.

The fundamental reason why the prediction problem remains unsolved lies in the large size of the conformational space that is accessible to a single protein (Branden and Tooze, 1999; Berg et al., 2001).

1.1 Amino Acids

The basic monomeric unit of a protein is an amino acid. There are twenty naturally occurring amino acids. All of the twenty amino acids have a central carbon atom (C_α), to which are attached a hydrogen atom, an amino group(NH_2), and a carboxyl group(COOH). The side chain which is attached as the fourth valency to the C_α differentiates the various amino acids. There are twenty different naturally occurring amino acids specified by the genetic code [*]. The twenty different side chains that occur in natural proteins are shown in Figure 1.1. Their names are commonly abbreviated with either a three-letter code or a one-letter code.

Amino acids are linked together by the formation of peptide bonds to form a chain. A peptide bond is formed when the carboxyl group of the first amino acid reacts with the amino group of the next to eliminate water, as shown in Figure 1.2. This process is repeated until the whole protein chain is synthesized. At the ends of the polypeptide chain, the amino group of the first amino acid and the

[*]There are very rare occurrences of some other amino acids in proteins.

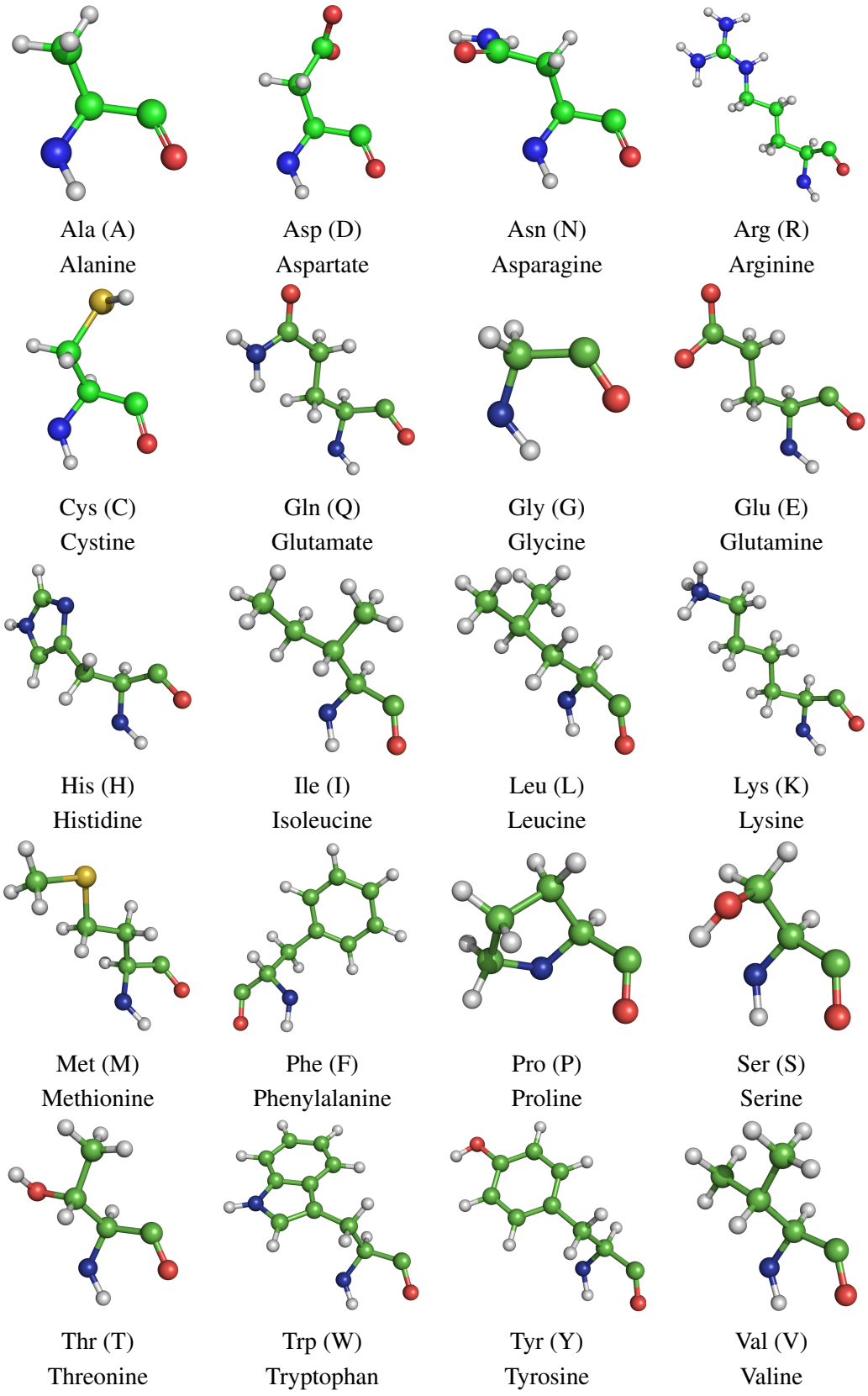


Figure 1.1: Twenty naturally occurring amino acids: Their structure, name, three-letter code and one-letter code. The structures are color coded with carbon(green), nitrogen(blue), oxygen(red) , hydrogen(white) and sulphur(orange)

carboxy group of the last amino acid still remain intact. Thus the chain is generally referred as to run from amino(N) terminus to carboxy(C) terminus. The formation of a succession of peptide bonds generates a “main chain” or “backbone” from which various “side chains” project outwards.

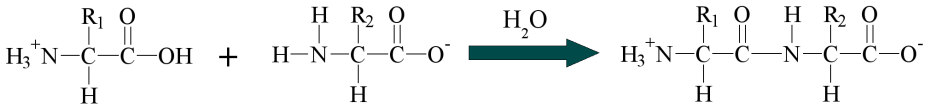
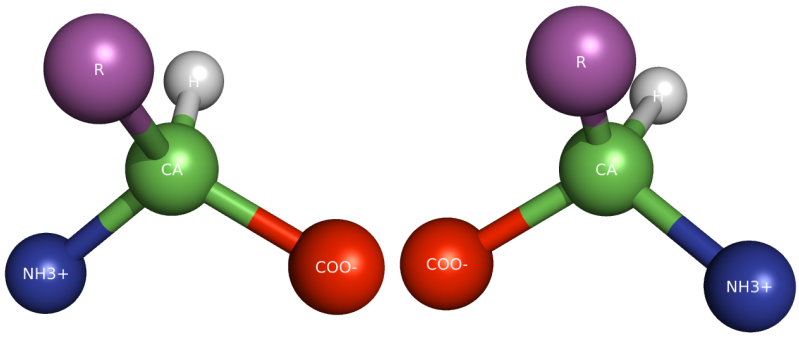


Figure 1.2: Formation of a peptide bond

The main chain atoms of a polypeptide chain are a carbon atom C_α to which the side chains is attached, a NH group bound to C_α, and a carbonyl group C=O, where the carbon atom C is attached to C_α. These units are called residues and are linked into a polypeptide chain by peptide bonds between the C atom of one residue and the nitrogen atom of the next. The basic repeating unit along the main chain is thus (NH-C_αH-CO), which is the residue of the common parts of amino acids after peptide bonds have been formed.

At the fourth valency of C_α is the side chain and depending upon the chemical structure of the side chain, the amino acids are divided into three different classes(Branden and Tooze, 1999). The first class comprises those with strictly hydrophobic side chains Ala(A), Val(V), Leu(L), Ile(I), Phe(F), Pro(P), and Met(M). The second class includes four charged residues Asp(D), Glu(E), Lys(K) and Arg(R) and the third class comprises those with polar side chains Ser(S), Thr(T), Cys(C), Asn(N), Gln(Q), His(H), Tyr(Y) and Trp(W). The amino acid glycine(G) has only a hydrogen atom as the side chain and thus is the simplest of all the twenty amino acids. The amino acid proline(P) is also different from the rest as it is the only amino acid where both ends of the sidechain are covalently bound to the main chain.



(a) L form (b) D form

Figure 1.3: The L and D chiral forms of amino acids. R(magenta) represents the sidechain.

All amino acids (except glycine) are chiral molecules which can exist in two different forms with

different hands, L or D-form (see Figure 1.3). Biological systems depend on specific detailed recognition of molecules involving differentiation between chiral forms. Amino acids are found in only one of the chiral forms, the L-Form, during protein synthesis. There is, however, no obvious reason why the L-form was chosen during the evolution and not the D-form(Weatherford and Salemme, 1979; Mason, 1984).

1.2 Conformation of the polypeptide chain

The linkage of amino acids produces a polypeptide chain, with the backbone atoms linked through the peptide bond which does not change in its chemical structure during folding. The folding pattern of the polypeptide chain can be described in terms of angles of internal rotation around the bonds in the main chain. The bonds in the polypeptide backbone between N and C_α and between the C_α and C, are single bonds. Internal rotations around these bonds are not restricted by the electronic structure of the bond, but only by possible steric collisions in the conformations produced. In contrast, the peptide bond itself has a partial double bond character, with restricted internal rotation(Lesk, 2001). This means that the NH and CO along with the two C_α's always remain in a peptide plane(see yellow regions in Figure 1.4).

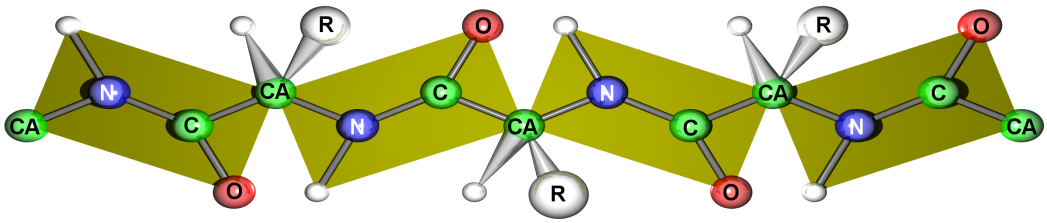
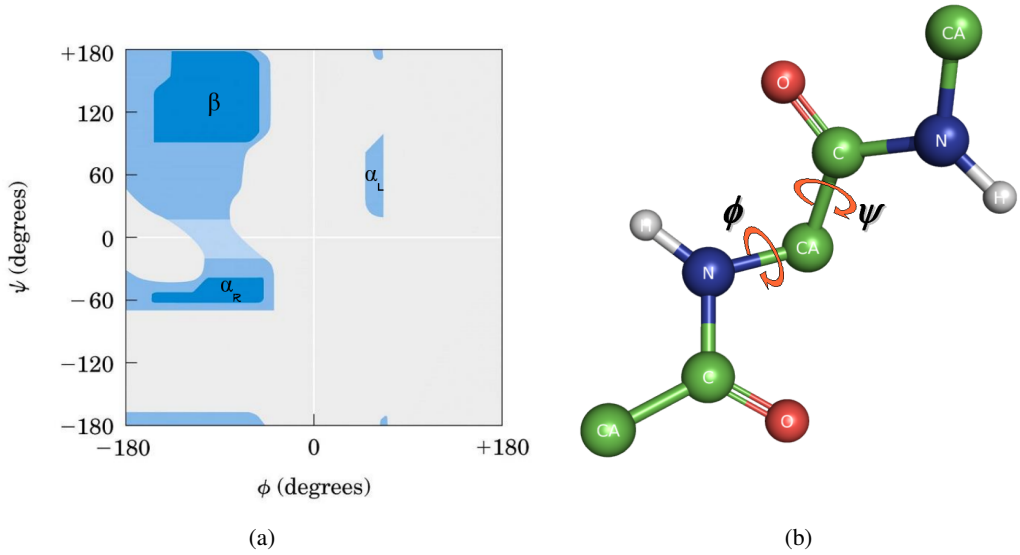


Figure 1.4: Peptide planes of a polypeptide chain. R represents the side chain and white spheres are hydrogens

The peptide group can occur in both *cis* and *trans* forms, with the *trans* isomer being the more stable. For all the amino acids except proline, the energy difference between *cis* and *trans* states is very large(Ramachandran and Mitra, 1976). For proline, the energy difference is only about 1.2 Kcal·mol^{-1}(Lesk, 2001). Taking a section of three peptide units having the sequences *trans-trans-trans* and *trans-cis-trans*, conformational energy calculations indicate that the latter can occur only to an extent of 0.1%, unless there occurs the sequence X-Pro, in which case it is of the order of 30%. This explains the extreme rarity of *cis* peptide units in proteins. As a result, virtually all the *cis* peptides in proteins appear between a proline and the residue preceeding it in the chain, however, it follows that even with nonprolyl residues, *cis* peptide units are not forbidden, but can occur in some rare examples(Ramachandran and Mitra, 1976).

1.3 The Sasisekharan - Ramakrishnan - Ramachandran Plot

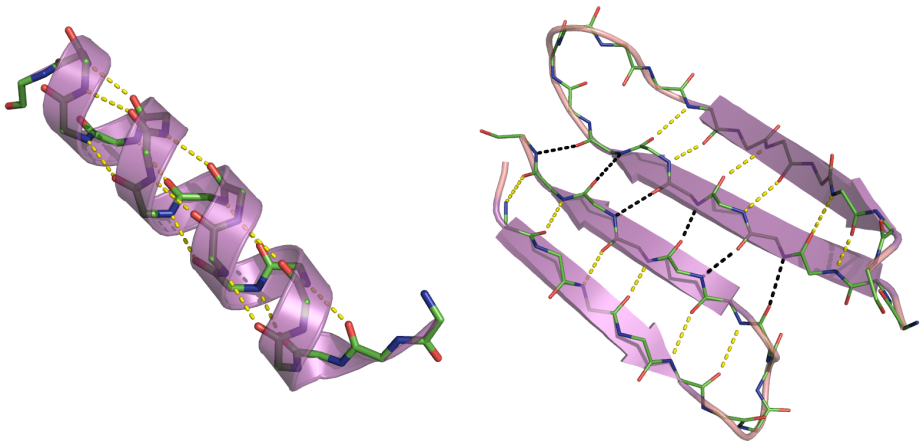
As most residues in proteins have *trans* peptide bonds, the main chain conformation of each residue is determined by two angles, commonly named as ϕ and ψ. The dihedral angle around the bond N-C_α is known as ϕ and the dihedral angle around the bond C_α-C is known as ψ. As ϕ involves a previous amino acid and ψ involves the next, the first amino acid and the last amino acid in the polypeptide chain have only one angle of rotation (ψ and ϕ respectively). The angles of rotation are shown in Figure 1.5(b). Many combinations of ϕ and ψ produce sterically disallowed conformations. V. Sasisekharan, C. Ramakrishnan and G.N. Ramachandran first plotted the "allowed" regions in a graph of ϕ and ψ (Ramachandran et al., 1963). The plot is generally known as the Ramachandran plot, shown in Figure 1.5(a). There are two main allowed regions, one around $\phi = -57°, \psi = -47°$ (denoted α_R) and the around $\phi = -125°, \psi = +125°$ (denoted β) with a neck like region between them. The mirror image of α_R, denoted α_L, is allowed equally for glycine residues only because glycine is achiral.



(a) (b)

Figure 1.5: Sasisekharan-Ramakrishnan-Ramachandran Plot (The figure for Ramachandran plot was taken from the internet from http://bifi.unizar.es/)

The two major allowed regions correspond to the two major types of secondary structures found in proteins, helix and sheet. A continuous stretch of residues, with all conformations in the α_R region, would form a right handed helix (A helix formed by a stretch of residues in the α_L region would form the corresponding left handed helix). In the β region, the chain is nearly fully extended. A continuous stretch of residues, with all the conformations in the β region, would form a single strand of a sheet (Both helix and sheet are discussed in detail in the next section). The conformations that correspond to low energy states of individual residues also permit the formation of structures with extensive main chain hydrogen bonding. The two effects thereby cooperate to lower the energy of the native state.



(a) α-Helix (b) β-Sheets

Figure 1.6: Secondary structural elements. Dashed lines indicate the presence of hydrogen bonds. in (b) yellow/black bonds are shown for antiparallel/parallel β-sheets respectively.

1.4 Protein Structure

Proteins are made up of unique sequences of the twenty naturally occurring amino acids. The protein structures are classified into four categories depending upon the amount of information known.

Primary Structure

Primary structure describes the sequence of amino acids starting from amino(N) terminus to carbonyl(C) terminus. The primary sequence is written in either the one-letter code or three-letter code, for example, SWTWEGNKWTWK or SER-TRP-THR-TRP-GLU-GLY-ASN-LYS-TRP-THR-TRP-LYS (three letter code and one letter code as described in Figure 1.1) for the tryptophan zipper protein (PDB code:1LE0).

Secondary Structure

Due to the "allowed" regions of the Ramachandran plot, polypeptide chains fold themselves into regularly repeating structures. In 1951, L. Pauling and R. Corey proposed two periodic structures called the α-helix and the β-pleated sheet. Later, other structures such as the β-turn and Ω-loop were also identified. Although not periodic, these common turn or loop structures were well defined and contribute along with α-helices and β-sheets to form the final protein structure.

- **Alpha Helix:** The α-helix is a spring like structure where tightly coiled backbone forms the inner part of the helix and the side chains project outwards in a helical array (see Figure 1.6(a)). The α-helix is stabilized by hydrogen bonds between the NH and CO groups of the main chain. In particular, the CO group of each amino acid forms a hydrogen bond with the NH group

Structure	ϕ	ψ	n	d(Å)
α-helix	-57	-47	3.6	1.5
3_{10}-helix	-49	-26	3.0	2.0
π-helix	-57	-70	4.4	1.1
Polyproline II helix	-79	+149	3.0	3.1
Parallel β-strand	-119	+113	2.0	3.2
Antiparallel β-strand	-139	+135	2.0	3.4

Table 1.1: Structural parameters for protein secondary structures. ϕ and ψ are the conformational angles of the mainchain, n is the number of residues per turn, d is the displacement between successive residuesalong the axis.

of the amino acid which is situated four residues ahead in sequence. Thus, except the amino acids near the ends of an α-helix, all the main chain CO and NH groups are hydrogen bonded. Each residue is related to the next one by a rise of 1.5 Å along the helix axis and a rotation of 100°, which gives 3.6 amino acid residues per turn of helix. The screw sense of a helix can be right handed or left handed. The Ramachandran plot reveals that both the right handed and left handed helices are among the allowed conformations. However, right handed helices are energetically more favorable because there is less steric clash between the side chains and the backbone. Essentially all α-helices found in proteins are right handed. There are also other types of helices, such as a 3_{10}-helix, a π-helix and polyproline II helix. The ideal parameters of these are given in Table 1.1.

- **Beta Sheet:** The β-sheet differs remarkably from the spring like α-helix. A polypeptide chain, called a β-strand, in a β-sheet is almost fully extended rather than being tightly coiled as in a helix. The distance between adjacent amino acids along a β-strand is approximately 3.5 Å in contrast with a distance of 1.5 Å along an α-helix(Lesk, 2001). The side chains of adjacent amino acids point in opposite directions. A β-sheet is formed by linking two or more β-strands by hydrogen bonds. Adjacent chains in a β-sheet can run in opposite directions (antiparallel β-sheet) or in the same direction (parallel β-sheet), shown in Figure 1.6(b). In the antiparallel arrangement, the NH group and the CO group of each amino acid are respectively bonded to the CO and NH group of a partner on the adjacent chain. In the parallel arrangement, for each amino acid, the NH group is hydrogen bonded to the CO group of one amino acid on the adjacent strand, whereas the CO group is hydrogen bonded to the NH group on the amino acid two residues further along the chain. Hydrogen bonding of parallel and antiparallel β-strands are shown is Figure 1.6(b) in black and yellow respectively. The ideal parameters are given in Table 1.1. Many strands come together to form β-sheets with minimum being two for a β-hairpin and as many as ten in β-barrel proteins. Such β-sheets can be purely antiparallel, purely parallel or mixed.

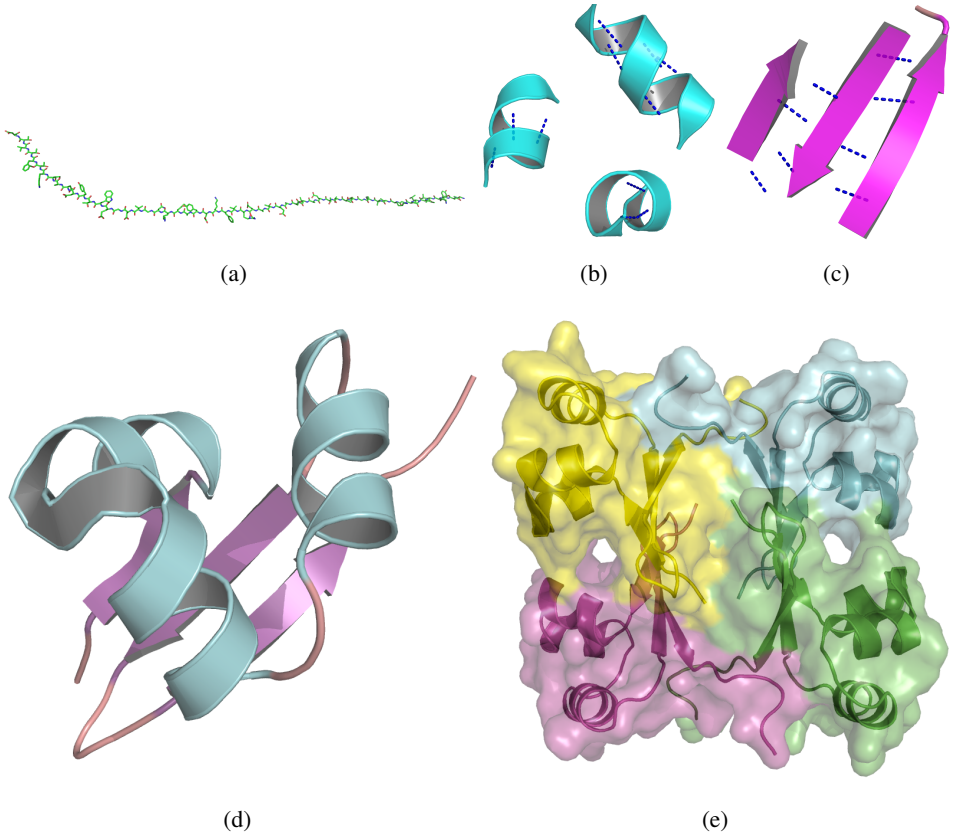


Figure 1.7: Protein Structure Classifications: (a) shows primary structure,(b) & (c) shows the secondary structural elements, helices and sheets respectively, (d) shows the tertiary structure and (e) shows the quaternary structure of a protein.

Tertiary Structure

The tertiary structure is formed by the assembly of secondary structural elements along with turns and loops into a three dimensional arrangement. The tertiary structure mainly has a hydrophobic core with charged residues on the surface of protein. The charged residues on the surface gives the protein its biological activity and is thus responsible for its biological function.

Quarternary Structure

Tertiary structures of proteins (independent folding chains) can still assemble themselves under physiological conditions in order to perform specific functions. These are termed as quaternary structure. For example four identical chains come together to form the hemoglobin complex. Figure 1.7 shows all four kinds of structure for a gene regulating protein (PDB code: 5CRO).

1.5 Dominant forces in Protein Folding

During folding, different sets of residues come in proximity of each other in different possible conformations of the same polypeptide chain. The interactions of side chains and main chain, with one another and with the solvent and with other surrounding proteins or ligands, determine the energy of the conformation. Proteins have evolved so that one folding arrangement of the backbone and its side chain produces a set of interactions that is significantly more favorable than all other possible conformations. This conformation is called the native state of a protein. The experiments of C.B. Anfinsen and coworkers showed that for many proteins, the protein structure is determined by the amino acid sequence alone.

Formation of the native state is a global property of a protein. In most cases, the entire protein (or at least a large part) is necessary for stability. This is because many of the stabilizing interactions involve parts of the protein that are very distant along the polypeptide chain, but brought into spacial proximity by the folding process.

Proteins are only marginally stable, and achieve stability only within narrow ranges of conditions of solvent and temperature. Outside of these regions proteins lose their definite compact structure, and even their helices and sheets, and take up states with disorder in the backbone conformation and specific interactions among residues(Hollecker and Creighton, 1982; Matthews, 1987).

Protein structures are stabilized by a variety of chemical interactions for their stability and for their affinity and specificity for ligands.

1. **Covalent and coordinate chemical bonds:** Some proteins contain covalent chemical bonds between side chains. These covalent bonds such as disulphide bridges between cystine residues are quite common and these sets of cystine residues “lock” the polypeptide chain together.

2. **Hydrogen bonding:** Certain groups in proteins can form hydrogen bonds with water or other protein groups. The main chain has one H-bond donor (N-H) and H-bond acceptor (C=O) for each amino acid. In addition, some polar side-chains can form hydrogen bonds. The main chain, containing peptide groups, must pass through the interior, and some polar side chains are also buried. They thereby lose their interactions with water. To recover the energy, buried polar atoms form protein-protein hydrogen bonds. The standard secondary structures, helices and sheets, are achieved by the formation of hydrogen bonds by the main chain atoms.

3. **Hydrophobic effect:** For proteins to take their native states in the aqueous environments, hydrophobic residues bury themselves in the interior and charged residues come on the surface. The accessible surface area of the protein, calculated from a set of atomic coordinates, measures the thermodynamic interaction between the protein and water.

4. **van der Waals forces and dense packing of protein interiors:** The packing of atoms in protein interiors contributes in two ways to the stability of structure. One is the exclusion of hydrophobic atoms from contact with water. The other is the dispersive attraction between the protein atoms.The cohesion of ordinary substances shows the existence of attractive forces between atoms and molecules. As the matter does not collapse, there must be limits to how far

it can be compressed. This observation leads to the presence of repulsive forces at short range. The most general type of interatomic force, the van der Waals force, reflects this principle: The nearer the atoms, the stronger the attractive force, until the atoms are in contact, at which the forces become repulsive and strong. To maximize the total cohesive force, therefore, as many atoms as possible must be brought as close together as possible. It is the requirement for a dense packing that imposes a requirement for structure in the interior of a protein. It produces a fit of the elements of secondary structure packed together in protein interiors.

1.6 Protein folding problem

In his pioneering work, C. B. Anfinsen, showed that the necessary information for the polypeptide chain to fold into its native structure is contained in its sequence of amino acids. Protein refolding especially demonstrated that the native conformation of many proteins is reproducibly formed even when the proteins are in isolation. This observation can be explained, if the native state is lower in free energy than all other conformations. This observation led to the thermodynamic hypothesis (Anfinsen, 1973) that the native state is the global minimum in the free energy. The stability of each possible conformation of a polypeptide chain depends on the free energy change between native and unfolded states given by equation:

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where $\Delta G, \Delta H$, and ΔS are the differences between free energy, enthalpy, and entropy respectively, of the native and unfolded conformation. The enthalpic difference is the difference associated with atomic interactions (electrostatic interactions, van der Waals potentials, hydrogen bonding) whereas the entropy term describes hydrophobic interactions, thereby including the dominant interactions in protein folding, namely, the hydrophobic effect, hydrogen bonding and configurational entropy. The free energy of stabilization of proteins under ordinary conditions is typically only a few Kcal $\cdot$ mol^{-1} (Privalov, 1979; Privalov and Gill, 1988; Lesk, 2001) and slight changes in the surrounding conditions can force a protein to adopt a completely different conformation.

In an unfolded protein, the polypeptide chain can adopt different rotameric positions around ϕ and ψ torsional angels, and side chain can adopt different rotamers around their dihedral angles. When folded, the ϕ and ψ dihedral angles of the polypeptide chain are nearly restricted to a narrow range of values, as are majority of χ angles. This loss of freedom translates into a loss of configurational entropy. This loss of configurational entropy must be overcome by favorable interactions, such as hydrogen bonding, increase in solvent entropy, etc, in order to fold a polypeptide chain into a stable conformation(Baldwin, 1986; Dill, 1990; Dill et al., 1995; Makhatadze and Privalov, 1996).

While the experiments by C.B. Anfinsen and co-workers demonstrated that many proteins can adopt their native conformation spontaneously, it immediately raised a fundamental problem known as Levinthal's paradox(Levinthal, 1968). Anfinsen's experiments suggested that the native state of a protein is thermodynamically the most stable state under biological conditions. But a polypeptide chain has enormous number of possible conformations (at least 2^{100} for an 100 amino acid protein considering are only two possible conformations per amino acid). If one estimates that each state

is reached in 1ps from a related conformation, such a chain would take $\sim 2^{100}$ ps (considering one ps per conformation) or $\sim 10^{10}$ years (even more than the estimated age of universe) to sample all possible conformations and to find the lowest energy state. Levinthal thus concluded that a specific folding pathway must exist and that protein folding is under kinetic control rather than thermodynamic control.

This issue can be resolved by considering a balance between kinetics and thermodynamics in an energy landscape perspective. According to the energy landscape paradigm, the free-energy landscape has a small gradient in all confromations towards the native state. Even in the absence of a unique folding pathway the protein dynamics is guided towards the native state. Projected to low dimension, the free energy surface thus has a funnel like slope. The landscape perspective explains the process of reaching a global minimum in free energy (satisfying Anfinsen's experiments) and doing so quickly (satisfying Levinthal's concerns) by multiple folding routes on funnel-like energy landscapes(Leopold et al., 1992) because the new view recognizes that "folding pathways" are not the correct solution to the kinetic problem Levinthal posed.. The funnel theory includes ruggedness on the funnel surface(see

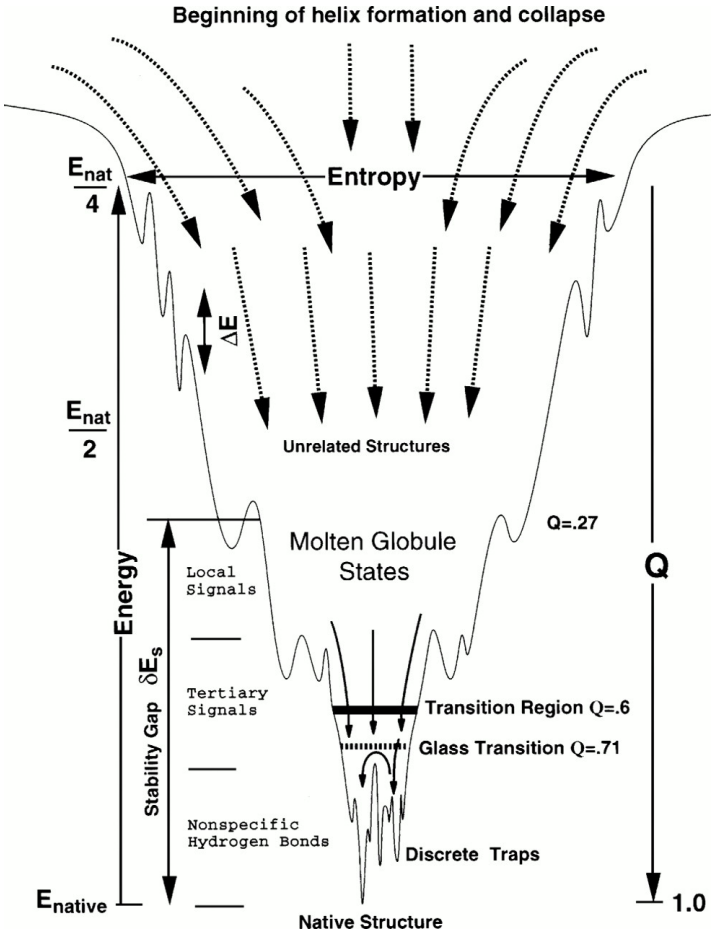


Figure 1.8: Schematic representation of the protein folding landscape (taken from Onuchic et al. (1997))

Figure 1.8). The main idea is that while the folding landscape resembles a funnel globally but is to some extent rugged locally, *i.e.* with traps in which the protein can be trapped along the folding pathway. The funnel guides the protein through many different sequences of traps toward the low-energy folded (native) structure. Here there is no pathway but a multiplicity of folding routes. For small proteins, discrete pathways emerge only late in the folding process when much of the protein has almost reached the native ensemble. The simple parts of the folding process, where most of the real molecular organization is going on, occur in the early events of folding and can be described using a few parameters statistically characterizing the protein folding funnel(Onuchic et al., 1997; Chan and Dill, 1998).

2

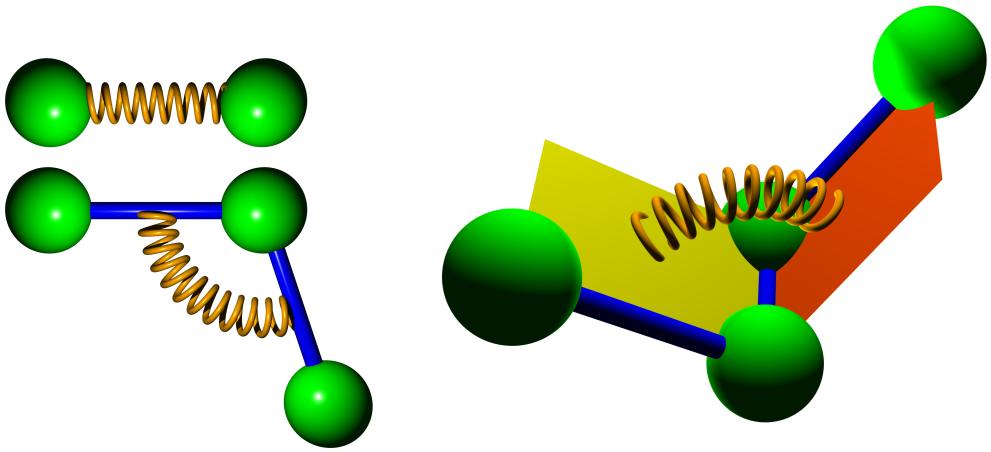
Forcefields and Biomolecular Simulation

Any computational approach to study a chemical system requires a mathematical model to calculate the energy of the system as a function of its conformation. Central to the success of the study is the quality of the mathematical model used. For smaller chemical systems studied in the gas phase, quantum mechanical(QM) approaches are appropriate and feasible. Walter Kohn and John A. Pople jointly won the Nobel prize in chemistry in 1998 for the development of the density-functional theory and computational methods in quantum chemistry. However, these methods are typically limited to system of approximately hundred atoms or less, although approaches to treat large systems are under development. Systems of biophysical or biochemical interest typically involve macromolecules that contain thousands of atoms plus their surrounding environment. In addition to the large size of the system, the inherent dynamical nature of biomolecules require long simulation times, *i.e.* many energy calculations. Many processes of biophysical relevance occur on microsecond to millisecond time scales, while the individual time step of the methods commonly used today are of the order of femtosecond. Thus the energy function might be subjected to over 10^8 energy evaluations in a single simulation.

Atomistic energy functions fulfill the demands required by computational studies of biochemical and biophysical systems. Empirical force fields use atomistic models, in which atoms are the smallest particles in the system rather than the electrons and nuclei used in quantum mechanical descriptions. The mathematical equations in these empirical energy functions include relatively simple terms to describe the physical interactions that dictate the structure and dynamical properties of biological molecules. These simplifications allow for the computational speed required to perform a large number of energy evaluations on biomolecules in their environment. Empirical energy functions were first used for small organic molecules, where it was referred as molecular mechanics, but are now regularly applied to biological systems (Mackerell, 2000).

Some of the standard force fields available for biomolecular simulations are :

AMBER	Assisted Model Building with Energy Refinement (Pearlman et al., 1995)
CHARMM	CHemistry at Harvard Macromolecular Mechanics (MacKerell et al., 1998)
GROMOS	GROningen MOlecular Simulation (Scott et al., 1999)
OPLS	Optimized Potentials for Liquid Simulations (Jorgeson, 1981)
CFF	Consistant ForceField (Hagler and Ewig., 1994)



(a) (b)

Figure 2.1: Schematic representations of the bonded interactions

2.1 Potential Energy Functions

A potential energy function is a mathematical model that parameterizes for the potential energy, V of a chemical system to as a function of its three dimensional(3D) structure $\overrightarrow{R}$. The equation includes terms describing the various physical interactions that are relevant to describe the structure and properties of a chemical system. The total potential energy of a chemical system with a defined 3D structure, $V_{total}(\overrightarrow{R})$, can be separated into terms for internal, $V_{internal}(\overrightarrow{R})$, and external $V_{external}(\overrightarrow{R})$, potential energy as described in the following equations.

$$V_{total}(\overrightarrow{R}) = V_{internal}(\overrightarrow{R}) + V_{external}(\overrightarrow{R}) \tag{2.1}$$

$$V_{internal}(\overrightarrow{R}) = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} K_\chi[1+cos(n\chi-\delta)] \tag{2.2}$$

$$V_{external}(\overrightarrow{R}) = \sum_{nonbonded} \left(\varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_D r_{ij}} \right) \tag{2.3}$$

with the three dimensional structure $\overrightarrow{R} = \{\overrightarrow{r_i} | i = 1 \ldots N\}$. Bond length b, valance angles θ, dihedral angle χ and distance between atoms r_{ij} can be calculated for the atomic positions.

2.1.1 Bonded interactions

Interactions are typically divided into "bonded" and "non-bonded" categories. Bonded interactions model chemically covalent bonds between atoms. They are called bond fluctuations, angular bends and dihedral bends with one, two and three covalent bonds involved respectively. Other interactions

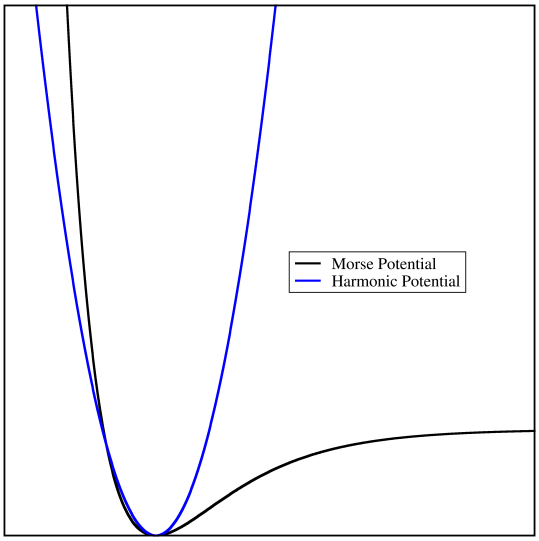


Figure 2.2: Comparison of Harmonic potential with Morse potential

are termed as non-bonded even though they might be atoms of same molecule, like, atoms of the same molecule participating in a intra-molecular hydrogen bond.

Bond stretch

Bond stretching and angle-bending terms are modelled with harmonic potentials, which keeps the bond lengths and angles near equilibrium values.

$$V_{bond} = K_b(b - b_0)^2$$

where b_0 is the equilibrium bond length and K_b is the force constants.

It is important to note that this is an approximation to the real bond stretching potential and that for large deviations from b_0 the harmonic approximation no longer holds true. For situations where the bond lengths may deviate far from b_0 or to accurately calculate molecular structures and vibrational frequencies it is necessary to go beyond the harmonic approximation and include higher order terms. Alternatively the more realistic Morse potential can be used for for little increase in complexity (Levitt, 1982).

$$V_{bond} = D\left(1 - e^{-\alpha(b-b_0)}\right)^2$$

where D is the dissociation energy and $\alpha = \sqrt{\frac{k}{2D}}$. The two functions are shown in Figure 2.2.

Typically molecular dynamics or Monte Carlo simulations are performed in the vicinity of room temperature, the harmonic energy surface represents the bond and angle distortions accurately. Thus

the use of harmonic terms is sufficient for the conditions under which biological computations are performed.

Angle bend

Angle bending terms are generally treated with a harmonic potential, which keeps the angles near their equilibrium values.

$$V_{angle} = K_\theta(\theta - \theta_0)^2$$

where θ_0 is the equilibrium bond angle and K_θ is the force constant.

The energy needed to distort an angle away from its equilibrium value is much lower than that needed to distort a bond length, therefore bond angle bending force constants tend to be smaller than those for bond stretching. As with the bond stretching potential, the accuracy can be improved by including higher order terms.

Torsional term

Dihedral or torsional angles represent the rotations about a bond, leading to changes in the relative postion of the first and fourth atom with the bond involving the second and third atom. These terms are different than bond stretching or angle bend and are oscillatory in nature, requiring the use of a periodic function to model them. Often used is the functional form

$$V_{tortional} = K_\chi[1 + cos(n\chi - \delta)]$$

where K_χ is the force constant, n is the periodicity of multiplicity and δ is the phase. The magnitude of K_χ is the barrier of the torsional potential. The periodicity n indicates the number of cycles per 360 °rotation about the dihedral angle and the phase δ describes the exact location of the minimum. A schematic representation of the potential is shown in Figure 2.3.

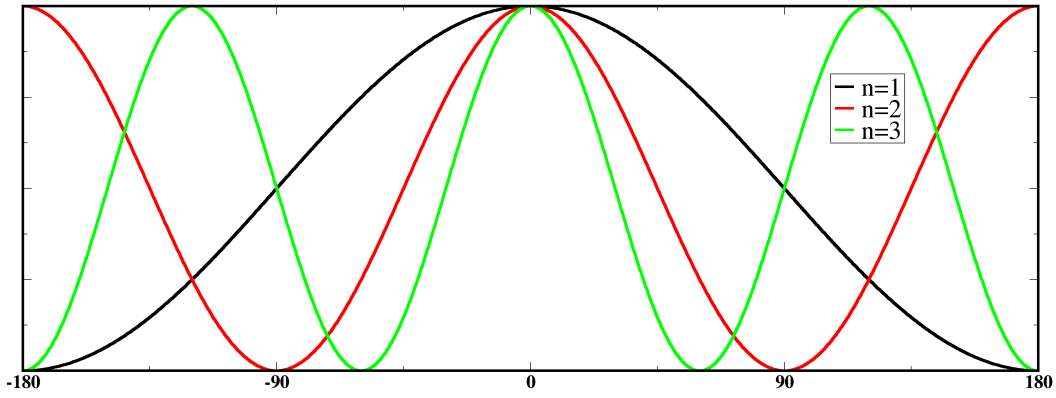


Figure 2.3: Schematic representation of torsional potential with n=1,2,3 periodicity.

A simple example to understand for this type of potential is the rotaion around the central C-C bond in a X-C-C-X configuration, like H_3C-CH_3, which changes the structure from a low energy staggered conformation to a high energy eclipsed conformation, then back to low energy conformation and so on.

2.1.2 Non-bonded interactions

Equation 2.3 describes the external or non-bonded interactions. A proper treatment of nonbonded interactions is essential for successful biomolecular computations. The mathematical model required to do so is relatively simple.

van der Waals interaction

The van der Waals interaction and static repulsion are treated with the Lennard-Jones (LJ) 6-12 potential.

$$V_{LJ} = \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right]$$

where ε_{ij} is the well depth, $R_{min,ij}$ is the minimum interaction radius , r_{ij} is the distance between atoms i and j and q_i is the partial atomic charge on atom i. The well depth ε_{ij} indicates the magnitude of the favorable London's dispersion interactions.

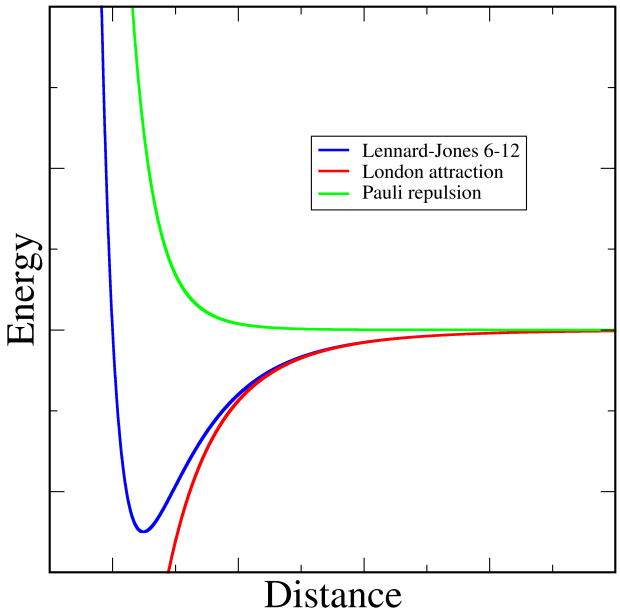


Figure 2.4: Schematic representation of Lennard-Jones 6-12 potential

The $1/r^{12}$ term represents the exchange repulsion between atoms associated with overlap of electron clouds of the individual atoms (*i.e.* Pauli's exclusion principle) as shown in Figure 2.4. The strong distance dependence of the repulsion arises from the 12^{th} power of this term. London's dispersion interactions or instantaneous dipole-induced dipole interactions are represented by the $1/r^6$ term, which is attractive.

Electrostatic Interaction

The electrostatic term involves the interaction between partial atomic charges q_i and q_j on atoms i and j separated by a distance r_{ij} in a dielectric medium, represented by the dielectric constant ε_D. The interaction is given by Coloumb's law

$$V_{coul} = \frac{q_i q_j}{\varepsilon_D r_{ij}}$$

One of the difficult parameters is ε_D which describes the effect of the medium. In case of proteins the medium is often not homogeneous. For example some water molecules might be trapped inside a protein which differ strongly in their dielectric constant (≈ 80) from the rest of the protein ($\approx$ 2-4). A distance depending dielectric constant might be used as well as taking approximate values for the interior of the protein (Herges, 2003). Another factor to take into account could be the charge fluctuations on atoms. The partial charges change according to their in environment, which would change the electrostatic contribution to total energy.

Hydrogen bonding

In proteins, a hydrogen bond appears when a donor(hydrogen) is bonded to a strong electronegative partner like oxygen in water or nitrogen in the backbone of a polypeptide chain. The positively charged hydrogen can interact with a negatively polarized partner like oxygen or nitrogen. This interaction is often modelled as pure electrostatic or as a dipole-dipole interaction. Hydrogen bonding is extremely important for protein folding as it stabilizes the formation of secondary structural elements, such as β-sheets or α-helices. Various model potentials describing the hydrogen bond are still a subject to scientific discussion. Some of the functional forms are

$$\begin{aligned} V_{HB1} &= \frac{q_i q_j}{\varepsilon r_{ij}} \\ V_{HB2} &= \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{10}} \\ V_{HB3} &= cos(\theta)\left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}\right) + (1 - cos(\theta))\left(\frac{C}{r_{ij}^{12}} - \frac{D}{r_{ij}^{10}}\right) \end{aligned}$$

where q_i is the charge on atom i, r_{ij} as the distance between atoms i and j, θ is the angle of the hydrogen bond and A, B, C and D are parameters determining the strength of the bond.

Solvent interaction

Another energetic contribution is the interaction of a polypeptide chain with its environment, mostly water (lipid bilayer for membrane proteins). Interactions with water can be treated in an explicit way by including many water molecules in the simulation. This is computationally expensive, since the number of water atoms is manyfold (as high as 50 fold) of that of the number of protein atoms itself. Another possibility is to treat the solvent implicitly, which is less accurate, but also computationally less demanding and can allow the study of reasonable size polypeptide chains. Following the work of Eisenberg and McLachen (Eisenberg and McLachlan, 1986), one can assume the contribution of an atom to the solvent energy to be proportional to its solvent accessible surface area (SASA)

$$V_{ImplicitSolvent} = \sum_{atoms} \sigma_{T_i} A_i$$

where σ_{T_i} are parameters in (cal/mol)$\cdot$ Å^{-2} which gives the energy contribution per SASA of each atom of type T_i t and A_i is the solvent accessible surface area. Comparing the accuracy of this approach with explicit solvent interactions(Skolnick and Kolinski, 1989) shows that applying implicit solvent interactions lowers the computational costs significantly with only a slight change in accuracy of the interaction. The bulk solvent is sometimes modeled with a Generalized Born approach and the multigrid method used for Coulombic pairwise particle interactions. It is faster than full explicit solvent simulations. Models like Generalized Born allow estimation of the electrostatic free energy but do not account for entropic effects arising from solvent-imposed constraints on the organization of surface-exposed regions of a macromolecule which is a major factor in the folding process of globular proteins with hydrophobic cores. Implicit solvation models may be augmented with a term that accounts for the hydrophobic effect. This effect is approximated in many studies by the solvent accessible surface area (SASA) approach.

2.2 Polarizable force fields

All force fields are based on numerous approximations and derived from different types of experimental data. One component of the force field is the electrostatic interaction, often modeled via a Coulomb pair potential between two sites assigned representative charges. The charge distribution is often determined from ab initio calculations on representative systems in vacuum and this assumes a specific environment of the molecule. Such approximations are not rigorous for treating systems with strong anisotropy, where charge fluctuations must be taken into account. Such charge fluctuations are taken into account in certain force fields (Patel and Brooks III, 2006). Various studies on fluctuating charge models (Rick et al., 1994) have shown encouraging results (Krishnan et al., 2001).

2.3 Biomolecular Simulations

Biomolecular systems are characterized by a large degree of flexibility. These atomic movements are correlated and may be essential for biological function. The biological action of molecules frequently involves large amplitude motions, called conformational changes, resulting in change in the

geometry of the molecule. As these changes occur on many different time scales, different strategies are required to answer various different questions. For example, to determine dynamical properties and to understand the processes at nanoseconds to microsecond time scale, deterministic methods can be used. In contrast, stochastic methods are better suited to treat problems such as protein tertiary structure prediction, which can be located at the free energy minimum.

There are two main approaches in performing molecular simulations: the stochastic (Monte Carlo) and the deterministic (Molecular Dynamics). Recent comparisons reveal that for polypeptide folding Monte Carlo takes ∽2-2.5 times smaller computational effort (Ulmschneider et al., 2006) than a comparable molecular dynamics study. Also comparing the time scale, a single molecular dynamics timestep of 2fs corresponds to one Monte Carlo scan (A scan is defined as the number of steps required to move, on average, each residue of the system once).

2.3.1 Monte Carlo

The stochastic approach, called Monte Carlo, is based on exploring the energy landscape by random changes in the geometry of the molecule. In this way, a large area of the configurational space is searched. A Monte Carlo simulation is composed of the following steps:

1. Specify the initial coordinates (R_0).
2. Generate new coordinates by random change to initial coordinates ($R^{'}$).
3. Compute transition probability $T(R_0, R^{'})$.
4. Generate a uniform random number RAN in range [0,1].
5. If $T(R_0, R^{'}) < RAN$, then discard the new coordinates and goto step 2.
6. Otherwise accept the new conformation and goto step 2.

The most popular realization of the Monte Carlo method for molecular systems is the Metropolis method (see flowchart in Figure 2.5):

1. Specify the initial atom coordinates
2. Select some atom i randomly and move it by a random displacement
3. Calculate the change of potential energy ΔV corresponding to this displacement.
4. if $\Delta V < 0$, accept the new coordinates and goto step 2.
5. Otherwise, if $\Delta V > 0$, select a random number RAN in the range [0,1] and:
   (a) if $e^{-\Delta V/kT} < RAN$, accept the new conformation and goto step 2.
   (b) if $e^{-\Delta V/kT} \geqslant RAN$, keep the original coordinates and goto step 2.

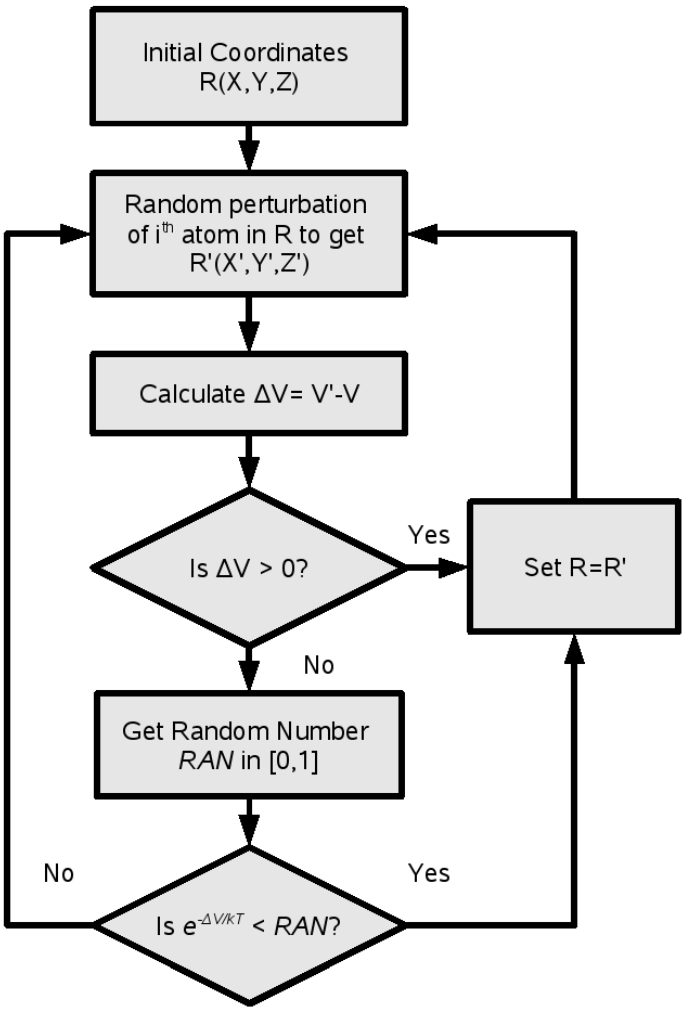


Figure 2.5: Schematic representation of Metropolis method.

In Monte Carlo simulations, the system has no "memory" between two steps, *i.e.*, the probability that the system might revert to its previous state is as probable as choosing any other state. As a result of stochastic simulation, the large number of configurations are accumulated and the energy function is calculated for each of them. This data can then be used to calculate thermodynamic properties of the system. Monte Carlo is not a deterministic method and does not offer time evolution of the system in a form suitable for viewing, but is well suited for investigating systems in certain ensembles. Monte Carlo simulations often gives rapid convergence of the calculated thermodynamic properties for small molecules (Leach, 2001).

2.3.2 Molecular Dynamics

The deterministic approach, called molecular dynamics, actually simulates the time evolution of the molecular system and provides us with a trajectory of the system. Newton's or Lagrange's equations are solved to obtain the coordinates and momenta along the simulation trajectory. Alternative approaches are based on solving Langevin's equations when the solvent is treated implicitly with added friction and noise terms corresponding to the solvent effect. The information generated from simulations can in principle be used to fully characterize the thermodynamic state of the system. In practice, most simulations are interrupted long before there is enough information to derive absolute values of thermodynamic functions (the non-ergodicity of simulation trajectory), however the differences between thermodynamic functions corresponding to different states of the system are usually computed quite reliably.

In molecular dynamics, the evolution of the molecular system is studied as a series of snapshots taken at very close time intervals (usually of the order of femtoseconds). For large molecular systems the computational complexity is enormous and supercomputers or special attached processors have to be used to perform simulations spanning long enough periods of time to be meaningful. Typical simulations of small proteins including surrounding solvent cover the range of tens to hundreds of nanoseconds, *i.e.*, they incorporate millions of elementary time steps.

Based on the potential energy function V, we can find the components, F_i of the force $\overrightarrow{F}$ acting on an atom with mass m as

$$F_i = \frac{-\partial V}{\partial x_i} \tag{2.4}$$

According to Newton's equation of motion, this force results in an acceleration $\overrightarrow{a}$

$$\overrightarrow{F} = m\overrightarrow{a} \tag{2.5}$$

From acceleration, the new velocity after time Δt can be calculated as

$$a = \frac{v' - v}{\Delta t}. \tag{2.6}$$

And from the new velocity, the new coordinates after time Δt can be calculated as

$$v' = \frac{x' - x}{\Delta t}. \tag{2.7}$$

And the cycle can be repeated until the desired simulation time is reached. A schematic flowchart is shown in Figure 2.6. It requires to evaluate forces and perform integration for every atom every timestep. Each picosecond of simulation time requires 1000 iterations of this cycle (with one femtosecond timestep). The number of evaluations can be huge in case of molecular dynamics simulations, e.g. with 50,000 atoms, each picosecond involves 100,000,000 evaluations.

To start the molecular dynamics simulation, an initial set of atom positions and atom velocities is needed. In practice, the acceptable starting state of the system is achieved by "equilibration" and "heating" runs prior to the "production" run. The initial positions of atoms are most often accepted

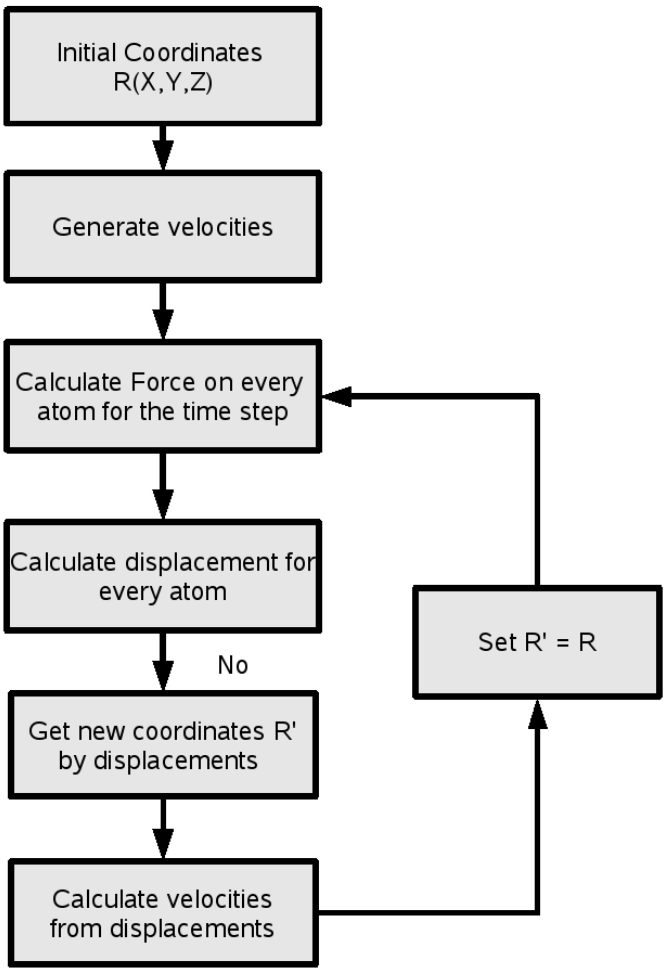


Figure 2.6: Schematic representation of molecular dynamics.

from the prior optimization of conformation with molecular mechanics. Formally, such positions correspond to the absolute zero temperature. The velocities are assigned randomly to each atom from the Maxwell distribution for some low temperature (say 20 K). The random assignment does not allocate correct velocities and the system is not at thermodynamic equilibrium. To approach the equilibrium the "equilibration" run is performed and the total kinetic energy (or temperature) of the system is monitored until it is constant. The velocities are then rescaled to correspond to some higher temperature (typically experimental temperature or room temperature), *i.e.*, the heating is performed. Then the next equilibration run follows. The absolute temperature, T, and atom velocities are related through the mean kinetic energy of the system:

$$T = \frac{2}{3Nk} \sum_{i=1}^{N} \frac{m_i |\overrightarrow{v_i}|^2}{2} \tag{2.8}$$

where N denotes the number of atoms, m_i represents the mass of i^{th} atom and k is the Boltzmann constant.

Molecular dynamics for larger molecules or systems in which solvent molecules are explicitly taken into account, is a computationally intensive task even for the most powerful supercomputers, and approximations are frequently made. The most popular is the SHAKE method which in effect freezes vibrations along covalent bonds. This method is also applied sometimes to valence angles. The major advantage of this method is not the removal of a number of degrees of freedom (*i.e.*, independent variables) from the system, but the elimination of high frequency vibrations corresponding to "hard" bond stretching interactions. In simulations of biological molecules with large conformational changes, these modes are usually of least interest, therefore their exclusion allows to increase the size of the time step, and in effect achieve a longer time range for simulations.

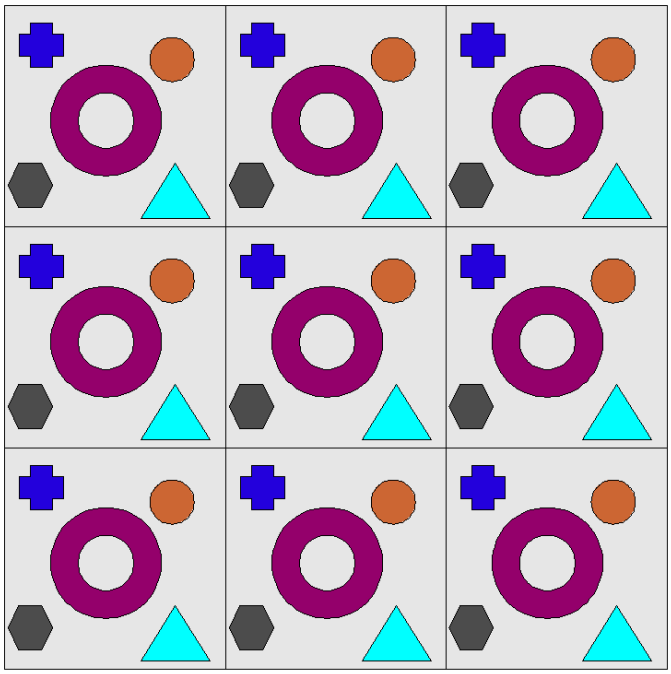


Figure 2.7: Schematic representation of periodic boundary conditions

Even supercomputers have their limitations and there is always some practical limit on the length (*i.e.*, simulated time) of the system. For situations involving solvent, the finite volume of the box in which the macromolecule and solvent are contained introduces undesirable boundary effects (as the system size of the simulation is many orders of magnitude smaller than the respective experiment). In fact, the results may depend sometimes more on the size and shape of the box than on the molecules involved.

To circumvent this difficulty arising from limited box size, periodic boundary conditions are generally used. This idea is represented in Figure 2.7. In this approach, the original box containing a solute and solvent molecules is surrounded with identical images of itself, *i.e.*, the positions and ve-

locities of corresponding particles in all of the boxes are identical. The common approach is to use a cubic or rectangular parallelepiped box, but other shapes are also possible (e.g., truncated octahedron). Using this approach, it is possible to approximate an infinite sized system. A particle (usually a solvent molecule) which escapes the box on the right side, enters it on the left side, due to periodicity.

Since molecular dynamics simulations are usually performed as an NVE (microcanonical) ensemble (*i.e.*, at constant number of particles, constant volume, and constant total energy) or an NVT (canonical) ensemble, the volume of the boxes does not change during simulation. The conservation in the number of particles is enforced by the periodicity of the lattice, e.g., a particle leaving the box on left side, enters it on the right side. There are also techniques for performing simulations in a NPT (isothermal-isobaric), and NPH (isobaric-isoenthalpic) ensembles, where the pressure conservation during simulation is achieved by squeezing or expanding box sizes. The constant temperature is usually maintained by "coupling the system to a heat bath", *i.e.*, by adding dissipative forces (usually friction forces in Langevin dynamics) to the atoms of the system which as a consequence affects their velocities.

With periodic boundary conditions one actually simulates a crystal comprised of boxes with ideally correlated atom movements. Longer simulations will be contaminated by these artificially correlated motions. The maximum length for the simulation, before artefacts start to show up, can be estimated by considering the speed of sound in water ($\approx$ 15 Å/ps at normal conditions). This means that for a cubic cell with a side of 60 Å simulations longer than 4 ps will incorporate artefacts due to the presence of images.

Choice of the Simulation Method

The choice of the simulation method depends on the system and properties under study. There has been continued discussion on the relative merits of Monte Carlo and Molecular Dynamics for biomolecules (Clarge et al., 1995; Jorgensen and Tirado-Rives, 1996).

Molecular dynamics is more appropriate when calculating time dependent quantities such as transport coefficients while Monte Carlo is most appropriate to investigate ensemble properties. The two methods also differ in their ability to explore the conformational space. Monte Carlo method can make non-physical moves that can significantly increase the capacity to explore the phase space while Molecular Dynamics might not be able to cross barriers between the conformations sufficiently often to ensure the correct statistical sampling. Thus Molecular Dynamics can be very useful in exploring local phase space whereas Monte Carlo method may be more effective for wider conformational changes (Leach, 2001; Schlick, 2002). Molecular Dynamics may requires large computational costs, but can ultimately yield detailed dynamic information such as folding pathways and rates of conformational changes.

As the two methods complement each other in their ability to explore phase space, there has been efforts to combine the two methods (Clamp et al., 1994). The simple idea attempts to combine the favorable properties of Molecular Dynamics simulations, *i.e.*, sampling phase space in directed manner guided by the shape of the gradient, with that of Monte Carlo, *i.e.*, sampling phase space

globally, to achieve better sampling of the phase space. A combination by moving some particles by Monte Carlo and others by Molecular Dynamics may be more effective than either Monte Carlo or Molecular Dynamics alone (LaBerge and Tully, 2000).

3

Free Energy Protein Forcefield

According to the thermodynamic hypothesis, the native state of many proteins is located at the global free energy minimum of its free energy surface (Anfinsen, 1973). The free energy comprises the internal energy and the entropic contributions of the system. The free energy force field PFF02 (Protein Force Field 02) is an all-atom free-energy force field* and is designed to locate the native state of various proteins at their lowest free energy. Thus it can be used to predict the native conformation of proteins following the thermodynamic hypothesis (Anfinsen, 1973). Using the combination of an effective free energy force field along with stochastic optimization algorithms (Schug et al., 2005a) the native state of a protein can be prediceted much faster than by direct simulation.

In the earlier studies PFF01 (Herges, 2003), a predecessor of PFF02, was able to predict the native state of various helical proteins at the global minimum of their free energy surface (Schug et al., 2003a, 2004b; Schug and Wenzel, 2004; Herges and Wenzel, 2005b), but could not model protein structure with β-sheet elements (Schug, 2005). A modification of the force field was thus required to have a more universal force field. The modified force field PFF02 has two additional terms in comparison to PFF01. The parameters of the other terms were not modified.

The force field models the physical interactions of a protein in an implicit solvent(water) environment at a fixed temperature of 300K.

3.1 PFF01

The protein force field PFF01 comprises four terms modelling electrostatics, hydrogen bonding, Lennard-Jones potential and solvent interaction. The solvent interaction is modelled by an implicit solvent model based on solvent accessible surface area.

The vibrational terms are not included in the force field and all bond lengths and peptide planes are kept fixed during the simulations. The only degrees of freedom available to the polypeptide chain are the dihedral angles of both mainchain and sidechains. Thus the spatial coordinates of the polypeptide chain, $\{\vec{r}\}$ specified by the internal coordinates of the dihedral angles $\{\vec{\theta}\}$. In order to increase speed

*Apolar groups of the type CH_N are modeled as united atoms with bigger radii. Modeling these hydrogens explicitely would increase the computational requirement without significantly increasing the accuracy of the force field. All other atoms are modeled explicitely in PFF01/02.

Amino acid	Potential type	Amino acid	Potential type
ALA	CME	ASP	CME CP O_2 O_2
ILE	4xCME	ARG	3xCME N_1 H CP 2x(N_1 H H)
LEU	4xCME	GLU	CME CME CP O_2 O_2
MET	CME CME S CME	HIS	CME CR N_1 H_2 CR CR N_1 H
PHE	CME 6x CR	LYS	3xCME CP N_3 3xH
PRO	3x CME	Main Chain	N_1 HM CME CP O_1
TRP	CME 3xCR N_1 H 5xCR	N-terminus	N_3 H H H CME CP O_1
VAL	3xCME	C-terminus	N_1 HM CME CP O_1 O_2
ASN	CME CP O_2 N_2 H H	SER	CP O_1 H
CYS	CME S	THR	CP CME O_1 H
GLN	2xCME CP O_2 N_2 H H	TYR	CME 6xCR O_1 H

Table 3.1: List of the different potential types according to the amino acids in PFF01/02. The list starts from the C_β atom outwards.

without significant loss of accuracy, apolar hydrogens in CH_N groups are modelled as larger united atoms.

The parameters for these terms in PFF01 force field were derived from a set of proteins. This set of proteins was selected to represent a wide range of different protein structures also including a wide range of protein folds. PFF01 was further optimized on the 36 amino acid Villin headpiece protein which had been intensively investigated by different groups using AMBER (Duan and Kollman, 1998) and ECEPP/2 (Hansmann, 2002).

Using these optimized parameters various other non-homologous helical proteins were successfully folded without any further modification of the parameters (Schug et al., 2003a, 2004b; Schug and Wenzel, 2004; Herges and Wenzel, 2005b).

Potential Types

The atoms in the polypeptide chain are classified according to their chemical characteristics. These potential types are used to obtain the values of the different force field parameters as in table 3.1.

Lennard-Jones

The van der Waals interactions are included in the force field as a Lennard-Jones 6-12 potential

$$V_{LJ}(\vec{r}) = V_0 \sum_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right]$$

where i, j represent the atoms included in the force field, r_{ij} is the distance between these atoms and R_{ij} are the Lennard-Jones radii ($R_{ij} = \sqrt{R_{ii} R_{jj}}$). The parameters for the Lennard-Jones potential were derived from a potential of mean approach to experimental data by fitting short-range (2Å - 5Å)

Potential type	R_{ii}	σ_i
CME	4.10	84
CP	4.10	-6
CR	3.28	93
N_1	3.55	-30
N_2	3.55	-15
N_3	3.55	-45
O_1	3.10	-30
O_2	3.10	-15
S	3.80	84
H	1.95	according to bound partner
HM	2.25	according to bound partner

Table 3.2: Lennard-Jones radii in Å and the solvation enthalpies in cal/(mol Å ^{2}) for potential types in PFF01/02.

radial distributions of a set of 138 different proteins [*]. The parametes corresponding of the potential types are given in Table 3.2.
In this force field the attractive part of the Lennard-Jones potential plays a very minor role compared to the repulsive part. The repulsive part prohibits clashing of atoms according to the Pauli-principle.

Electrostatics

The electrostatic interaction in PFF01/02 is split into main chain and side chain contributions and uses a model with group-specific dielectric constants. The electrostatic contribution can be written as

$$V_{ele}(\vec{r}) = V_{main}(\vec{r}) + V_{side}(\vec{r}) = \sum_{ij} \frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}}$$

where i, j represent the atoms included in the force field, q_i and q_j are the corresponding partial charges, r_{ij} is the distance between these atoms and $\varepsilon_{g(i)g(j)}$ are group-specific dielectric constants.

The group specific dielectric constants represent the characteristics of the atoms as being part of different amino acids and takes their specific partial charges, orientation or accessibility to the solvent into account. This is a strong approximation to the real environment, as only the interacting amino acids and not the complete environment is taken into consideration. The parameters for $g(i)$ and $g(j)$ are given in Table 3.4, the parameters for $\varepsilon_{g(i)g(j)} = \varepsilon_{g(j)g(i)}$ are given in Table 3.3. This parameterization excludes some parts or even complete sidechains (like PHE, GLY, MET, PRO) from contributions to the electrostatics.

The parameters for $g(i) = 1,2$ are used to describe the hydrogen bonding for the main chain as dipole-dipole interaction and constitute the biggest contribution from electrostatics. $g(i) = 3,4,5$ describe interactions of the partially charged OH, CO and NH_2 groups of the (ASN,GLN, SER, THR,

[*]These proteins are believed to represent a wide span of different folds(Avbelj and Moult, 1995).

g	1	2	3	4	5	6
1	0.375731	0.375731	0.000000	0.143396	0.143396	0.043222
2		0.375731	0.161852	0.143396	0.143396	0.031012
3			0.000000	0.000000	0.161852	0.045452
4				0.143396	0.143396	0.043222
5					0.143396	0.031012
6						0.025000

Table 3.3: Parameters for the inverse group-specific di-electrical constants $\varepsilon^{-1}_{g(i)g(j)} = \varepsilon^{-1}_{g(j)g(i)}$.

TRP)-sidechains, which are smaller in their contributions. The interaction of the charged COO^- and $NH_x^{(+)}$ of (ASP, GLU, ARG, LYS, HIS, TRP) are the smallest contributions to the electrostatic interaction.

The electrostatic contributions of the sidechains contribute only in minor quantities to the total free energy of the protein.

Hydrogen Bonding

Hydrogen bonding is a very vital contribution in protein folding (Berg et al., 2001), especially important for the formation of secondary structure in proteins. The experimental measurements of the strength of hydrogen bonding in proteins vary between -2.8 kcal/mol to +1.9 kcal/mol (Avbelj, 1992; McDonald and Thornton, 1994). Hydrogen bonding can be modeled partly by electrostatics and partly by Lennard-Jones interactions. Such approach is used in some versions of CHARMM or AMBER force fields. However in PFF01/02 hydrogen bonding and solvent interaction are considered the two major contributions to protein folding and thus special emphasis is placed to include some quantum-mechanical effects which are not modeled by the pure electrostatics.

Considering only the dipole-dipole interaction of the amino- and carboxyl groups of the main-chain, longrange interaction are overemphazised due to cooperative effects

$$V_{hydrogen-ij-dipole} = \frac{0.38 \cdot 0.28 e^2}{4\pi\varepsilon\varepsilon_0}\left(\frac{1}{r_{C_iH_j}} - \frac{1}{r_{C_iN_j}} - \frac{1}{r_{O_iH_j}} + \frac{1}{r_{O_iN_j}}\right)$$

(where i, j counts the amino acids with i belonging to the carboxyl- and j the amino group, e equals one elementary charge, $r_{X_iY_j}$ gives the distance of the atoms X from amino acid i and Y from amino acid j). This cooperative effect gets stronger for longer helices. Therefore an additional short-ranged corrective term for hydrogen bonding was included. It considers the alignment of the hydrogen bond with respect to the donor and acceptor groups (Sippl et al., 1984). The hydrogen bonding term can thus be written as

$$V_{hb} = \lambda V_{hydrogen-ij-dipole} + (1-\lambda)V_{corr}$$

where λ gives the strength of correction between $[0,1]$ with $\lambda = 1$ meaning that the hydrogen bonding is modelled by pure dipole-dipole interaction. In PFF01/02 the value of λ is 0.75. The correction

Group	Atoms	Potential	g	Group	Atoms	Potential	g
Main chain	N	n1	1	HIS	CB	cme	6
Main chain	HN	hn	1	HIS	CG, CD2, CE1	cr	6
Main chain	C	co	2	HIS	ND1, NE2	n1	6
Main chain	CO	o1	2	HIS	HD1, HE2	h	6
ASN	CG	cp	5	SER	CB	cme	3
ASN	OD1	o2	5	SER	OG	o1	3
ASN	ND2	n2	4	SER	HOG	h	3
ASN	HNA, HNB	h	4	ARG	CD	cme	6
ASP	CB	cme	6	ARG	NE	n1	6
ASP	CG	cp	6	ARG	HNE, HHA,	h	6
ASP	OD1, OD2	o2	6	ARG	HHB, HHC, HHD	h	6
GLN	CD	cp	5	ARG	CZ	cp	6
GLN	OE1	o2	5	ARG	NH1, NH2	n1	6
GLN	NE2	n2	4	THR	CB	cme	3
GLN	HNA, HNB	h	4	THR	OG1	o1	3
GLU	C,CDG	cme	6	THR	HOG	h	3
GLU	OE1, OE2	o2	6	TYR	CZ	cr	3
LYS	CD	cme	6	TYR	OH	o1	3
LYS	CE	cp	6	TYR	HOH	h	3
LYS	NZ	n3	6	TRP	NE1	n1	6
LYS	HZA, HZB, HZC	h	6	TRP	HNE	h	6

Table 3.4: The parameters for g according to the atoms of the different amino acids. Please note that for all other atoms not listed above $g = 0$.

term V_{corr} used in PFF01/02 is

$$V_{corr} = V_0 \sum_{ij} R(r_{H_i O_j}) \Lambda(\alpha_{ij}, \beta_{ij})$$

where $V_0 = -2.12$ kcal/(mol Å), α is the NHO angle, β the angle between the CO and NH-dipoles, $R(r)$ gives the radial and $\Lambda(\alpha)$ the angular dependence to the correction potential. $R(r)$ and $\Lambda(\alpha,\beta)$ are defined as

$$\begin{aligned} R(r) &= s_{2.4,0.075}(r) \\ \Lambda(\alpha,\beta) &= s_{45,5}(\alpha) s_{40,5}(\beta) s_{1.5,0.05}\left(\sqrt{\frac{\alpha}{30}^2 + \frac{\beta}{24}^2}\right)^2 \text{ where} \\ s_{A,B}(x) &= \frac{1}{2}\left(1 - tanh\left(\frac{x-A}{B}\right)\right) \end{aligned}$$

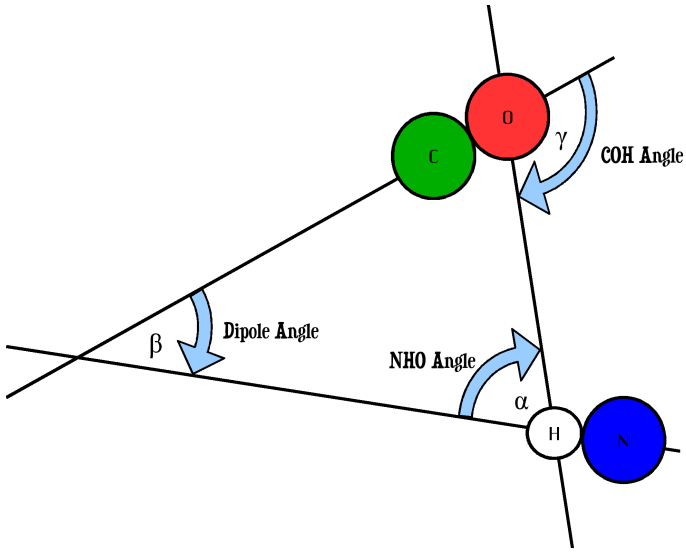


Figure 3.1: Definition of the angles α, β, γ occurring in hydrogen bonding.

Solvation effect

Since PFF01 is a free energy force field for proteins the entropic effect of the solvent needs to be incorporated. This effect is modelled with an implicit solvent model. Implicit solvent model means that the water molecules are not treated explicitly in the simulations, instead the effect of water molecules and the resulting hydrogen bonding between water molecules and protein is included in an averaged way. Thus the contribution of the entropy of the protein and the interactions of water with protein is slightly less accurate than in force fields that include explicit solvent. However this simplification saves large computational times in the simulation of the protein and is essential to obtain an estimate of the free energy of a protein conformation.

In order to estimate the contribution of solvent effect on protein in water, different physical/chemical properties and surface interaction with water are considered for the atoms. On the surface there are two different kind of interactions that are important:

- hydrophobicity, entropy of the water molecules
- entropic contributions from configurational entropy of the protein, especially from the sidechains

The entropy of water molecules contributes towards the total free energy of the solvent. The entropic contributions of the protein charge solvation comes from its configurational entropy. On the surface of the protein sidechains are less restricted in movement when compared to the inner part because of their dense packing. Therefore when a sidechain is buried inside the protein from its surface, there is a loss in configurational entropy. Since the mainchain is much more restricted in movement, the contribution from the mainchain to configurational entropy is less significant.

From the work of (Eisenberg and McLachlan, 1986) which are widely used in biophysics we can consider:

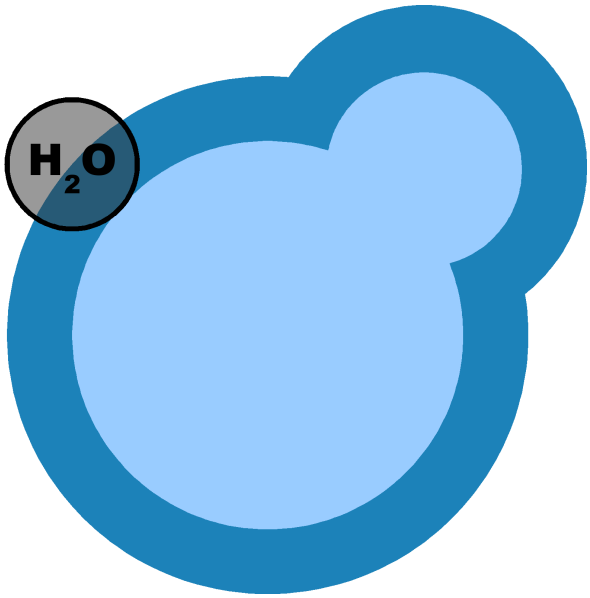


Figure 3.2: Schematics representing the calculations of the SASA-surface of a protein by rolling a water-sphere of 1.4 Å radius over two atoms. Each point on the surface belongs to the closest atom and contributes to its SASA-surface.

- transfer energy of each atom is proportional to its surface exposed to water
- transfer energy of an amino acid is the sum of the transfer energies of the individual atoms

In PFF01/02 solvent interactions are modelled by calculating the Solvent Accessible Surface Area (SASA) of each atom of the protein (Lee and Richards, 1971). In order to calculate SASA, water is considered as sphere with the radius of 1.4 Å and rolled over the protein surface defined by the Lennard-Jones radii. The area is defined by the the surface spanned by the center of the water sphere as it is rolls over the protein. Each point of the surface is associated with the nearest atom of the protein. The definition of solvent accessible surface area is illustrated in a schematic repsentation in Figure 3.2. The surface area of the dark blue region is the solvent accessible surface area.

The contribution from solvent can thus be written as

$$\Delta F = \sum_i \sigma_{PT(i)} A(i)$$

where, i counts all atoms, $PT(i)$ is the potential type of atom i, $\sigma_{PT(i)}$ gives the atomic solvent parameter (ASP) according to the potential type and $A(i)$ gives the SASA of the atom i. The parameters σ are calculated using the above equation to fit data from experiment. These data are the transfer energies from tripeptides in the form Gly-X-Gly from water to n-octanole (Fauchere and Pliska, 1983). The solvent contribution contains hydrophobic effect, configurational entropic effect (Fauchere and Pliska, 1983) and charge solvation. N-octanole is larger than water and limits the movements of the sidechains significantly. Therefore a correction has to be made since the configurational contribution in n-octanole can be estimated to be effectively zero. Later works include further corrections for

volume of the different solvents and hydrophobicity of different proteins (Sharp et al., 1991; Casari and Sippl, 1992). The solvation parameters depend on the Lennard-Jones radii and were recalculated for PFF01 (Herges, 2003). These parameters for atomic solvation parameters are given in Table 3.2. As these parameters are measured at 300K in experiment the implicit solvent model is fixed at the physical temperature of 300K.

3.2 PFF02

PFF01 was successful in predicting the native state of various helical proteins as their global free energy minimum. There were however no predictions made for proteins with beta sheet elements. PFF01 was inherently biased toward helices and thus folded to predict helices also in β-sheet regions (Schug, 2005).

Using PFF01 we investigated three non-homologous hairpin models: a structured β-peptide (pdb code: 1K43 (Pastor et al., 2002)), a mutant peptide from the first N-terminal 17 amino acid of ubiquitin (pdb code: 1E0Q (Zerella et al., 2000)) and the hairpin of the wildtype barnase (residues 85-102, pdb code: 1A2P (Mauguen et al., 1982)) with 14, 17 and 17 amino acids respectively. For each system we performed ten independent basin hopping simulations starting from completely unfolded conformations. For all the three hairpins we find no near-native conformations in the low-energy ensemble. The conformations with lowest energy have a RMSD of 6.07, 6.20 and 5.11 Å for 1K43, 1E0Q and 1A2P respectively (see Figure 3.3). Near-native conformations generated independently in basin hopping simulations starting with the native conformation also had higher energy than the terminal energy of the folding simulations. In addition, most of these simulations ultimately unfold the peptides. The occurrence of non-native conformations with lower energies than the native conformations in a free-energy force field violates the thermodynamic hypothesis and points towards deficiencies of the force field.

A free-energy force field approximates the internal free energy of the peptide/protein and must therefore account for differential solvation effects between protein microstates in the folded, the partially folded and the unfolded ensemble. Entropic contributions to the hydrophobic effect, *i.e.* changes in the solvent entropy upon exposure of the aliphatic groups of the protein, are described in an implicit solvation model. In addition the electrostatic model must be adapted to account for the nontrivial screening of electrostatic interactions by the solvent.

Since all hairpin folding simulations resulted in helical conformations at the lowest free energy, a combination of energetic contributions in PFF01 must overemphasize helical content. Several studies have investigated the differences in electrostatic stabilization of β-sheet secondary structure over helical conformations resulting from the differences in the alignment of the backbone dipoles in both types of conformations (Ripoll et al., 2005; Avbelj, 1992). We therefore use the 'local' correction to the backbone electrostatics (E_{local}) proposed to account for this effect. This correction can be interpreted as a modification of the short range dielectric constant / polarizability of the participating groups and is easily incorporated in the model.

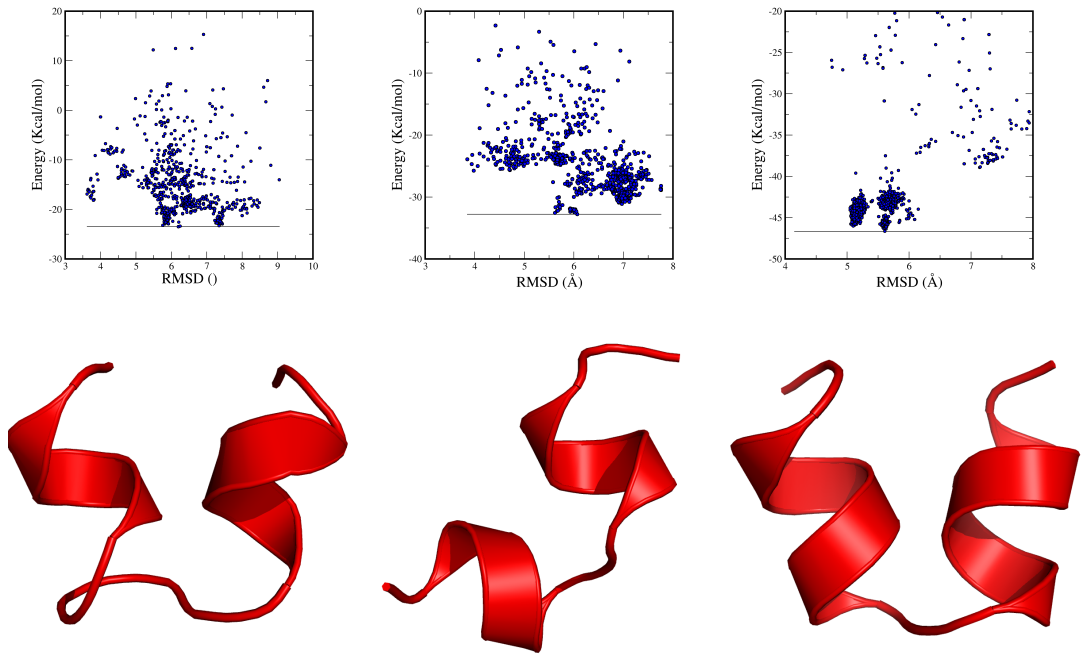


Figure 3.3: Simulations of 1E0Q, 1K43 and 1A2P in PFF01 and respective Energy vs RMS plots

Local Electrostatics

Different amino acids have been shown to have certain preferences for α-helical, β-sheet and main chain conformational states. These α-helical, β-sheet and other main chain conformational states of a residue have significantly different electrostatic energies of interaction between adjacent peptide groups (Avbelj and Moult, 1995). These energy differences are so large that electrostatics must play a role in conformational preferences. E_{local} is thus defined as the electrostatic energy of the mainchain CO and NH groups of a residue arising from interactions with the main chain CO and NH groups within that residue and with the adjoining peptide groups as shown in Figure 3.4. Thus for NH_i interactions are calculated for CO_{i-2}, NH_{i-1}, CO_i & NH_{i+1} and for CO_i the intereactions are calculated for CO_{i-1}, NH_i, CO_{i+1} & NH_{i+2} with blue and orange arrows respectively.

As Figure 3.5 shows, the nearest CO and NH dipoles in a β-strand are aligned antiparallel, whereas in an α_R-helix* these dipoles are parallel. It also suggests a mechanism by which the energy difference between the β and α_R-conformations is reduced. The electric fields produced by the parallel dipoles of peptide groups adjoining a residue in the α_R-conformation reinforce each other, resulting in strong interactions with the dipoles of water molecules or other polar protein groups nearby. If these groups are oriented favorably, the resulting energy may overcome the unfavorable backbone-backbone contribution. Conversely, for a residue in the β-conformation, the peptide dipoles are antiparallel and result in a weak electric field and thus weaker interactions with other groups in the vicinity. The degree to

*As the commonly found helix in proteins are right handed helices, the helical regions considered here are primarily right handed helices.

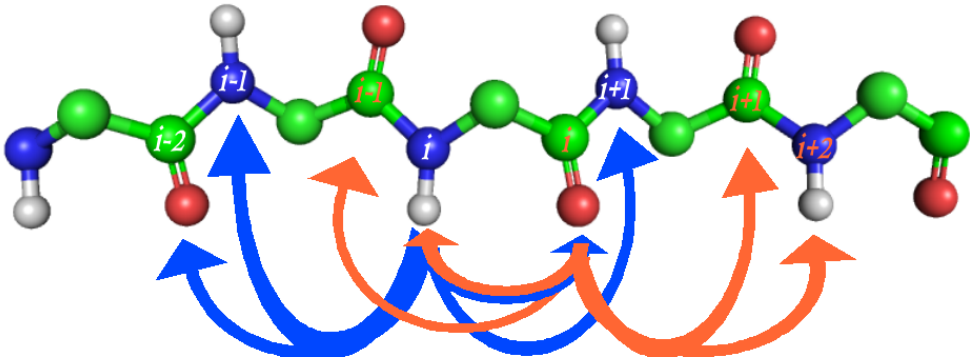


Figure 3.4: Schematic representation of interactions involved in calculation of E_{local}

which such compensation can take place also depends on the access of water and protein groups to the backbone. Such access depend partly on the type of the side chain involved.

The local electrostatics is then calculated as

$$E_{\text{local}} = \lambda_{\text{local}} \frac{332.150625 \times \zeta_A}{2} \sum_{j \in B} \sum_{i \in A} \frac{q_i q_j}{r_{ij}}$$

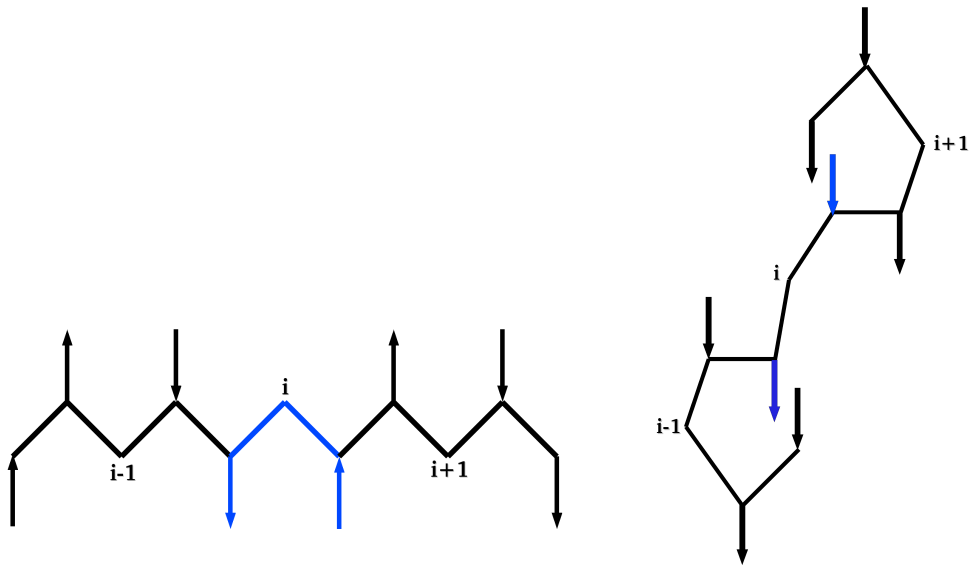
where q_i is the charge on the atom and r_{ij} is distance between the atoms. PRO group is considered equivalent to the NH group. The parameter ζ is amino acid specific parameter and the values of these parameters are given in Table3.5

Name	ζ	Name	ζ	Name	ζ	Name	ζ
GLY	0.12	PHE	0.37	SER	0.17	HIS	0.21
ALA	0.17	PRO	0.00	THR	0.18	ASP	-0.01
VAL	0.40	MET	0.34	ASN	0.11	GLU	0.11
ILE	0.43	TRP	0.21	GLN	0.21	LYS	0.19
LEU	0.29	CYS	0.23	TYR	0.28	ARG	0.22

Table 3.5: Amino acid specific parameters for local electrostatic interaction (Avbelj and Moult, 1995)

Simulations with E_{local}

Using this model for E_{local} (with $\lambda_{\text{local}} = 1$) we repeated the folding simulations for the three hairpins. In this model 1K43 folded into a near-native conformation with a RMSD of 2.8 Å, but the five of ten simulations result in helical conformations with energy differences that are only 0.5-1.2 kcal/mol higher than their misfolded conformations. The RMSD of the lowest energy structure of the other two peptides was 7.14 Å and 5.12 Å for 1E0Q and 1A2P respectively. Many conformations with backbone hydrogen bonding with beta-sheet topology emerged in the simulations but these conformations



(a) β-sheet (b) α-helix

Figure 3.5: Dipole arrangement of a residue with its adjoining residues in helix and sheet conformations

are energetically higher than the helical conformation. However, the energetic difference between the misfolded helical structures and the near-native hairpin conformations is significantly reduced in comparison to PFF01.

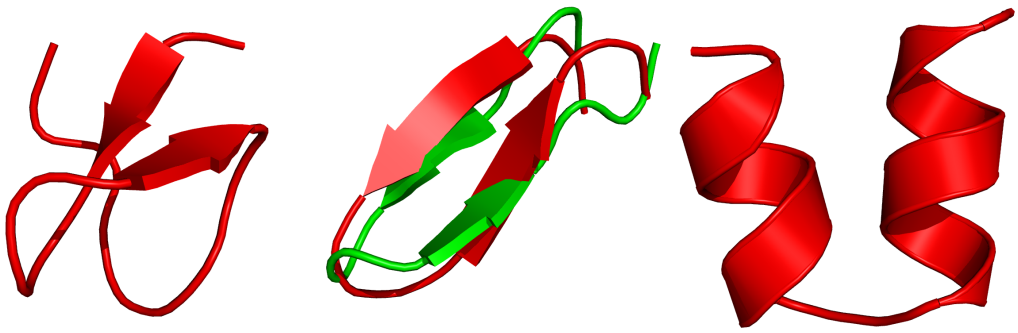


Figure 3.6: Lowest energy conformations from simulations of 1E0Q, 1K43 and 1A2P in PFF01 with E_{local} and $\lambda_{\text{local}} = 1$. Overlay with native conformation(green) is only shown for 1K43.

Several studies have investigated the impact of dynamic flexibility on backbone propensity of alpha-helix and beta-sheet proteins, suggesting a larger flexibility of beta-sheet conformation. Free-energy force fields approximate the internal free-energy of the peptide, but cannot directly account for backbone conformational entropy, because only a single backbone conformation is considered. Indeed, even in the presence of the local correction, the beta-peptides fold into extremely compact tightly packed conformations (see Figure 3.6), some of which even develop backbone hydrogen bonding with

beta-sheet topology. We hypothesize that a small free-energy correction resulting from a combination of this effects and other torsional components might be required to stabilize the beta-hairpin conformations in a free-energy force field. We have therefore incorporated backbone torsional potential, which accounts for such an entropic bias.

Torsional term

The torsional term is included as an angle dependent term to introduce stabilization of the beta sheet regions of a protein. The β region of the Ramachandran plot is around $\phi = -110°$and $\psi = 130°$as shown in the right panel of Figure3.7. This term stabilizes β-sheets by providing a weak potential when the angles are in around the above mentioned region of the Ramachandran Plot.

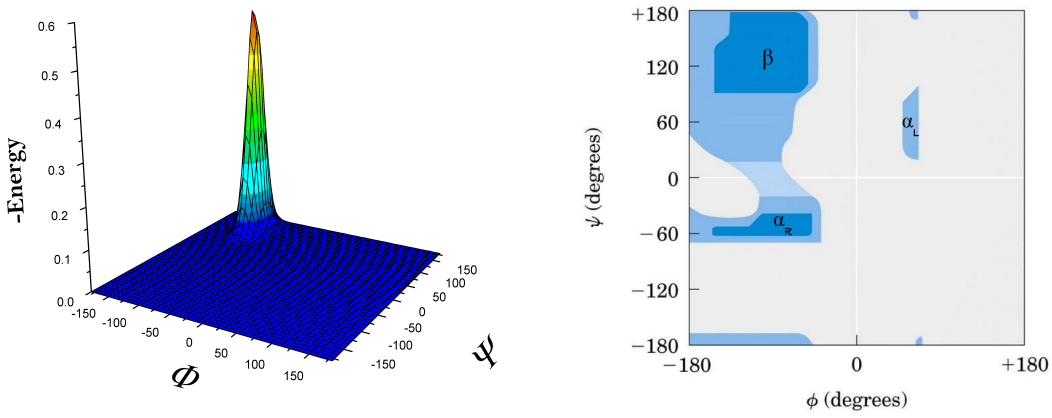


Figure 3.7: Energy contribution of E_{tor}

The torsional term has a functional form of

$$E_{\text{tor}} = \lambda_{\text{tor}} \sum_i e^{\gamma_\phi(\phi_i-\phi_0)^2+\gamma_\psi(\psi_i-\psi_0)^2}$$

for all amino acids except proline and glycine. The dihedral angles ϕ and ψ are defined in Figure 1.5(b). For proline and glycine $E_{\text{tor}} = 0$. ϕ_i and ψ_i are the backbone dihedral angles of amino acid i. We used $\phi_0 = -110°$, $\psi_0 = 130°$, $\gamma_\phi = 5 \times 10^{-3}$ deg^{-2} and $\gamma_\psi = 1.25 \times 10^{-3}$deg^{-2}.

Simulations with E_{tor} and E_{local}

Again we conducted ten independent folding simulations with E_{tor} and E_{local} (with $\lambda_{\text{local}} = 1$ and $\lambda_{\text{tor}} = 0.6$ Kcal/mol). Now all three peptides folded close to their respective native conformations(see Figure 3.8). The RMSD's of the lowest energy structure was 2.67 Å , 3.47 Å and 2.53 Å for 1K43, 1E0Q and 1A2P respectively.

We find that all four native backbone hydrogen bonds are reproduced for 1K43 and two out of three native backbone hydrogen bonds are reproduced for 1A2P as shown in Table 3.6. Even when

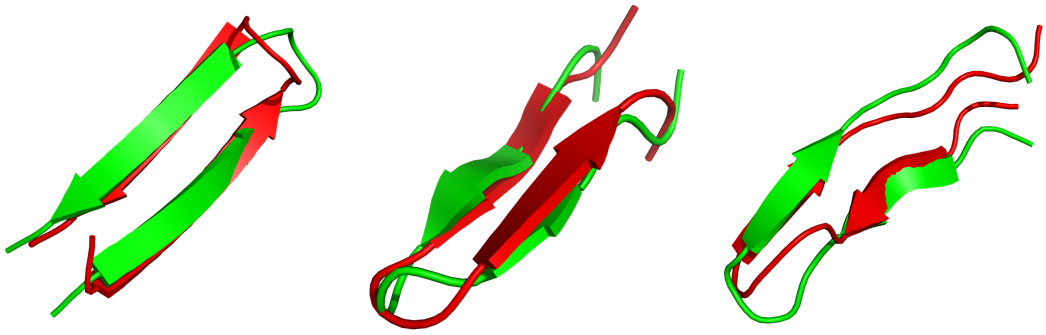


Figure 3.8: Overlay of the lowest energy conformations found for 1E0Q, 1K43 and 1A2P from left to right. Green indicates the native conformations while red is the lowest energy conformation.

Hydrogen bond	Nat	Pred
07 SER HN→ 11 LEU O	X	X
10 TRP HN→ 07 SER O	X	
13 TYR HN→ 05 LEU O	X	X

Hydrogen bond	Nat	Pred
04 TRP HN→ 11 TYR O	X	X
06 TYR HN→ 09 ILE O	X	X
09 ILE HN→ 06 TYR O	X	X
11 TYR HN→ 04 TRP O	X	X

Table 3.6: Native Backbone hydrogen bonds (left: 1A2P; right: 1K43) between native and predicted conformations. *X* represents the presence of the bond.

the lowest energy state is a β-hairpin in 1E0Q, none of the only two hydrogen bonds of 1E0Q are reproduced which is the reason for the larger deviation in RMSD.

These observations demonstrate that PFF02 can predictively and reproducibly fold beta-hairpins. The size of the entropic correction for beta-hairpin stabilization is small, favoring β-sheet conformations over α-helices by approximately 0.3 kcal/mol per amino acid for which such a difference occurs. By varying the prefactor of E_{tor} for the decoy sets generated in the all hairpin folding simulations, we find that values below 0.2 kcal/mol are insufficient to generate hairpin conformations for all peptides, while larger values destabilize helical proteins.

Thus, there are two new terms in the modified force field PFF02 (Verma and Wenzel, 2006c), namely E_{local} and E_{tor}. All the other terms, along with their parameters are kept at their original values.

4

Decoy sets in PFF02

In this chapter we show that the correction introduced in PFF02 do not destabilize the helical proteins and study its selectivity for various proteins. The accuracy and predictivity of free-energy protein force fields can be investigated using decoy sets (Park and Levitt, 1996), a method that works even for proteins that are too large or too complex to be folded from random initial conformations. In such studies a large library of protein conformations is generated to approximately span all relevant low-energy regions of the free energy surface. The conformations in the library (decoy set) are then ranked according to their energies in different force fields. If near native conformations emerge lowest in the free-energy function, the force field differentiates between native and near-native conformations. In the limit of completeness of the decoy set, which is rarely reached in practice, this test alone is sufficient to show that the force field stabilizes the native conformation of the protein against all competing metastable conformations and corresponds to the global optimum of the free-energy force field.

We investigate two decoy sets in this study. The first set of decoys consisted of all conformations generated in previous folding simulations in PFF01. The second set was taken from the decoy sets of 32 proteins generated using Rosetta (Bonneau et al., 2001). These decoy sets were then ranked according to the new force field.

4.1 PFF01 Decoys

We compiled decoy sets for the engrailed homeodomain 1ENH (∼900 decoys), the trp-cage protein 1L2Y (∼1200 decoys), designed three helical protein 2A3D (∼1000 decoys), the villin headpiece 1VII (∼4000 decoys) and bacterial ribosomal protein 1GYZ (∼1000 decoys) from earlier folding studies using PFF01.

All of these decoys sample the native ensemble as well as many competing low-energy metastable states. Because these competing metastable conformations lie just a few kcal/mol in energy above the native conformation in PFF01, this is a strong test for the predictivity of PFF02. As Figure 4.1 indicates PFF02 stabilizes near-native conformations of all investigated proteins against the decoy sets. The RMSD of the lowest energy conformation deviates by 2.33, 2.42, 2.68, 4.59 and 3.76 Å from the native conformation for the the trp-cage protein, the engrailed homeodomain protein, a

Protein	Z-score
1L2Y	-1.90
1VII	-4.56
1GYZ	-3.24
1ENH	-2.04
2A3D	-1.96

Table 4.1: Z-scores of proteins in the PFF01 decoy set

designed three helical bundle, the villin headpiece and bacterial ribosomal protein respectively. This result also indicates that the resolution of PFF02, as that of PFF01, is limited to a range of 3-4 Å , which is most likely an inherent limitation of the implicit solvent model used. The energy vs. RMSD plots and the overlay of lowest energy structures are shown in Figure 4.1.

For decoy sets generated within unbiased methods, the computation of the Z-score (the difference between energies of near-native decoys to the mean energy of the decoy set in units of its standard deviation) gives a quantitative measure of the selectivity of the force field. The Z-score is defined as

$$Z = \frac{E_{\text{ref}} - \langle E \rangle}{\sigma} \tag{4.1}$$

where E_{ref} is the reference energy, *i.e.*, the energy of the native conformation and $\langle E \rangle$ is the average energy of the decoy set and σ is the standard deviation of the decoy set. The Z-score simply measures the distance from the native state of a protein in terms of standard deviations. The lower the Z-score, the better is the discrimination between native and non-native conformations in the decoy set. The histograms show the distribution of decoys over a wide energy range (see Figure 4.1). The bars in cyan represent the distribution of near-native decoys generated from native structure and red bars represent all the decoys from the decoy set.

The Z-scores for the proteins from the decoy set are listed in Table 4.1. The average Z-score for this decoy set is -2.74 which indicates the good selectivity of PFF02 for the previously folded helical proteins.

4.2 Rosetta decoys

Encouraged by these results we explored the range of the protein which are stabilized by PFF02 using the large all atom Rosetta decoy sets (Tsai et al., 2003). For the calculation of Z-scores we generated near-native conformations for 32 proteins of the latest Rosetta decoy library. The proteins range between 32-85 amino acids in size and span all secondary structural classes. We excluded only proteins that are stabilized by transition metal clusters or other ligands as such interactions are yet to be implemented in the present force field. The resulting near-native conformations deviate 1-4 Å from the experimental conformation, except for 1am3 and 1utg, where deviations of 4.05 and 5.4 Å respectively are observed (top panel of Figure 4.2, Table 4.2 for all data). Since both of these proteins

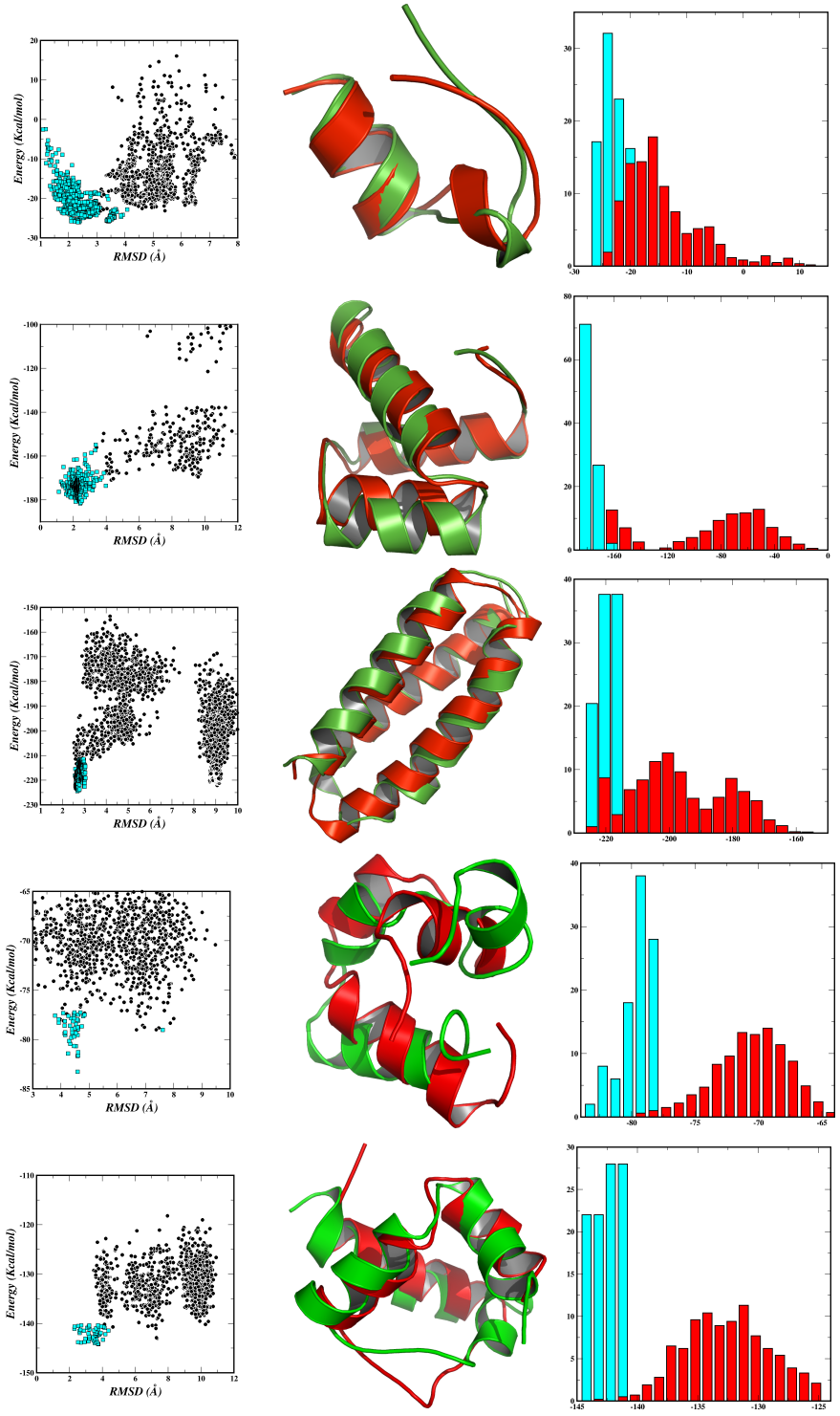


Figure 4.1: Scatter plots, overlays and Histograms for the 5 proteins; Top to bottom: 1L2Y,1 1ENH, 2A3D, 1VII, 1GYZ. For histograms Cyan indicates the distribution of decoys generated from native conformation and red indicated the distribution of all other decoys. In overlays green is native and red is lowest energy conformation

are dimeric, this difference arises because the molecules are relaxed here in isolation. The average deviation between experiment and near-native conformation in the force field for the threat of 32 proteins was 2.14 Å. The figure also indicates that there is little correlation between the size of the protein and the accuracy with which the local minimum of the force field agrees with the experimental conformation.

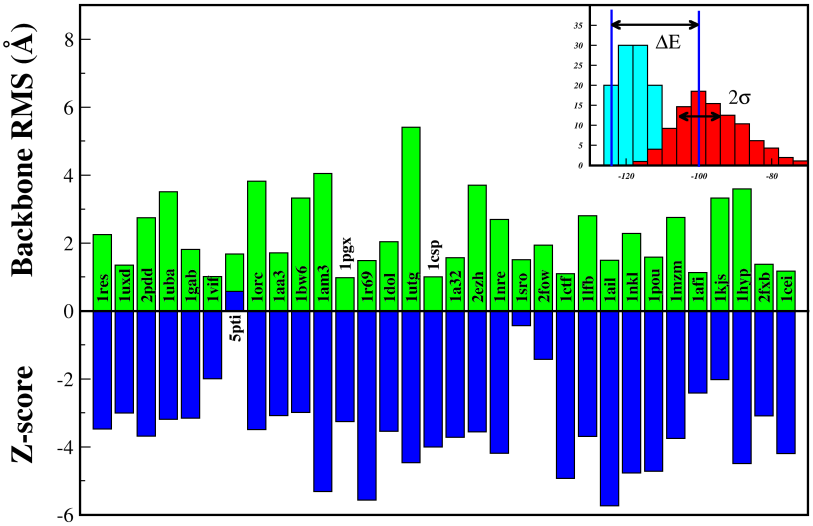


Figure 4.2: Lowest Energy RMSD and Z-scores of proteins in the Rosetta decoy set

In order to arrive at a meaningful comparison of the energies we relaxed the approximately 2000 decoys for each of the proteins in the decoy library in PFF02. This procedure maps each decoy to a local minimum of the force field of similar structure, the average change in RMSD between the starting and relaxed conformation was less than 0.02 Å. This means that the decoys are not changed in the relaxation process. Figure 4.3 shows the distribution of the 32 decoy sets over the energy range. Red bars indicate the distribution of Rosetta decoys and cyan indicate the distribution of near native decoys generated from the native conformation. Figure 4.4 shows the overlay of the lowest energy confromation (red) with the native conformation(green).

The Z-scores for 29 out of the 32 proteins in the decoy set are less than -2.0 (top panel of Figure 4.2). This indicates a good selectivity of the force field for these proteins. The average the score of -3.46 is lower than that of any previously reported alternate scoring function for the same decoy set. The average Z-score for the same set of proteins in PFF01 was -3.06 (Verma and Wenzel, 2007b). This indicates the improvement of the force field for this set of proteins which spans all kinds of secondary structural elements, with the only exception of 5PTI. Since the Rosetta decoy sets were specifically generated to span a wide range of near-native and non-native conformations for each proteins in or-

der to evaluate the selectivity and predictivity of different scoring functions these data indicate that PFF02 stabilizes near-native conformations of a large family of small and medium-size proteins of all secondary structure classes as its global optimum.

PDB ID	Z-Score$_{PFF02}$	Z-Score$_{PFF01}$	PDB ID	Z-Score$_{PFF02}$	Z-Score$_{PFF01}$
1a32	-3.72	-2.66	1nre	-4.19	-3.36
1aa3	-3.08	-2.88	1orc	-3.49	-3.31
1afi	-2.41	-3.23	1pgx	-3.26	-3.93
1ail	-5.73	-4.90	1pou	-4.72	-3.82
1am3	-5.32	-4.80	1r69	-5.57	-4.76
1bw6	-2.98	-2.94	1res	-3.47	-2.68
1cei	-4.19	-3.60	1sro	-0.43	-1.88
1csp	-4.01	-4.13	1uba	-3.19	-1.70
1ctf	-4.93	-4.38	1utg	-4.47	-3.64
1dol	-3.54	-3.18	1uxd	-3.00	-2.38
1gab	-3.16	-2.17	1vif	-2.00	-2.56
1hyp	-4.49	-4.20	2ezh	-3.56	-2.72
1kjs	-2.02	-1.32	2fow	-1.43	-1.62
1lfb	-3.69	-2.86	2fxb	-3.09	-3.14
1mzm	-3.75	-2.84	2pdd	-3.69	-2.98
1nkl	-4.77	-3.83	5pti	0.58	0.48

Table 4.2: Z-scores in PFF02 and PFF01 for Rosetta decoy set

All these results suggest that PFF02 emerges as a more universal force field which can be used for protein structure prediction. The selectivity of PFF02 is very good for a wide range of protein structures. The average Z-score for the PFF01 decoy set and the Rosetta decoy set is -2.74 and -3.56 respectively quantifying the good selectivity of near native decoys from a decoy set achieved in PFF02.

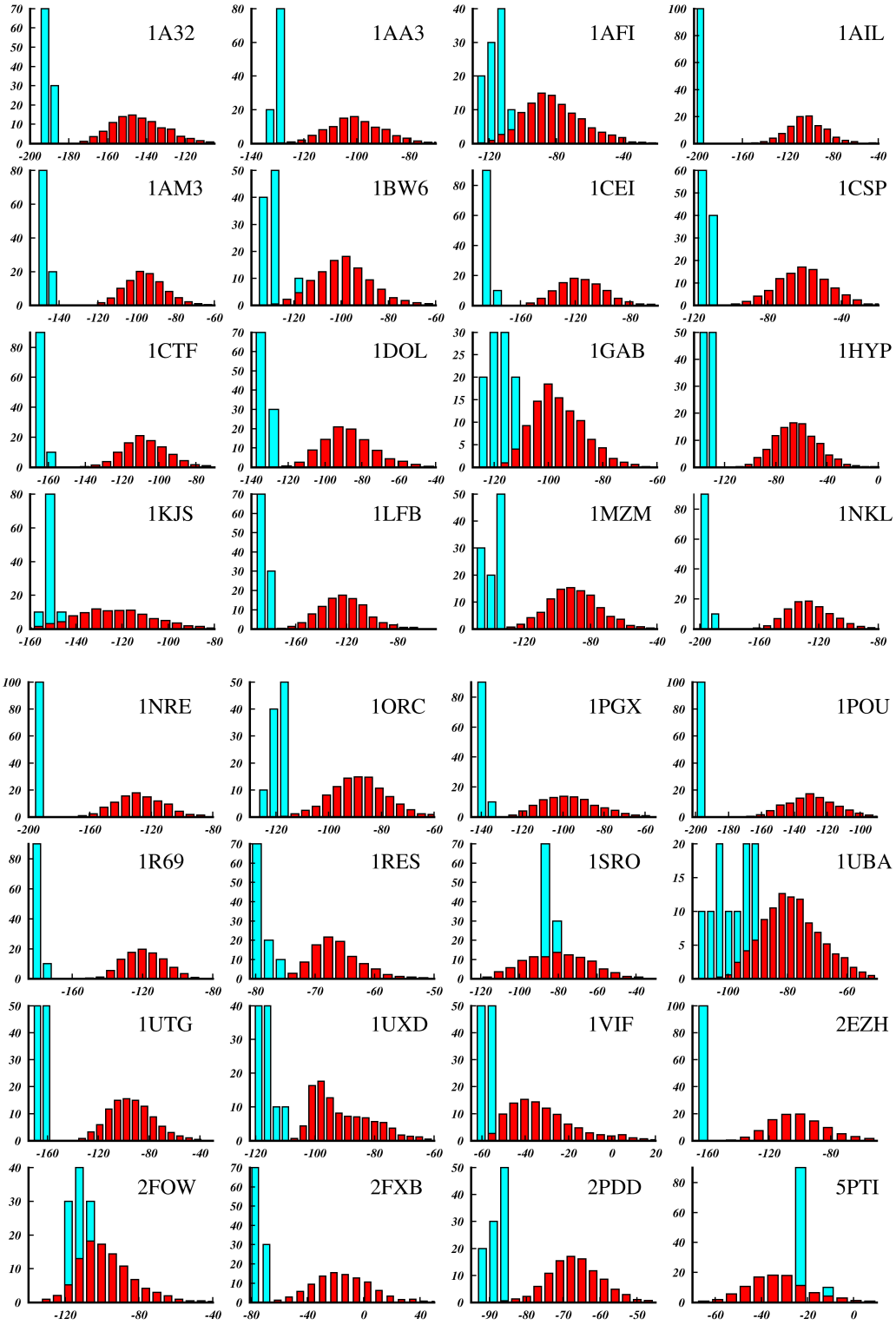


Figure 4.3: Histograms for the 32 proteins from Rosetta decoy set. Cyan indicates the distribution of decoys generated from native conformation and red indicates the distribution of Rosetta decoys.

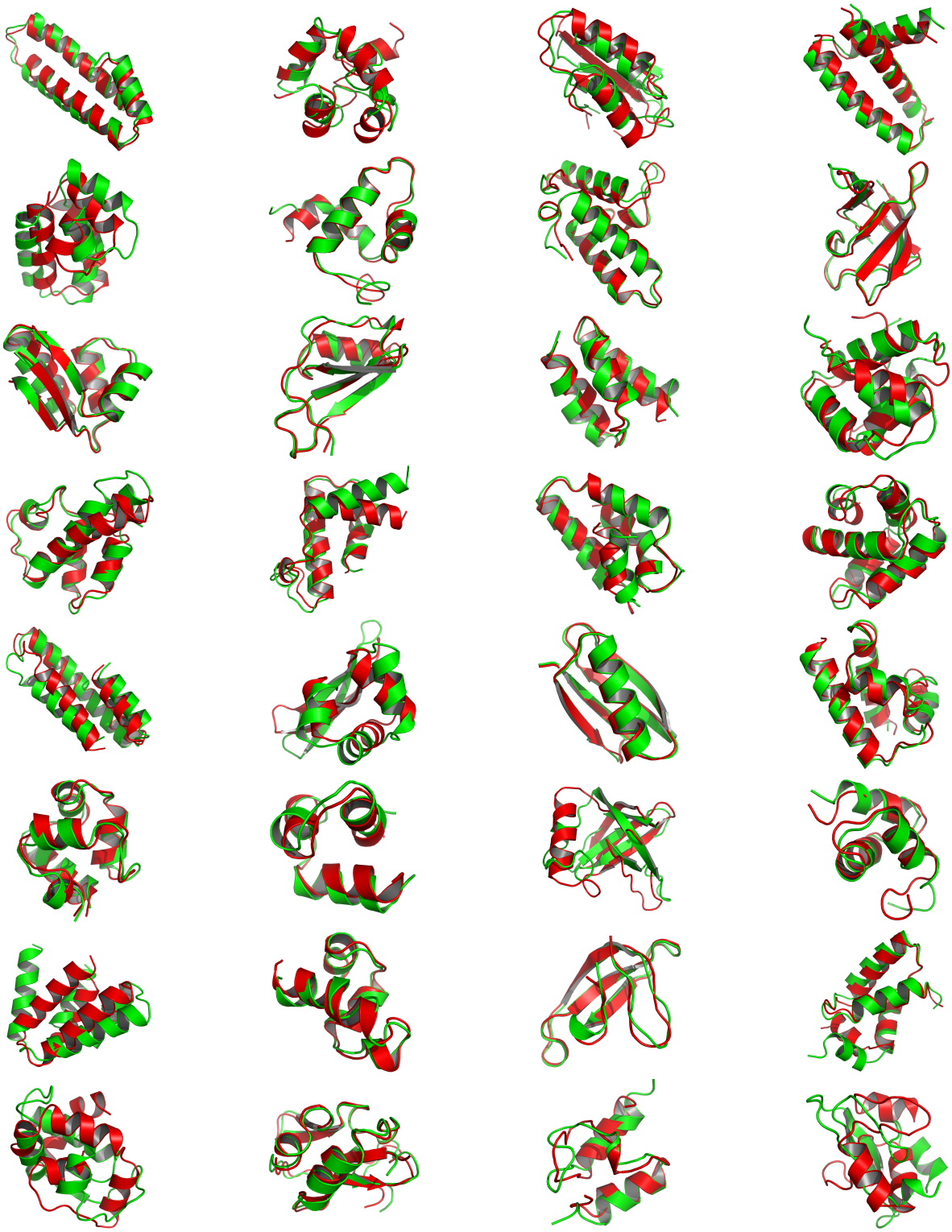


Figure 4.4: Overlay of the lowest energy conformation (red) with the native conformation (green) for the 32 proteins (corresponding to Figure 4.3)from Rosetta decoy set

5

Stochastic Methods

> "...everything that living things do can be understood in terms of the jiggling and wiggling of atoms." - Richard P Feynman, Nobel Prize in Physics, 1965

Proteins assume unique three dimensional structures after being synthesized into a linear chain of amino acids following the thermodynamic hypothesis. The thermodynamic hypothesis states that the native state of a protein corresponds to the global minimum of its free energy surface. The low-energy region of the free-energy landscape of proteins is extremely rugged due to the close packing of the atoms in the native conformation. Even if suitable free-energy force fields are available to correctly predict the native structure of a protein, the lack of optimization methods to reliably locate the associated global minima of the free-energy surface is a central bottleneck to fold proteins starting from sequence information alone. As the complexity of the free energy landscape increases with the size of a protein; this task becomes more challenging as the size of proteins increases. Furthermore, all available free-energy force fields have an inherent error, which separates the global free-energy minimum from the experimental structure. This error arises from the approximations (like point charges, fixed bond length, implicit solvation, etc.) made in the classical force fields presently used. For PFF01/02, as for most implicit solvent molecular-dynamics (MD) force fields, this deviation between the global optimum and the experimental structure is about 2-4 Å depending on the protein. Even a perfect optimization method cannot predict structures to a better resolution than the inherent resolution of the underlying force field. Thus, a root mean square deviation of less than 4-4.5 Å is considered a very good prediction of the structure of a protein.

The free-energy surfaces of many proteins have metastable conformations with energies only a few kcal/mol above the global optimum. In *de novo* protein structure prediction with free-energy models the predicted structure is selected solely on the basis of its energy in comparison to all other conformations. It is therefore important to develop techniques that can identify the global optimum of the force field to such an accuracy. Because all-atom protein folding requires substantial computational resources it is important to investigate various optimization strategies. The basin hopping technique (BHT) (Nayeem et al., 1991; Abagyan and Totrov, 1994; Wales and Doye, 1997), which has been used to fold the conserved 40-amino-acid headpiece of the HIV accessory protein (Withers-Ward et al., 2000; Herges and Wenzel, 2004), emerges as one suitable approach to perform such

simulations. In this investigation we develop a new, robust parameterization for BHT, which reduces the computational effort to fold (Verma et al., 2006). This new parameterization reduces the number of free parameters of the method.

5.1 Basin Hopping Technique

BHT (Nayeem et al., 1991) employs a relatively straightforward approach to eliminate high-energy transition states of the free-energy surface: The original free-energy surface is simplified by replacing the energy of each conformation with the energy of a nearby local minimum. In many applications the additional effort for the minimization step is more than compensated by the improved efficiency of the stochastic search. This process leads to a simplified potential on which the simulations search for the global minimum. This replacement eliminates high-energy barriers in the stochastic search that are responsible for the freezing problem in simulated annealing. A one dimensional schematic representation of BHT is shown in Figure 5.1. Every basin hopping cycle (minimization step) tries to locate a local minima and thus it simplifies the original PES (black curve) into an effective PES (blue curve) which is then searched for the global minima.

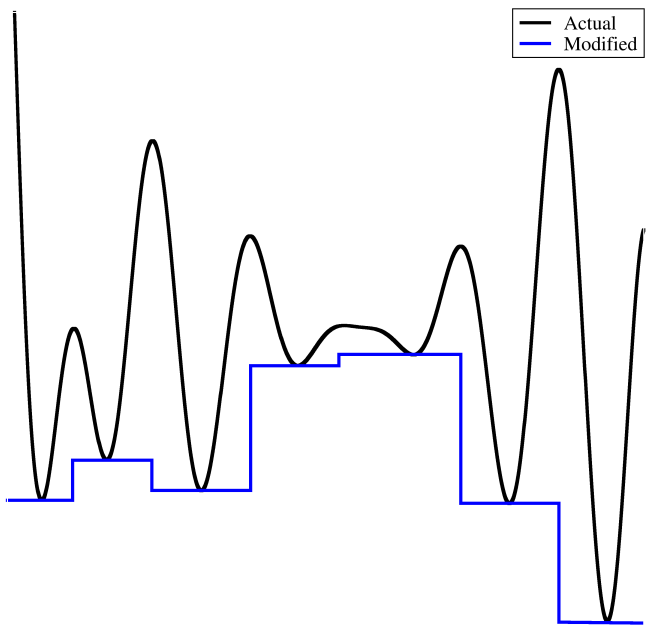


Figure 5.1: Schematic representation of Basin Hopping technique. The modified potential is obtained by replacing every point on the curve to its neared local minimum.

The basin hopping technique and derivatives have been used previously to study the potential-energy surface of model proteins and polyalanines using all-atom models (Abagyan and Totrov, 1994; Wales and Dewsbury, 2004; Mortenson and Wales, 2001; Mortenson et al., 2002). Here we replace

the gradient-based minimization step used in many prior studies with a simulated annealing run (Kirkpatrick et al., 1983), because local minimization generates only very small steps on the free energy surface of proteins. In addition the computation of gradients for the SASA (Solvent Accessible Surface Area) model is computationally prohibitive. Within each simulated annealing simulation, new configurations are accepted according to the Metropolis criterion, while the temperature is decreased geometrically from its starting to the final value.

The starting temperature and cycle length determine how far the annealing step can deviate from its starting conformation. The final temperature must be chosen small compared to typical energy differences between competing metastable conformations, to ensure convergence to a local minimum. The annealing protocol is thus parameterized by the starting temperature T_S, the final temperature T_F, and the number of steps. We investigated various choices for the numerical parameters of the method but have always used a geometric cooling schedule. At the end of one annealing cycle the new conformation is accepted if its energy difference to the current configuration was no higher than a given threshold energy ε_T, an approach recently proven optimal for certain optimization problems (Schneider et al., 1998). Throughout these studies we have used a threshold acceptance criteria of 1-3 kcal/mol.

Here we studied the choice of parameters for BHT with the folding of twenty amino acid tryptophan-cage protein (PDB code: 1L2Y). The efficiency of the local optimizer emerges as the central element in algorithms based on the basin hopping technique. Since the protein free-energy surface is very rugged for large scale moves and local gradients may not be available or computationally very expensive, stochastic local optimizers emerge as a natural choice. In order to be useful for a wide range of proteins, these methods should require as few tunable parameters as possible. Once the annealing schedule is fixed, the simulated annealing method is parameterized by its starting and final temperatures and the number of steps. The final temperature must be low enough to ensure the convergence to a local minimum and was chosen as 2K for all the simulations reported here. The starting temperature must be high enough to escape from local minima, but not too high to literally “boil away” the information in the presently accepted conformation.

Simulations with fixed starting temperature and length

We performed twenty independent basin hopping simulations with fixed starting temperatures and fixed cycle lengths. Figure 5.2 illustrates the difficulties in all-atom protein folding using this technique. The graph shows the backbone root-mean-square deviation (RMSD) of all accepted conformations at the end of the annealing cycle and their corresponding energy for twenty independent basin hopping simulations. In each simulation, an annealing cycle comprised 10,000 energy evaluations; the starting and final temperatures were 800 and 2K, respectively. The data of the simulation which converged to the lowest energy for this parameter set are shown in red. The best accepted conformation had a backbone RMSD of 2.54 Å and an energy of -25.6 kcal/mol. The figure demonstrates that there are several competing local minima, most of which are found in several simulations. Energetically these minima are very close, so that a single simulation is insufficient to determine whether the native structure of the protein was correctly predicted or not.

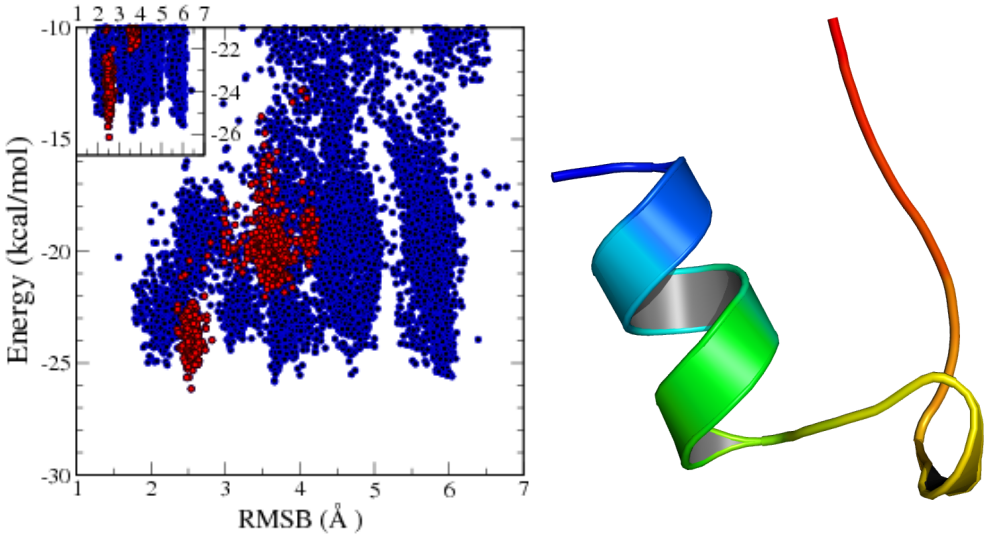


Figure 5.2: Energy vs. RMSD plot of the energy vs backbone RMSD of all accepted conformations of twenty basin hopping simulations (blue) for 1L2Y and of the simulations that found the lowest energy (red). The inset shows the low energy region with higher resolution. Each simulation comprised 100 BHT cycles of 10,000 steps each, starting at 800K and cooling to 2K.

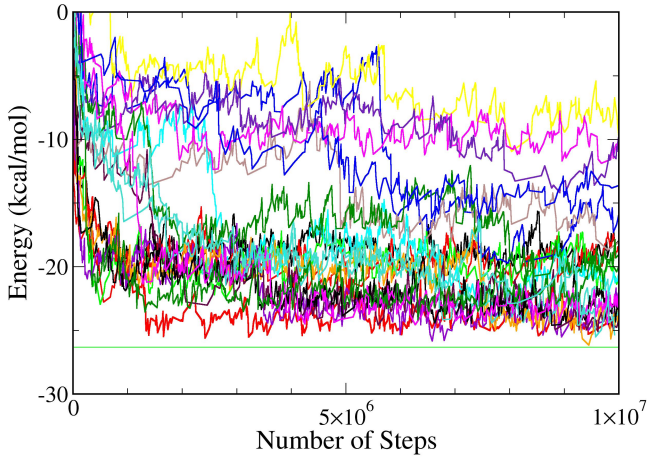


Figure 5.3: Energies of the accepted conformations at the end of each basin hopping cycle for twenty independent basin hopping simulations.

Figure 5.3 shows the dependence of the energy for all simulations on the total numerical effort. Clearly, not all simulations converge to the global optimum of the free energy surface, but the native conformation is visited independently by a number of simulations. In some simulations the optimal energy is reached quite early, while others take many cycles to converge. For this reason, the use of just the best energy of the best simulation may be misleading to judge the efficiency of different BHT parameterizations. Table 5.1 summarizes the best energies and the associated backbone RMSDs for sets of twenty simulations each. Each set of simulations used a different starting temperature. We find that nearly all simulations explore conformations very close to the native state (last three columns), which are often ranked highly in the population of final conformation. In three of five cases the best energy is associated with a conformation with a backbone RMSD of less than 3 Å to the native state; in the other two cases the second best conformation is near native. We find that very high starting temperatures are needed to ultimately converge to low energies; the best results were obtained with T_S=800 and 1000 K, respectively.

T_s (K)	E_{lowest} (Kcal/mol)	Δ_{best} (Å)	Δ_{min} (Å)	Rank
200	-25.33	4.90	2.32	(2)
400	-26.00	2.46	2.46	(1)
600	-24.96	2.63	2.63	(1)
800	-26.15	2.54	2.22	(5)
1000	-26.12	4.06	2.51	(2)

Table 5.1: Summary of simulations with fixed cycle length ($N = 10000$) and starting temperature $T_S(K)$: E_{lowest} designates the lowest energy found in twenty independent simulations after $m = 100$ iterations, and Δ_{best} is the corresponding backbone RMSD. Δ_{min} is the smallest backbone RMS deviation that was found with the lowest energy in any of the simulations, followed by the rank (in energy) of the simulation in which it was found. Total number of steps for each of the simulations were $N_{tot} = 10^6$

Since ranking of the performance of different parameterizations based on the best energies alone may mislead the interpretation, we also plot the weighted average of the sorted energies, of the best-accepted energies $E_k(t)$ as a function of time/ numerical effort t for all simulations $k = 1 - N_{sim}$.

$$E_W(t) = \frac{\sum_{k=1}^{N_{sim}} E_k/k^2}{\sum_{k=1}^{N_{sim}} 1/k^2} \tag{5.1}$$

This formula was chosen to give an unbiased measure of the success of all simulations. We must presently accept that some BHT simulations will go completely astray, but the energy difference in comparison to the best simulations allows us to identify such failures (see Figure 5.3). The weighting accounts for this fact and emphasizes the better simulations. The number and quality of low-energy conformations is a measure of the overall performance of the protocol. Note the weighted averages of top set of curves fall monotonically and slowly with increasing starting temperatures (Figure 5.4),

despite the fact that none of the twenty simulations with a starting temperature of 600 K found a very good energy. The best energies, in contrast, are dominated by the rare event of single simulation, finding a very good conformation at some point during the simulation.

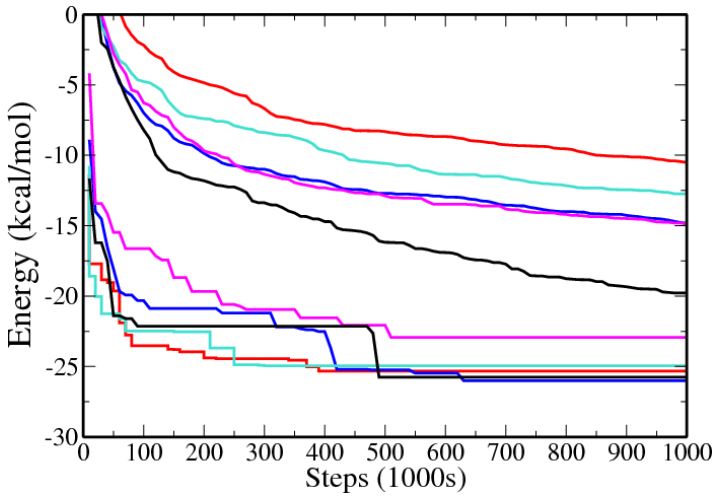


Figure 5.4: Best energy (lower set of curves) and weighted energy (upper set of curves) for the accepted conformations at the end of each basin hopping cycle for twenty independent basin hopping simulations vs. the number of energy evaluations. Red(200K), blue(400K), light blue(600K), cyan(800K) and black(1000K) show the data for different starting temperatures.

We conclude that high starting temperatures are required and will permit a significant fraction of the simulations to converge to low-energy conformations. The $T_S = 1000$ K set of simulations clearly performed best in this regard but not for the best energy. We also note that the ranking of the weighted averages is independent of time. Because we are ultimately interested in the asymptotic behavior of the method for long simulation times, such a consistency in rank simplifies the interpretation of the data obtained for finite simulation time.

While there is overall good convergence, it is striking that for all the starting temperatures there are little changes in the best energies after about 5×10^5 energy evaluations. The best set of simulations is shown in Figure 5.3, which shows that most of the energy gain occurs quite early in the simulations. This indicates that the annealing step fails to improve the energy in late cycle conformation. Indeed, the acceptance ratio of the threshold acceptance criterion drops with the number of cycles performed.

Simulations with fixed starting temperature and increasing length

To improve the performance of the algorithm without dramatically increasing its computational cost, we have increased the number of steps of the annealing simulations as a function of cycle number(m) as $N_m = N_0\sqrt{m}$. Again we performed twenty simulations for each set of parameters. The results of these simulations are summarized in Table 5.2 and the best and weighted energies as a function of time are shown in Figure 5.5. Note that the horizontal axis is the number of function evaluations, not the

cycle index, which makes both sets of data comparable. Table 5.2 demonstrates that the simulations explore an energy range that were never reached in simulations with constant cycle length. This is clearly demonstrated by the data in Figure 5.5, see, for instance, the simulations with $T_S = 800$ K blue line, which generate new optimal conformations drops in the line at step numbers 4, 5, and 6.2×10^6, while running with constant length (pink line) generate the last such jump at step 1.4×10^6 and fail to improve afterwards.

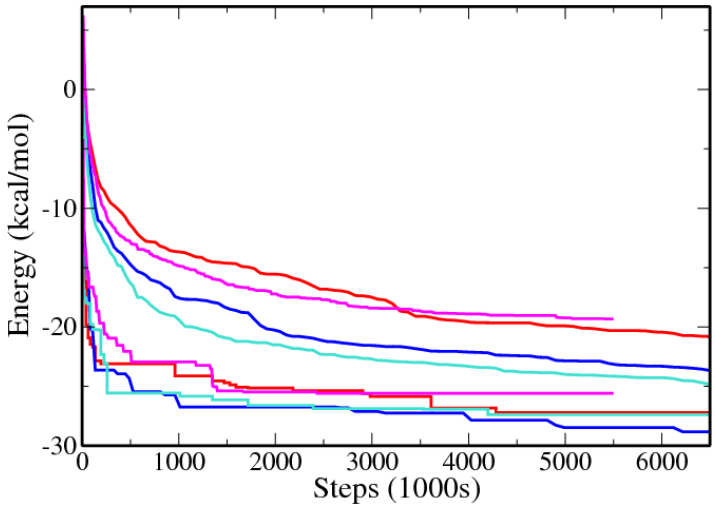


Figure 5.5: Best energy (lower set of curves) and weighted energy (upper set of curves) for the accepted conformations at the end of each basin hopping cycle for twenty independent basin hopping simulations. vs the number of energy evaluations. Red(400K), blue(600K), light blue(800K), cyan(800K) and black(1000K) show the data for different starting temperatures.

The total computational effort of a basin hopping simulation is the sum of all its constituent local optimizations. The number of steps per cycle must be chosen to balance the number of cycles in the simulation, *i.e.*, opportunities to accept a better configuration over the present starting configuration, against the improvement obtained in one cycle. High starting temperatures are required to escape from metastable states but lead to long relaxation times to anneal to competitive final conformations. A gradual increase in the cycle length permits a more thorough local search; it also essentially simplifies the method as the prefactor of the cycle length N_0 becomes a weak parameter. If the cycle length is chosen too small initially, some early cycles are wasted. Its gradual increase nevertheless ensures that ultimately long enough simulations will be performed. Monitoring the weighted average of the energy Figure 5.5 permits an assessment of the degree of convergence of the simulation set.

Simulations with random starting temperatures

Having thus essentially eliminated the cycle length as a sensitive parameter of the algorithm, we next turn to the choice of the starting temperature. When simulated annealing is used as a local optimizer the basin hopping cycle can escape only metastable minima with transition states of $O(K_B T_S)$. If T_S is

T_s (K)	E_{lowest} (Kcal/mol)	Δ_{best} (Å)	Δ_{min} (Å)	Rank
200	-24.24	3.75	3.51	(16)
400	-26.45	3.89	1.54	(7)
600	-28.50	5.19	2.17	(5)
800	-27.39	1.79	1.79	(1)
1000	-28.04	3.01	1.85	(11)

Table 5.2: Summary of simulations with increasing cycle length ($N(m) = 10000\sqrt{m}$) and fixed starting temperature $T_S(K)$: E_{lowest} designates the lowest energy found in twenty independent simulations after $m = 100$ iterations, and Δ_{best} is the corresponding backbone RMSD. Δ_{min} is the smallest backbone RMS deviation that was found with lowest energy in any of the simulations, followed by the rank (in energy) of the simulation in which it was found. Total number of steps for each of the simulations were $N_{tot} = 6.7 \times 10^6$

chosen too small, an individual simulation may never be able to escape from a deeper local minimum. The data obtained so far indicates that simulations at physiological temperatures are essentially always trapped in local metastable states, a fact also well established in molecular dynamics simulations. Choosing T_S too high, on the other hand, will lead to a rapid destruction of tertiary and even secondary structures early in the simulation; in the limit $T_S \to \infty$ each basin hopping cycle starts from a random conformation. An optimal implementation of the basin hopping technique must thus compromise between these two limits.

Little is presently known about the structure of the energy surface and the distribution of the depth of local minima. Assuming that metastable conformations exist on many energy scales, one can improve the convergence of the method by choosing the starting temperature of each individual cycle from a distribution of temperatures. This procedure was originally suggested for the exploration of glassy models in condensed matter physics which also feature ruggedness on many scales. Depending on the choice of the distribution (fixed starting temperature corresponding to a delta distribution), very high starting temperatures are sometimes chosen to permit trapped simulations to escape, while most of the computational effort is concentrated on small energy scales. For small $K_B T_S$ the basin hopping cycle is confined to a search in the neighborhood of the starting conformation. In our simulations with a fixed starting temperature of T_S=200 K initial and final conformations seldom deviate more than 1.5 Å backbone RMSD from one another.

Here we have investigated simulations where the starting temperatures were chosen from an exponential distribution $p(T_S) = exp(T_S/T_0)$. With this choice, the simulation has the chance to escape from local minima of arbitrary depth, although most of the effort is concentrated on optimization on an energy scale $O(K_B T_0)$. This approach reduces the sensitivity of the method to the choice of T_S, which will be difficult to recalibrate for large systems.

An impressive validation of the concept of random starting temperatures is demonstrated by comparing the simulations with T_0=250K (red) and T_0=750K (blue) in Figure 5.6, see also Table 5.3. Both

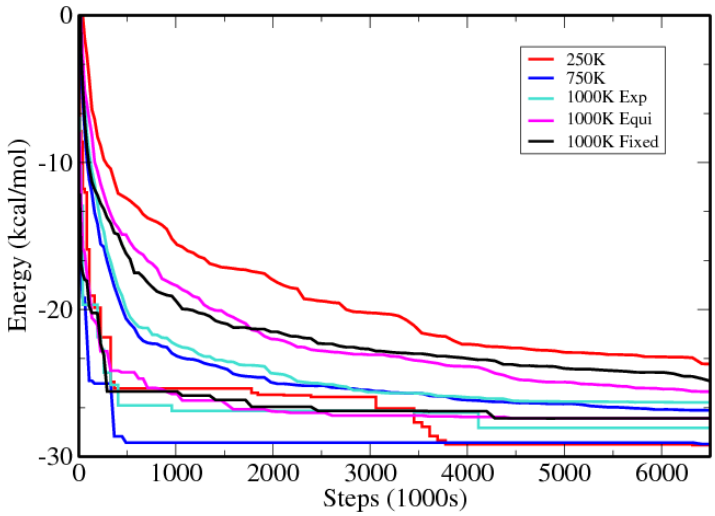


Figure 5.6: Best energy (lower set of curves) and weighted energy (upper set of curves) for the accepted conformations at the end of each basin hopping cycle for twenty independent basin hopping simulations vs. the number of energy evaluations. Red(250K), blue(750K), light blue(1000K with exponential), magenta(1000K with equidistributed) and black(1000K with fixed) show the data for different starting temperatures.

reach the same optimal energy, *i.e.*, even the low-energy simulation manages to escape from a deep local minimum. We note that the performance of the algorithm is much improved, the best energy is almost 2 kcal/mol lower than all the previously found energies. It is now reached after a mere 500,000 energy evaluations, more than an order of magnitude faster than with the initial parameterization (see Figure 5.5). We note that now the method saturates with T_0, indicating that T_0 of about 700 K is optimal for at least this protein.

The conformation with the lowest energy was obtained in a simulation with a random choice of initial starting temperatures. The lowest energy conformation had an energy of -29.28 kcal/mol and a backbone RMSD of 3.16 Å to the native conformation. The main contribution to this difference arises from the disordered tail of the protein, which bends around in the experimental conformation but not in our simulations. This conformation is illustrated in Figure 5.7, along with the C_β-C_β matrix indicating the existence of long-range native contacts. The off-diagonal dark areas in the C_β-C_β matrix indicate the crucial longrange contacts for the formation of tertiary structure.

Table 5.3 also shows data for constant cycle length, but these are inferior to the results of increasing cycle length. We have also found that simulations wherein the starting temperatures are drawn from an exponential distribution are superior to those with equidistributed starting temperatures.

Discussion

The development of reliable free-energy force fields and efficient optimization techniques offers an increasingly viable route for protein structure prediction at the all-atom level. Presently, the availability

T_0 (K)	E_{lowest} (Kcal/mol)	Δ_{best} (Å)	Δ_{min} (Å)	Rank
	increasing cycle length			
50	-26.43	4.93	1.71	(5)
100	-27.10	3.19	3.19	(1)
150	-27.18	3.75	1.97	(4)
200	-26.73	3.75	1.79	(10)
250	-29.28	3.16	1.72	(6)
500	-27.77	2.79	2.12	(2)
750	-28.90	2.86	1.78	(5)
1000	-28.80	3.02	2.18	(5)
	constant cycle length			
50	-20.90	4.08	2.45	(3)
100	-21.05	4.32	3.35	(9)
150	-22.70	1.64	1.64	(1)
250	-27.86	3.07	1.79	(2)
500	-26.38	4.01	1.80	(2)
750	-28.26	1.87	1.87	(1)
1000	-28.26	1.91	1.91	(1)

Table 5.3: Summary of simulations with random starting temperature $T_S(K)$, chosen randomly from an exponential distribution $p(T_S) \propto exp(T_S/T_0)$. We performed simulations with increasing cycle length ($N(m) = 10000\sqrt{m}$) or constant ($N = 10000$) cycle length. E_{lowest} designates the lowest energy found in twenty independent simulations after $m = 100$ iterations, and Δ_{best} is the corresponding backbone RMSD. Δ_{min} is the smallest backbone RMS deviation that was found with lowest energy in any of the simulations, followed by the rank (in energy) of the simulation in which it was found. Twenty independent simulations with $m = 670$ (constant cycle length or $m = 100$ (increasing cycle length) were performed, comprising $N_{tot} = 6.7 \times 10^6$ function evaluations each

of efficient optimization methods, rather than inaccuracies of the force field, appear to be the bottleneck towards the treatment of larger helical proteins. The free-energy optimization approach pursued here complements MD simulations because it offers a rational criterion for unbiased protein structure prediction. While sacrificing insights into the folding kinetics, the native conformation can be obtained orders of magnitude faster than by methods that require the simulation of the folding pathway. Nevertheless *de novo* protein structure prediction remains a formidable computational challenge even for moderate size proteins. Since large computational resources are required for predictive folding studies, the investigation and analysis of different stochastic optimization methods for this problem is important. In this study we have developed a new parameterization of the basin hopping technique that reaches the native conformation of the trp-cage protein about one order of magnitude faster than the original protocol. The new parameterization is less dependent on the choice of tunable parameters

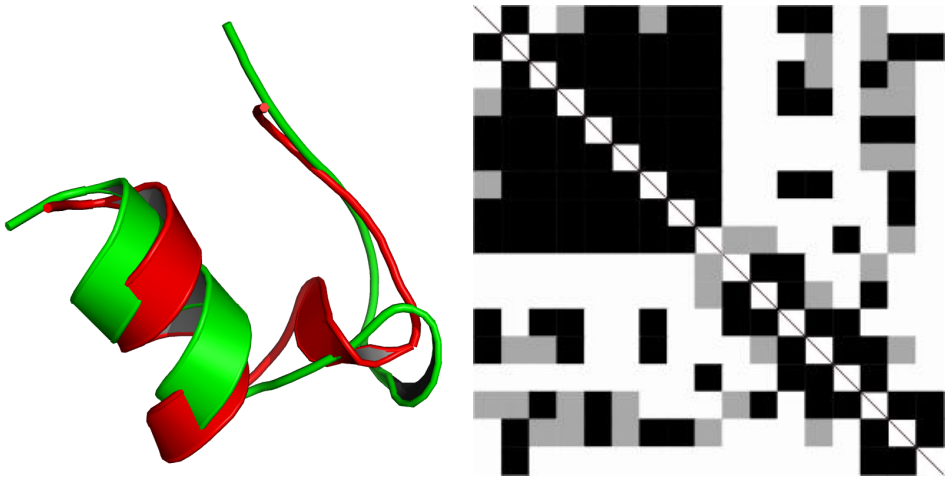


Figure 5.7: Overlay (left) and C_β-C_β distance map (right) of the folded structure and the experimental structure of the trp-cage protein(top) and the potassium channel blocker(bottom). A black(grey) box in row i and column j of the distance map indicates the difference in that the C_β-C_β distance of the native and the other structure differ by less than 1.5(2.25) Å respectively. White squares indicate larger deviations.

T_0 and cycle length and was demonstrated to be transferable to another system.

The basin hopping technique, when viewed as a single long simulation, gives up this continuity: Starting from a given conformation the search process is abandoned if no good competing conformation emerges in the next cycle and restarted from the "memory" of the process. This speeds the search process because optimal conformations are never discarded. The successive relaxation of the constraints of (i) realistic kinetics, (ii) thermodynamic equilibrium, and (iii) temporal continuity leads to increasingly better optimization methods. The results of this study offer a benchmark to calibrate and compare different optimization techniques for proteins that are readily treated with present day computational resources. The basin hopping technique emerges as a powerful yet simple workhorse for predictive all-atom *de novo* protein folding with free-energy models.

5.2 Distributed Computing

The search for efficient and predictive methods to describe the protein folding process at the all-atom level remains an important grand-computational challenge. The development of multi-teraflop architectures, such as the IBM BlueGene used in this study, has been motivated in part by the large computational requirements of such studies. Here we report the predictive all-atom folding of the forty-amino acid HIV accessory protein using an evolutionary stochastic optimization technique. We implemented the optimization method as a master-client model on an IBM BlueGene, where the algorithm scales near perfectly from 64 to 4096 processors in virtual processor mode (Verma et al., 2007). Starting from a completely extended conformation we simulated a population of sixty-four

conformations of the protein in our all-atom free-energy model PFF01. Using 2048 processors the algorithm predictively folds the protein to a near-native conformation with an RMS deviation of 3.43 Å in less then 24 hours (see Figure 5.8).

The low-energy region of the free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Suitable optimization methods must therefore be able to speed the simulation by avoiding high energy transition states, adapt large scale moves or accept unphysical intermediates. The basin hopping technique described in the previous section has proved to be a reliable workhorse for many complex optimization problems (Wales and Doye, 1997), including protein folding (Abagyan and Totrov, 1999; Mortenson and Wales, 2001; Mortenson et al., 2002; Herges and Wenzel, 2005a), but employs only one dynamical process. We have generalized this method to a population of size N which is iteratively improved by P concurrent dynamical processes. The whole population is guided towards the optimum of the free energy surface with a simple evolutionary strategy in which members of the population are drawn and then subjected to a basin hopping cycle. At the end of each cycle the resulting conformation either replaces a member of the active population or is discarded. Similar strategies, employing a conformation stack, have previously been explored in simulations of the 23 amino acid BBA5 protein (Abagyan and Totrov, 1994, 1999).

This algorithm was implemented on a distributed master-client model in which idle clients request a task from the master. The master maintains a list of open tasks comprising the active conformations of the population. The client then performs a simulated annealing simulation of specified length (N=40,000 steps) on the conformation. The simulated annealing runs used a geometric cooling schedule reducing the temperature from 1200 K to 2K.

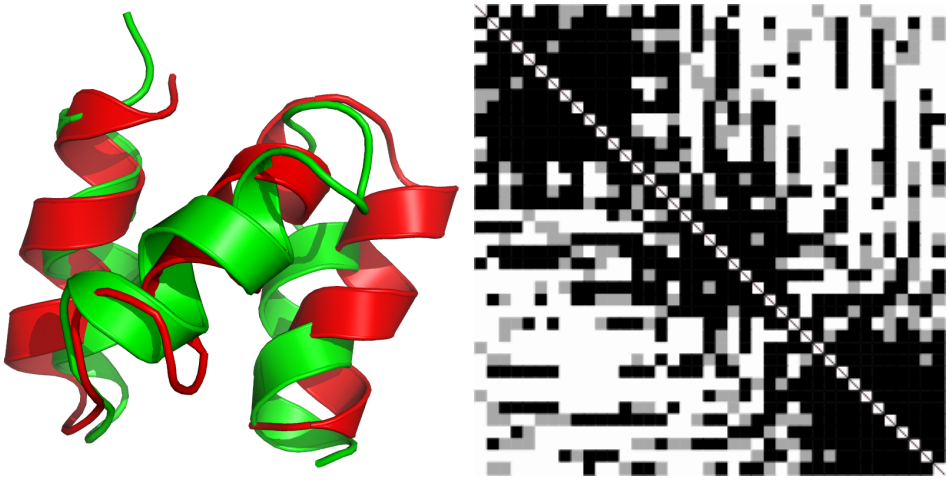


Figure 5.8: Overlay of the predicted (lowest energy, red) to native conformation (green) and the C_β-C_β distance map for 1F4I in evolutionary algorithm.

Conformations are drawn randomly according to some probability distribution from the active population. The acceptance criterion for newly generated conformations must balance the diversity of

the population against the enrichment of low-energy decoys. Since one can in principle account for the number of times a given conformation was found (not employed here), there is no need to store duplicates. We therefore accept only new conformations which are different by at least 4 Å RMSD (root mean square backbone deviation) from all members of the active population. If we find one or more members of the population within this distance, the new conformation replaces all the existing conformations if its energy is lower than the best, otherwise it is discarded. If the new conformation differs by at least the threshold from all other conformation it replaces the worst conformation of the population if it is better in total (free) energy. If a merge operation has reduced the size of the population, the energy criterion for acceptance is waived until the population size for the simulation is restored.

Scalability

For timing purposes we have performed simulations using 64, 128, 256, 512, 1024, 2048 and 4096 processors on an IBM BlueGene in virtual processor mode. Here we report data for a population size P=64 for simulations of the 40 amino acid HIV accessory protein (sequence: QEKEAIERLK ALGFEESLVI QAYFACEKNE NLAANFLLSQ, pdb-id: 1F4I) (Withers-Ward et al., 2000). As demonstrated in Figure 5.9 the algorithm scales well up to 4096 processors.

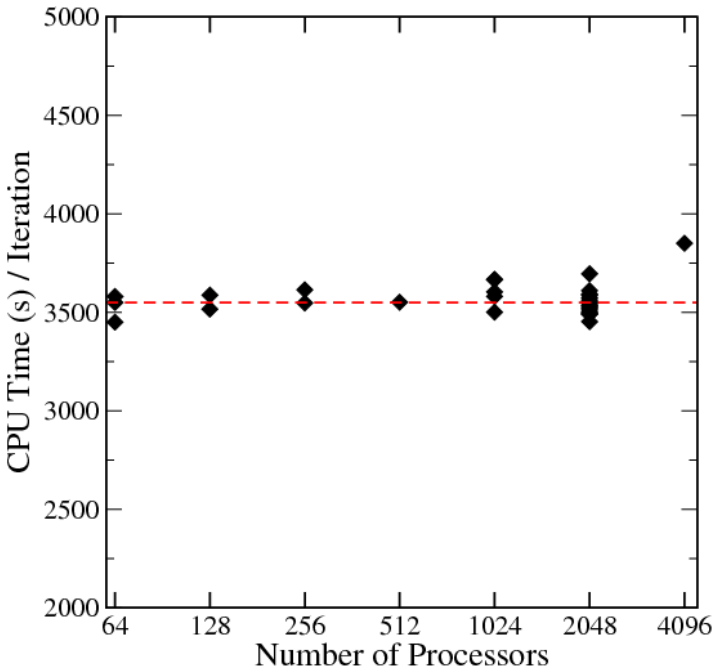


Figure 5.9: Wall-clock time per iteration for the evolutionary algorithm as a function of the number of processors; a constant time dependence indicates perfect scaling. The red line indicates the average of all iterations for N=2048 processors as a guide to the eye.

The control loop is implemented employing a synchronous simulation protocol, where tasks are distributed to all processors of the machine, each drawing a member of the presently active conforma-

tion with equal probability. Each processor then performs a simulated annealing simulation in which the present conformation is optimized independently of all others. For each step of the process the energy evaluation is optimized to compute only those energy terms in the model that have changed from the previous conformation, clashing conformations are rejected outright. For this reason the simulation time varies slightly from processor to processor even though the number of simulation steps is identical for each processor (N=40,000). As the simulations finish, their conformations are transferred to the master, which decides whether to accept (average probability: 57%) the conformation into the active population or disregard the conformation. Then a new conformation is immediately given to the idle processor. Because the processors are processed sequentially some processors wait for the master before they get a new conformation.

Fluctuations in the client execution times (see Figure 5.10) induce a waiting time before the next iteration can start. This waiting time is the largest in the first few iterations, because a processor in subsequent iterations have slight starting offsets along the time axis, which increase the likelihood that the results are returned in the same order of processors that they were issued. In this scenario there would be no waiting time even in a synchronous processing mode.

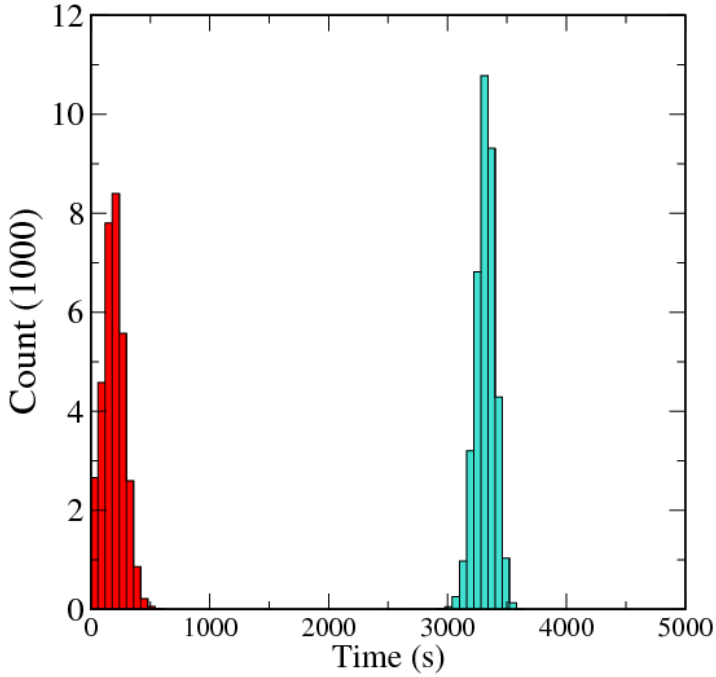


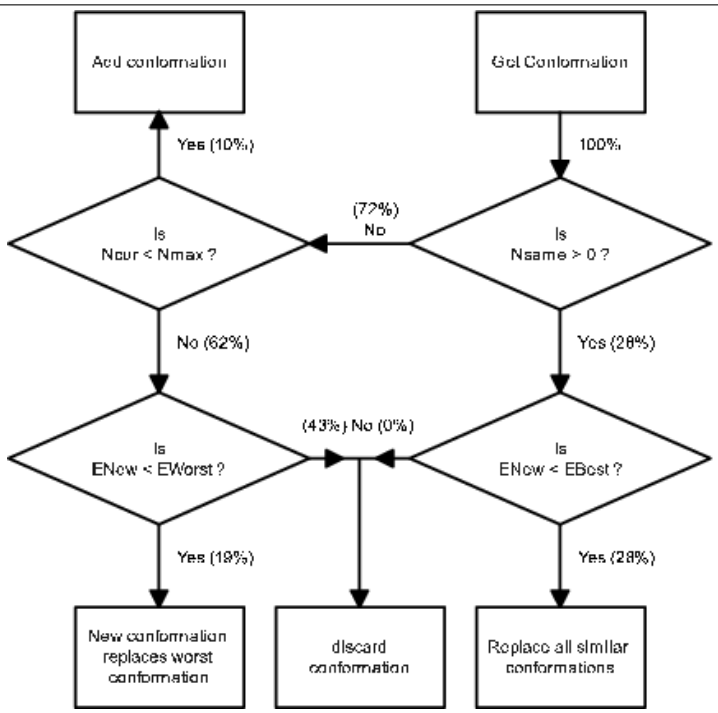
Figure 5.10: Histogram of the distribution of client execution time (blue) and client idle time (red) in seconds for 20 iterations of the EA on 2048 processors.

For the realistic simulation time chosen in these runs, the average waiting time is less than 10% of the execution time and nearly independent of the number of processors used. An asynchronous implementation of the master loop would probably reduce these fluctuations further.

The success of the algorithm to fold the protein can be rationalized by analyzing the flow of information through the decision tree (See Algorithm 1). We have annotated the arrows of the tree

to show the fraction of total new conformations flowing through the various branches. About 30% of the returning conformations are similar to at least one of the active conformations and all of these are accepted into the active population (refinement). This implies that the simulated annealing step is highly successful to improve existing conformations. We find that 10% of simulations lead to the replacement of more than one conformation (merge operation) in the decision tree, which indicates a narrowing of the folding funnel as the simulation proceeds. The protein is not just folded once, but many simulations converge to the same intermediate structure. The merge operation is therefore useful to avoid replication of the information.

Algorithm 1 Schematic illustration of the decision tree for the evolutionary algorithm employed in this investigation: new conformations enter the decision tree with energy ENew, the number of conformations in the population with an RMSD $<$ CutOff RMSD is designated as Nsame. Nmax/Ncur are the maximal/current number of conformations in the population. The highest energy of all conformations in the population is designated by EWorst. The arrows in the tree are annotated by the total probabilities of the conformation flow in the folding simulation described in section.



From the remaining 72% conformations, 10% conformations (the same as the fraction of merge operations) are added to the population because it has shrunk. The algorithm thus succeeds to continuously reseed itself, this generates a high likelihood that the simulation is not stuck in an uninteresting metastable area of the folding landscape. 19% of the new conformations are dissimilar to all other conformations of the population, but nevertheless better than the worst conformations. These new structural templates are then the candidates for further local refinement in the steps discussed above. About 43% of the basin hopping cycles go astray, which is commensurate with earlier basin hop-

ping investigations. We note that the balance of refinement and new structural templates generate a dynamic population that slides as a whole towards the global optimum of the free energy funnel.

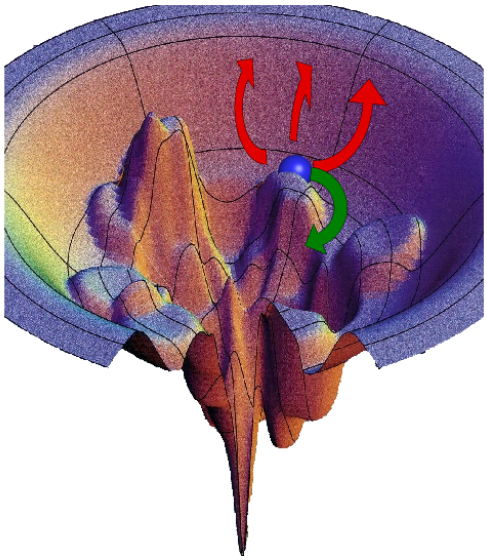


Figure 5.11: Schematic illustration of the different routes of the evolutionary algorithm on a two-dimensional model protein folding funnel. Most simulations explore the vicinity of the starting conformation, but with increasing dimension of the search space, many go astray (red), only a few find new conformations (green), that are refined in later iterations. This inherent limitation of the local search process (here the simulated annealing run) makes it possible to employ algorithms that start many simulations from the same conformation without wasting computational resources. The funnel landscape was taken from K. A. Dill's homepage (www.dillgroup.ucsf.edu).

The evolutionary algorithm evolves not one, but an active population containing many conformations concurrently. Considering the limiting cases, it is apriori unclear, how such a strategy can succeed to efficiently fold the protein. For small population size (P) many processors (N) construct short trajectories emanating from the same conformation ($P \ll N$). If the energy gain for each such step is small compared to the total folding energy, many cycles will be required to complete the simulation even if many processors are available. A large fraction of the computational resources would be wasted in such a scenario. In the opposite limit ($N \ll P$) most conformations sample high energy regions of the free-energy surface that are unrelated to the native conformation. Improvement of such conformations is irrelevant to the folding process. The latter limit is therefore unattractive for large scale distributed computational architectures, where N is large.

The key to convergence lies therefore in the exploitation of the specific characteristics of the protein free energy landscape of naturally occurring proteins. Following the current funnel paradigm (Onuchic et al., 1997; Dill and Chan, 1997) the protein explores an overall downhill process on the energy landscape, where the conformational entropy of the unfolded ensemble is traded for enthalpic gain of the protein and free energy gain of the solvent (Becker and Karplus, 1997; Lazaridis and Karplus, 1997). Using one- or low-dimensional indicators the complex folding process appears for many small

proteins as a two-state transition between the unfolded and the folded ensemble with no apparent intermediates. This transition has been rationalized in terms of the funnel paradigm, where the protein averages over average frictional forces on its downhill path on the free-energy landscape. In this context one cycle of the evolutionary algorithm in the $P \ll N$ limit attempts to improve many times each of the conformations of the active population.

Due to the effective friction and local frustration on the free-energy landscape most of these simulations explore the vicinity of their respective starting points. Because of the actual high dimensionality (D) of the search space ($D = 160$ free dihedral angles for 1F4I) most of them terminate higher in free-energy than their starting conformation. For a rugged two-dimensional free-energy surface this is illustrated schematically in Figure 5.11. These conformations are rejected by the energy criterion. Most of the remaining simulations that improve upon the starting conformation stay within the distance acceptance threshold of the evolutionary algorithm and replace their starting conformation in the active population. The distance acceptance threshold thus ensures that the population is not overpopulated by nearly identical conformations of the same region in conformational space. In the rare event, that the simulation improves the energy and generates a genuinely new conformation, the energetically worst conformation of the active population is replaced. This conformation is the starting point for further local refinement in subsequent iterations.

This analysis reveals the mechanism for the effectiveness of the evolutionary algorithm: The move generator, in this case the simulated annealing run in the individual step, generates an 'acceptable' new conformation with a probability $p(D)$ that falls rapidly with the dimension of the search space and the quality of the present population. As long as $p(D) < P/N$ each cycle of the evolutionary algorithm will improve each member of the active population at most once on average. As long as no genuinely better move generator exists (higher $p(D)$), all computation effort is, on average, efficiently directed towards folding the protein. Only when N becomes so large that the above relation no longer holds, several attempts per cycle will improve the same member of the active conformation, even though only one of these improvements can be kept according to the acceptance rules, leading to a duplication and hence waste of computational resources. This is good news for the scalability of the evolutionary algorithm for larger proteins: Because $p(D)$ drops rapidly with the size of the protein, the number of processors that can be effectively employed for folding can be further increased using thousands, possibly hundreds of thousands of processors concurrently.

6

Folding studies in PFF02

In Chapter 4, it was shown that the modified force field, PFF02 was able to differentiate between native and non-native decoys for different decoy sets including helical, sheet and both secondary structural elements. The average energy of the native decoy set was lower than the average energy of the complete decoy sets, which demonstrates the selectivity of the force field. We have also developed and implemented stochastic optimization methods which could reliably locate the global minimum (Chapter 5). In this chapter we investigate the ability to use the stochastic methods in the new force field PFF02 to predict the native state of various proteins starting from random conformations.

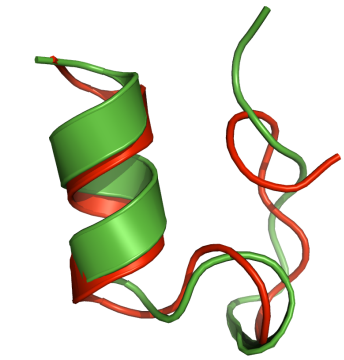
6.1 Helical proteins

In the earlier studies it had been shown that PFF01 correctly predicts the native state of various helical proteins. As we have introduced new terms in PFF02, in the first step towards all atom protein folding we undertake the study of helical proteins. This is done to ensure that the proteins investigated in earlier studies with PFF01 are not destabilized and can be reproduces in PFF02. We studied all atom folding of five helical proteins in PFF02, including three proteins which were earlier folded in PFF01 and two large helical proteins (over 50 amino acids).

6.1.1 Tryptophan cage - 1L2Y

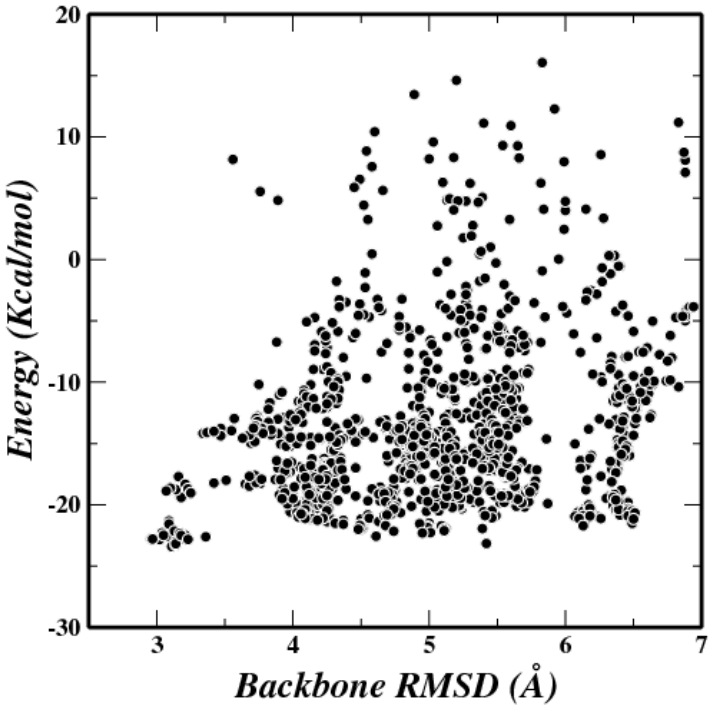
Tryptophan cage or trp-cage protein (Neidigh et al., 2002) has been subjected to various theoretical studies and it has been of great scientific interest. It had been reported to fold with PFF01 and STUN and replica exchange MD simulations (Snow et al., 2002; Schug et al., 2003b; Ding et al., 2005; Linhananta et al., 2005; Schug et al., 2005b, 2006; Juraszek and Bolhuis., 2006).

Here we study the all atom folding of tryptophan cage protein. We performed 20 independent basin hopping simulations starting with the completely extended conformations in PFF02 with 100 cycles. The starting conformation had a RMS of 12.94 Å to the native conformation and was completely extended manually (by setting all backbone dihedral angles except proline to 180°). The starting temperatures were chosen from a distribution of exponentially distributed temperatures and the number of steps increased with the BHT cooling cycle by $10^4\sqrt{n_m}$ where n_m is the number of



(a) Overlay

(b) C_β-C_β matrix



(c)

Figure 6.1: 1L2Y: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

minimization cycles.

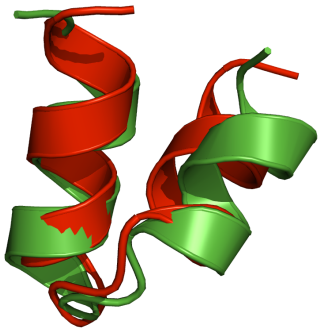
The lowest energy structure converges to a native like conformation with RMSD of 3.11 Å to the native conformation. For the sake of uniformity in case of NMR resolved experimental structures, we compare the RMSD to the first model in the protein data bank file. The lowest energy structure had an energy of -23.4 Kcal/mol. Figure 6.1(c) shows the scatter plot of the conformations visited by the basin hopping simulations on the free energy surface. The overlay of native conformation (green) with the lowest energy conformation (red) is shown in Figure 6.1(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 6.1(b). The C_β-C_β overlay matrix quantifies the tertiary alignment along with secondary structure formation by taking the difference between all C_β distances of predicted and native conformation. Black regions indicate excellent agreement in the formation of native contacts while white regions indicate larger deviations.

We were able to locate the global minimum of tryptophan cage protein in PFF02 but there were many non-native conformations which are within 1 Kcal/mol to the lowest energy conformation. Thus we do not conclude the simulation to be predictive.

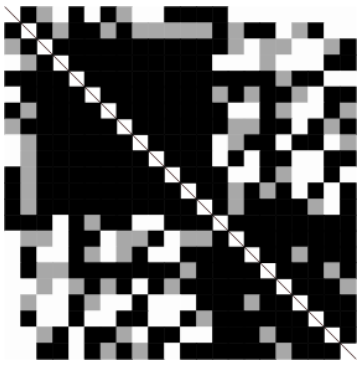
6.1.2 Potassium channel blocker - 1WQE

Potassium channel blockers, 1WQC, 1WQD, 1WQE (Chagot et al., 2005) are of specific interest due to their unusual fold for ion channel blockers. They are toxic venom peptides involved in blocking of potassium channel in cells. They have two helices which are 'locked' together with a disulphide bond. Here we study the folding of 1WQE, a two helical protein which had earlier been folded using PFF01 (Wenzel, 2006). The starting conformations for this study were completely extended conformations with RMSD of 20.6 Å to the native conformation.

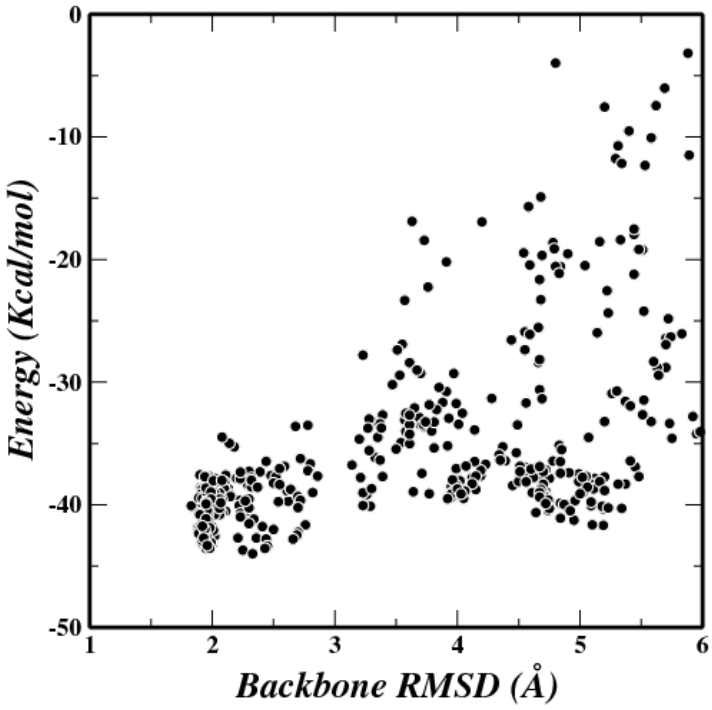
We did ten independent basin hopping simulations from the extended conformation in PFF02. Nine out of ten independent BHT simulations converge to conformations that differ by less than 3 Å RMSD to the native conformation. The lowest energy structure found in the simulations has a RMSD of only 2.33 Å to the native conformation with an energy of -44.0 Kcal/mol. This is very encouraging, the contribution to the formation of disulphide bridges is yet to be incorporated in PFF02.



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.2: 1WQE: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

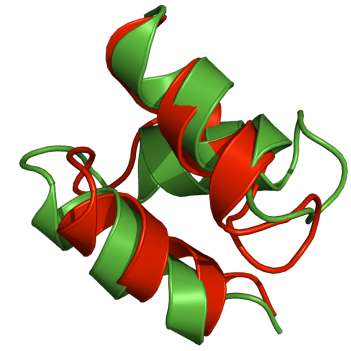
Figure 6.2(c) shows the scatter plot of conformations visited during the simulations. There are many conformations visited around the native state and the next metastable state can be seen at around 5Å and is about 4 Kcal/mol higher in energy. This metastable conformation also correctly predicts

the two helices but arranges itself in an orthogonal packing instead of up-down arrangement. Independently the two helices PRO3-THR12 and VAL12-CYS22 have an RMSD of 0.7 and 0.44 Å only. Figure 6.2(a) shows the overlay of the native conformation (green) to the lowest energy conformation (red) encountered in the simulations. The overlay shows the perfect agreement of the lowest energy conformation to the experimental structure and the C_β-C_β matrix (Figure 6.2(b)) illustrates the tertiary alignment of the overlay with many black regions.

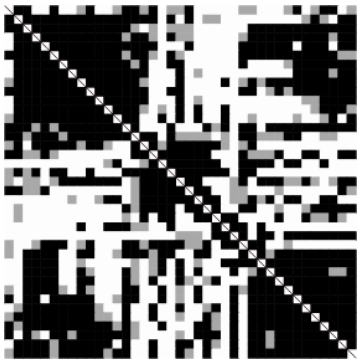
As nine out of ten simulations converge to native like conformation and the metastable conformation is 4 Kcal/mol higher in energy, we conclude the folding of 1WQE as predictive and reproducible.

6.1.3 HIV accessory protein - 1F4I

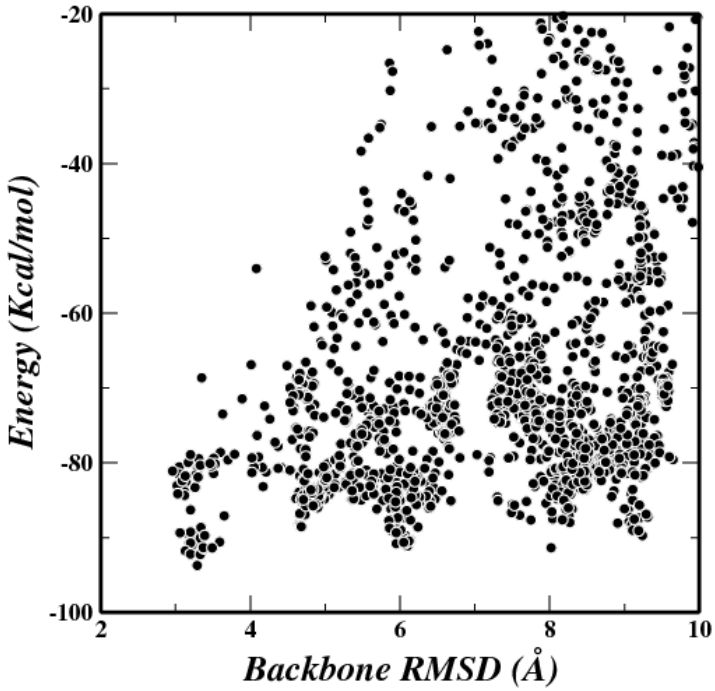
HIV accessory protein destroys the host cell's ability to survive by binding to a host receptor and restricting an important enzyme to activate the cell's immune system. The 40 amino acid HIV accessory protein 1F4I (Withers-Ward et al., 2000) was earlier folded using PFF01 starting from random starting conformations (Herges and Wenzel, 2004; Schug et al., 2004a).



(a) Overlay



(b) C_β-C_β matrix



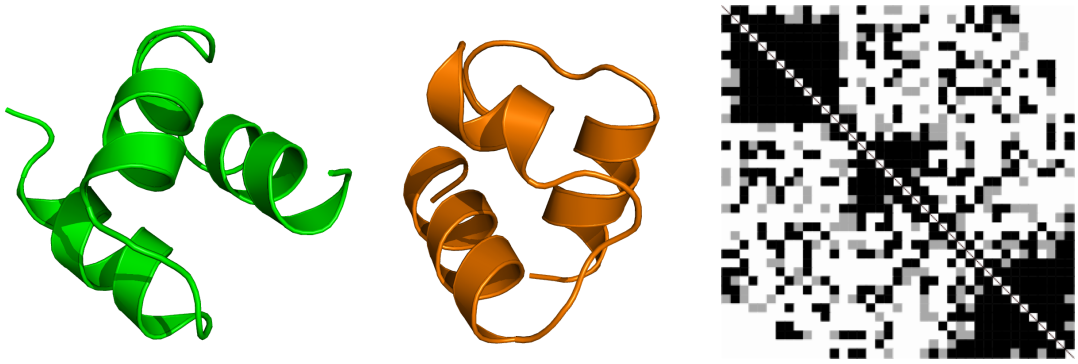
(c) Erms Plot

Figure 6.3: 1F4I: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

We studied the folding of 1F4I in PFF02 with basin hopping technique. We did twenty inde-

pendent runs with completely extended conformations in PFF02 for 150 basin hopping cycles. The number of cycles are larger than 1L2Y and 1WQE as the complexity of the search space increases with the size of protein. The temperatures were chosen from an exponential distribution and the cooling cycle length was increased as described above. The lowest energy structure encountered in the simulations had an RMSD of 3.29 Å to the native conformation and had an energy of -93.7 Kcal/mol.

The scatter plot of the conformations visited during the simulations is shown in Figure 6.3(c). Apart from the native-like conformations, there are clusters of low energy conformations around 6 and 8 Å. The first non-native conformation is 2 Kcal/mol higher in energy than the lowest native-like conformation. While this misfolded structure differs significantly, it has the same secondary structure. The misfolded conformation is shown in Figure 6.4. The corresponding C_β-C_β overlay matrix also shows the secondary structure formation with a different tertiary arrangement. Independently, helix-1(LYS3-LEU12), helix-2(GLU16-PHE24) and helix-3(ASN31-SER39) in this misfolded conformation are nearly perfectly predicted and have RMSD's of only 0.53, 2.0 and 0.52 Å respectively.



(a) (b) (c)

Figure 6.4: 1F4I: Overlay of misfolded (orange) structure to experimental (green) structure and the overlay of the C_β-C_β distance matrix.

Figure 6.3(a) shows the overlay of the native conformation(green) to the lowest energy conformation encountered in the simulations. The overlay shows the agreement of the predicted conformation to the native structure.The C_β-C_β matrix (Figure 6.3(b)) shows the tertiary alignment of the overlay with dark regions. Both the starting (LYS3-LEU12) and the end (ASN31-SER39) helix were correctly predicted, but the middle helix (GLU16-PHE24) had a different tertiary arrangement because of a wrongly predicted turn region. This can also be observed from the C_β-C_β overlay matrix.

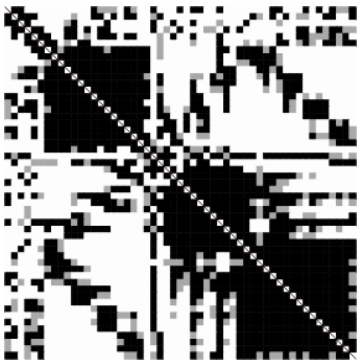
Only one of the twenty basin hopping simulations converged to native-like conformation, but the energy of native-like conformation was significantly lower (less than 2Kcal/mol) than any other conformation and thus we conclude the folding study to be predictive but not reproducible.

6.1.4 Engrailed Homeodomain - 1ENH

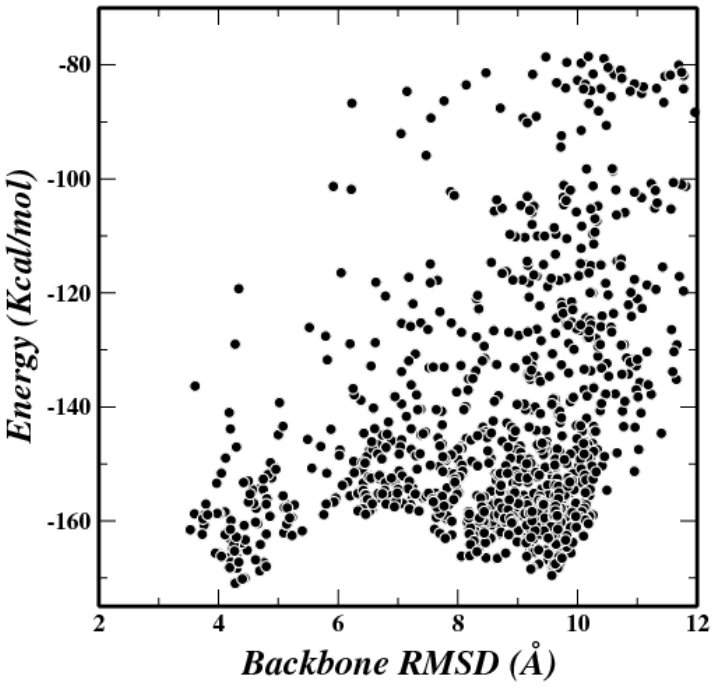
The 54 amino acid engrailed homeodomain protein (Clarke et al., 1994) is a three helical orthogonal bundle protein which has been subjected to detailed molecular dynamics simulations (Mayor et al., 2003; Daggett and Fersht, 2003). It was not possible to fold this protein using basin hopping technique and the simulations never reached the energies of the native-like conformations.

Here we studied the folding of engrailed homeodomain in PFF02 using evolutionary algorithm with a maximum population of 64 conformations and 512 processors (Verma and Wenzel, 2006a). The lowest energy structure converges to 4.28 Å to the native conformation with the energy of -170.95 Kcal/mol. 1ENH has a unstructured tail at the N-terminus, excluding this seven amino acid region, the RMSD reduces to only 3.4Å.

(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

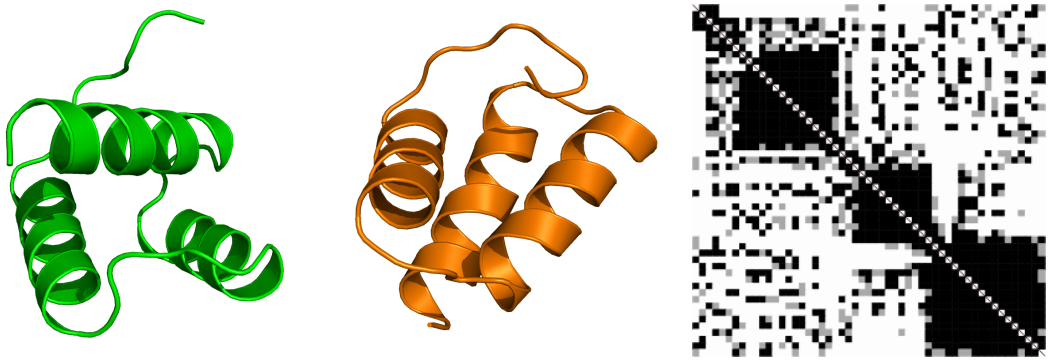
Figure 6.5: 1ENH: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

The scatter plot of conformations visited during the simulation are shown in Figure 6.5(c). Seven out of the total population of 64 structures are less than 4.5 Å RMSD to the native conformation. The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 6.5(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 6.5(b). There are also competing conformations (within 2 Kcal/mol) with large RMS deviations encountered in the

simulations. One such conformation is shown in Figure 6.6). These conformations have the same secondary structure, but a different tertiary structure alignment. The C_β-C_β overlay matrix for the misfolded conformation also confirms that all the three helices are properly predicted but their tertiary arrangement is completely different.

No two helices in the misfolded conformation are in agreement with the respective helices in the native state. Independently, helix-1 (E8-E20), helix-2 (E26-L36) and helix-3 (A40-K43) are nearly perfectly predicted and have RMS of only 0.56, 0.42 and 0.47 Å respectively.

As about 10% of the population is native-like and the misfolded conformations we can conclude that the folding is reproducible.



(a) (b) (c)

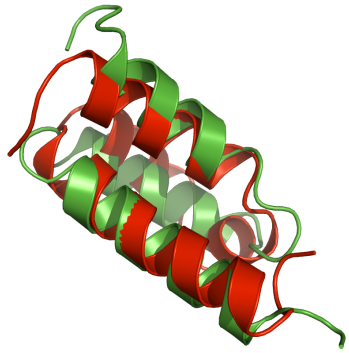
Figure 6.6: 1ENH: Overlay of misfolded (orange) structure to experimental (green) structure and the overlay of the C_β-C_β distance matrix.

6.1.5 E domain of Staphylococcal Protein A - 1EDK

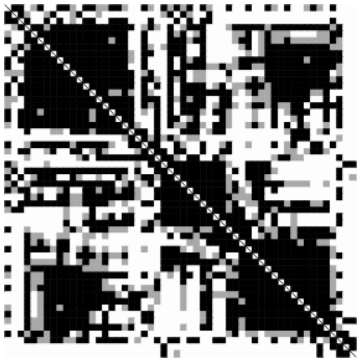
The E-domain of staphylococcal protein A is one of five homologous Immunoglobin G-binding domains designated E, D, A, B, and C that comprise the extracellular portion of protein A (Starovasnik et al., 1996). Its architecture is classified as an up-down bundle in CATH (Orengo et al., 1997), which makes the topology of the three helical bundle different from the one in engrailed homeodomain protein or HIV accessory protein.

As the final example of helical proteins, we studied the folding of protein A in PFF02 using evolutionary algorithm with a maximum population of 64 conformations and 256 processors for 50 cycles. The lowest energy structure converges to 4.05 Å to the native conformation with the energy of -154.78 Kcal/mol. Excluding the unstructured regions from both the N-terminus and C-terminus, the RMSD of the structured region (GLU5-SER52) reduces to only 2.99 Å.

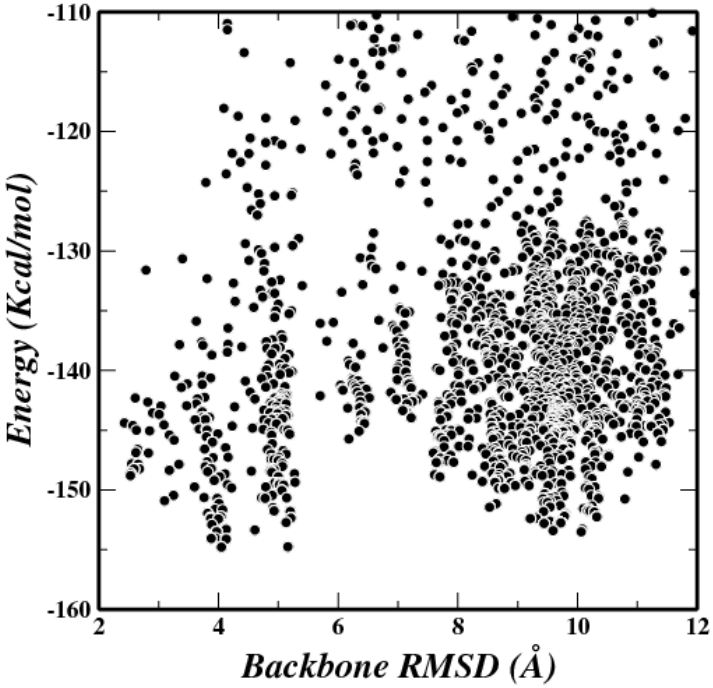
The scatter plot of conformations visited during the simulation are shown in Figure 6.7(c). It shows two funnel like regions. Seven out of the total population of 64 structures are less than 4.05 Å RMSD to the native conformation. The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 6.7(a) and the corresponding C_β-C_β overlay matrix is



(a) Overlay

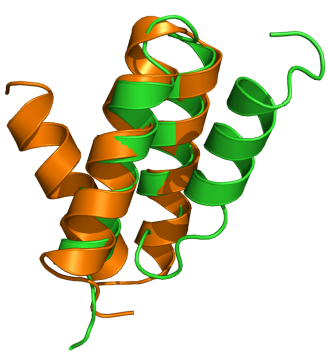


(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.7: 1EDK: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.



(a)

(b)

Figure 6.8: 1EDK: Overlay of misfolded (orange) structure to experimental (green) structure and the overlay of the C_β-C_β distance matrix.

shown in Figure 6.7(b).

There are also competing conformations (within 2 Kcal/mol) with large RMS deviations ($\sim$ 10 Å) encountered in the simulations. One such conformation is shown in Figure 6.8. This conformation is the mirror image of the native conformation as helix1 (GLU5-LEU15) and helix2 (ALA22-ASP34) align perfectly but helix3 (ALA40-SER52) is in the opposite direction because of the wrong turn. Independently helix1, helix2 and helix3 have RMS deviations of only 0.40, 0.49 and 0.46 Å respectively.

Again about 10% of the population included native-like structures indicating reproducible folding of protein A.

For all proteins which had competing metastable states, the secondary structure was always correctly predicted. This indicates that for proteins, in the low energy region the secondary structure is almost always formed correctly and what lacks is the tertiary arrangement of these secondary structure elements (Herges and Wenzel, 2005b).

6.2 Hairpins

Hairpins are the simplest beta sheet structures with only two strands in antiparallel directions that are connected together with a turn. Hydrogen bonding and the packing of the protein itself plays a crucial role here in the folding of such small polypeptides. There are not many hairpin proteins that are not stabilized by external interaction with ions or with the formation of disulphide bridges.

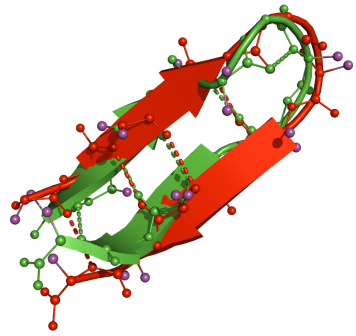
In this section we report the folding studies of various polypeptide chains which are stable in physiological conditions and have no other stabilizing contributions like disulphide bonds arising from cystine sidechains. Folding of such small polypeptides are the next step towards the more universal force field, as the force field selectivity is tested between a helical conformation and a sheet conformation. The helical conformation gives greater contribution with hydrogen bonding energy as it has more number of hydrogen bonds as compared to the sheet (every hydrogen bond in PFF02 gives a contribution of about 2 Kcal$\cdot$mol^{-1}). This hydrogen bonding energy should be compensated by the inclusion of new terms in PFF02 and change in other interactions.

6.2.1 Tryptophan zipper - 1LE0

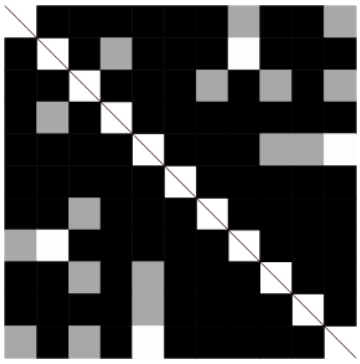
Tryptophan zippers are one of the smallest monomeric, stable β-hairpins that adopt an unique tertiary fold without requiring metal binding, unusual amino acids, or disulfide crosslinks (Cochran et al., 2001). We were able to fold various tryptophan zippers using PFF02 and basin hopping technique (not shown here).

We studied the folding of 1LE0 using 128 processors on Marenostrum cluster at the Barcelona supercomputer center starting from completely extended conformations. We performed twenty cycles of evolutionary algorithm. The lowest energy conformation reached in the simulation had a RMS of only 1.5 Å to the native conformation with the energy of -29.97 Kcal/mol.

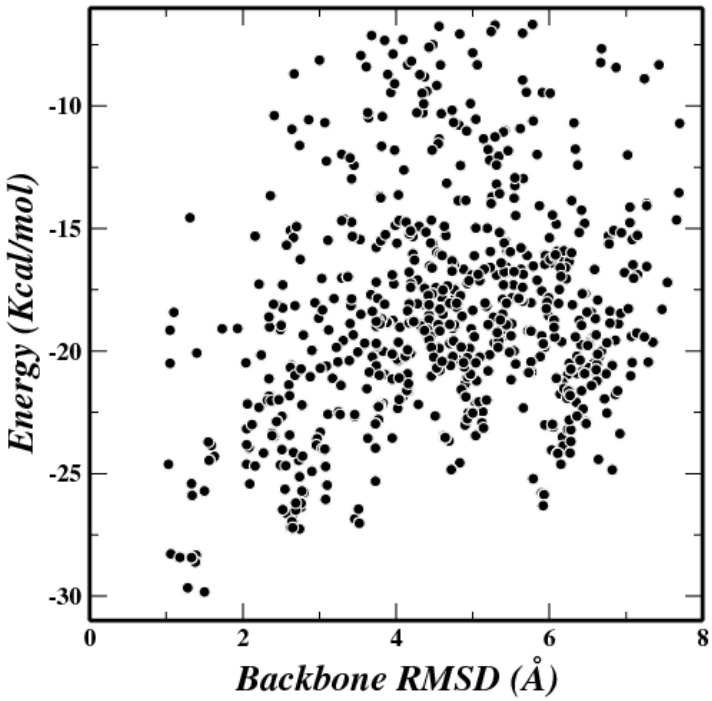
The scatter plot of the conformations visited during the simulations is shown in Figure 6.9(c). The scatter plot shows that the native-like conformations lie significantly below any other conformation.



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.9: 1LE0: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs RMSD plot.

Twelve out of the 64 conformations from the final population are less than 3.0 Å to the native conformation. The protein folds in less than 90 minutes using 128 processors in parallel using the twenty cycles of evolutionary algorithm amounting to 77×10^6 function evaluations or about 9 CPU days *.

The overlay of the predicted conformation (red) with the native conformation (green) is shown in Figure 6.9(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 6.9(c). Large black regions in the C_β-C_β overlay matrix indicates the agreement of native contacts between the two conformations.

As hydrogen bonding plays an important role in the formation and topology of β-sheet structures, it is important to compare the hydrogen bonding pattern in the lowest energy conformations as two β-sheet conformations might look very similar to the eye, but they might have completely different topology resulting from shifting of backbone hydrogen bonds.

The pattern of backbone hydrogen bonds are shown in Table 6.1 for the native and the predicted conformation. These were calculated using standard definitions with MOLMOL (Distance=2.4Å and

*We also performed the same simulation with same number of processors and same population size but smaller number of steps. This simulation converged to less than 3 Å in less than 15 minutes, but all the conformations had higher respective energies.

Hydrogen bond	Native	Predicted
03 THR HN ⟶ 10 THR O	X	X
05 GLU HN ⟶ 08 LYS O	X	X
07 ASN HN ⟶ 05 GLU O	X	
10 THR HN ⟶ 03 THR O	X	X
12 LYS HN ⟶ 01 SER O	X	X

	Secondary Structure	RMSD (Å)
Native	CEEECSSSEEEC	-
Predicted	CEEEETTEEEEC	1.52

Table 6.1: 1LE0: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

angle=35°). Four out of the five backbone hydrogen bonds of the native structure are predicted correctly in the lowest energy structure found in the simulations.

As about 20% of the population converged to native-like conformations with much lower energies, we conclude the folding of tryptophan zipper as reproducible and predictive.

6.2.2 HIV-1 V3 loops

Here we study the folding of two HIV-1 V3 loops. The V3 loop of the HIV-1 envelope glycoprotein gp120 is involved in binding to the CCR5 and CXCR4 coreceptors (Sharon et al., 2003). The structures of an HIV-1 V3 peptides bound to the respective antibody were found to be a beta hairpin. The hairpin structure with specific sidechains on the side is responsible for the binding of viral protein to its receptor. Only when the protein is properly bound, the virus can enter the cell. We studied the folding of two such loops, 1NIZ and 1U6U, which differ themselves by only an insertion of two amino acids in the 1niz sequence. The insertion changes the loop structure resulting in different sidechains getting exposed to the receptor. This change in the structure of V3 loops is considered to be responsible for coreceptor selectivity by the virus protein.

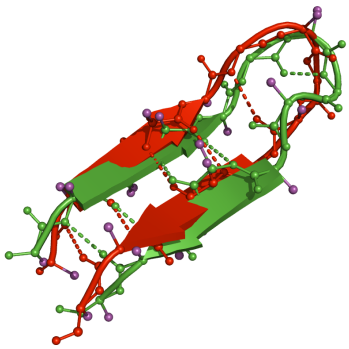
The $V3_{MN}$ loop - 1NIZ

We studied the folding of 14 amino acid HIV-1 $V3_{MN}$ loop 1NIZ (Sharon et al., 2003) in PFF02 using a greedy version of basin hopping technique (Verma and Wenzel, 2006b).

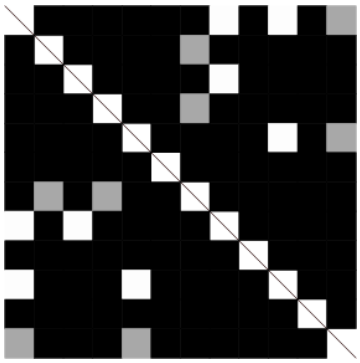
In basin hopping simulations there is a threshhold energy acceptance criterion at the end of every basin hopping cycle. In our simulations, we have used this threshhold acceptance criterion of 1-3 Kcal/mol depending upon this size of the protein. In the greedy version of basin hopping the threshold energy is varied depending upon the best energy found so far in the simulation. Here we calculated the threshold criteria as $(\varepsilon_S - \varepsilon_B)/4$, where ε_S is the starting energy and ε_B is the best energy found so far in the simulation. This choice implies that the conformation with the best energy is never replaced with a conformation that is higher in energy and thus introduces a "memory effect" in the simulation. For

the simulations that are higher in energy, the increased threshold value implies a higher acceptance probability of conformations with higher energy.

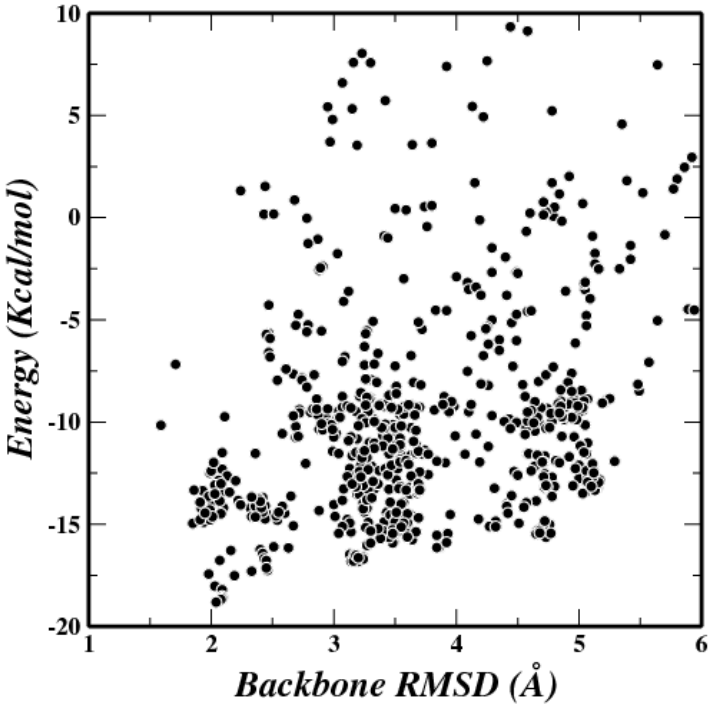
We did 200 cycles of greedy basin hopping simulations in PFF02. The simulations were started with completely extended conformation which had a RMS of 12 Å to the native state. The lowest energy structure found in the simulation had a RMSD of only 2.04 Å to the native state.



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.10: 1NIZ: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

The scatter plot of the conformations visited during the simulations is shown in Figure 6.10(c). The scatter plot shows a single downhill folding funnel for this hairpin. Eight out of the ten independent simulations converged to less than 3.5 Å RMSD to the native conformation.

The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 6.10(a) and the corresponding C_β-C_β distance matrix is shown in Figure 6.10(c). Large black regions in the C_β-C_β overlay matrix indicates the agreement of native contacts between the two conformations.

Again we did the backbone hydrogen bond analysis and four out of the five backbone hydrogen bonds of the native structure are correctly predicted in the lowest energy structure found in the simulations. The pattern of backbone hydrogen bonds is shown in Table 6.2. The secondary structure of the predicted and native conformation is also shown in Table 6.2. The letters in the secondary structure

Hydrogen bond	Native	Predicted
02 ARG HN ⟶ 13 THR O	X	X
04 HIS HN ⟶ 11 PHE O	X	X
06 GLY HN ⟶ 09 ARG O		X
08 GLY HN ⟶ 06 GLY O	X	
11 PHE HN ⟶ 03 HIS O	X	X
13 THR HN ⟶ 01 ARG O	X	X

	Secondary Structure	RMSD (Å)
native	CEEEECSSCEEEEC	-
predicted	CEEEECSSCEEEEC	2.04

Table 6.2: 1NIZ: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

correspond to DSSP definitions (see appendix A).

As eight of the ten simulations converged to native-like conformation without any competing metastable conformations, the folding is concluded as reproducible and predictive.

The HIV-1 $V3_{IIIB}$ loop - 1U6U

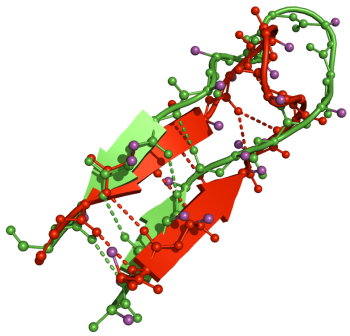
Comparison of the known V3 structures leads to a model in which a 180 degrees change in the orientation of the side chains and the resulting one-residue shift in backbone hydrogen bonding patterns in the N-terminal strand of the β-hairpins markedly alters the topology of the surface that interacts with antibodies and that can potentially interact with the HIV-1 coreceptors (Rosen et al., 2005).

We studied the folding of 17 amino acid HIV-1 $V3_{IIIB}$ loop-1U6U in PFF02 using a greedy version of basin hopping technique for same 200 cycles as for 1NIZ. The simulations were started with completely extended conformation which had a RMS of 15 Å to the native state.

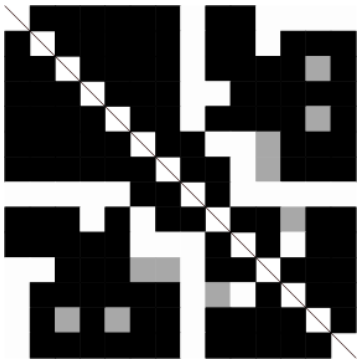
All the ten independent simulations after 200 cycles of greedy basin hopping found the β-sheet like conformations. The lowest energy conformation (-32.9 Kcal/mol) found in the simulation had a RMS of 4.57 Å to the native state, which is relatively higher for a beta hairpin. This happens because of an overall bend in the loop resulting from solvent interactions, which can be expected as the peptide is a fragment of a larger protein.

The scatter plot of all the conformations visited during the simulations is shown in Figure 6.11(c) which shows the single funnel-like landscape. The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 6.11(a) and the corresponding C_β-C_β distance matrix (Figure 6.11(b)) shows that the the two strands are correctly predicted and has the correct tertiary arrangement. The lowest energy structure still correctly predicts four out of five native backbone hydrogen bonds, thus indicating the correct pattern found in PFF02.

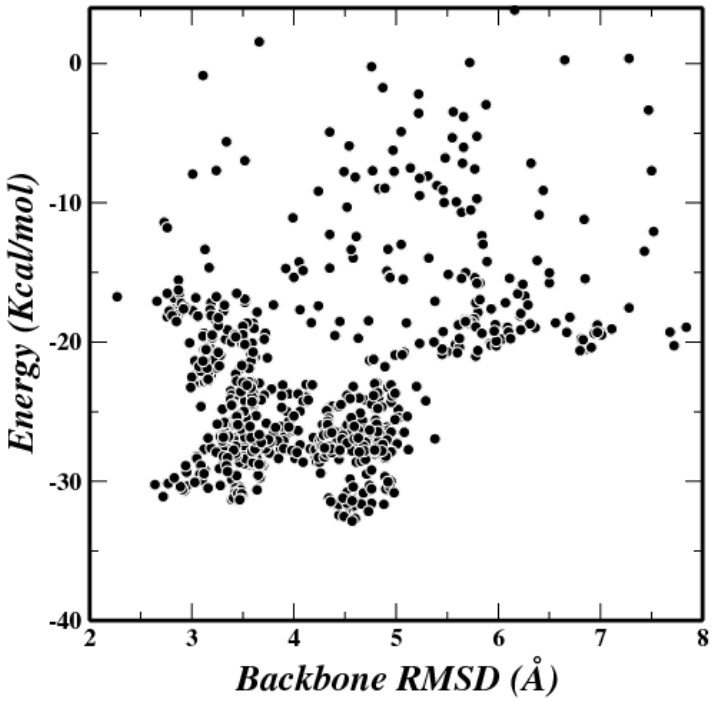
The hydrogen bond analysis helps us understand the topology better as the lowest energy conformation had larger deviations from the native structure. The backbone hydrogen bonds of both these



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.11: 1U6U: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs RMSD plot.

Hydrogen bond	Native	Predicted
02 SER HN ⟶ 16 ILE O	X	X
04 ARG HN ⟶ 14 VAL O	X	X
14 VAL HN ⟶ 04 ARG O	X	X
16 ILE HN ⟶ 02 SER O	X	X

	Secondary Structure	RMSD (Å)
Native	CEEEECCSTTCCEEEEC	-
Predicted	CEEEEEEETTTEEEEEC	4.57

Table 6.3: 1U6U: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

conformations are shown in Table 6.3. As all four backbone hydrogen bonds are predicted correctly in the lowest energy conformation, it is evident that this conformation has correct topology regardless of its high RMS deviation which occurs due to dislocated turn region.

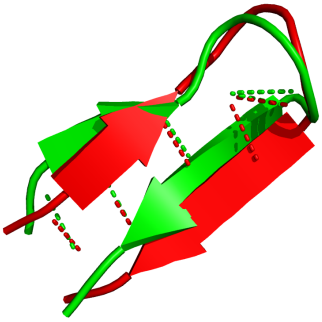
We had predictively and reproducibly folded two very similar(sequence) proteins with different topologies in PFF02. PFF02 can thereby differentiate between these two HIV-1 V3 loops.

6.2.3 HP7, a 12-residue β-hairpin - 2EVQ

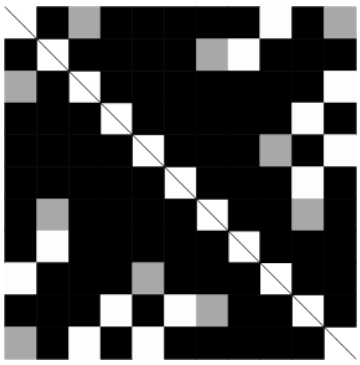
HP7 is a twelve amino acid designed β-hairpin (Andersen et al., 2006).

Here we studied the folding of this protein with the greedy version of basin hopping simulations (Verma and Wenzel, 2007a). We performed ten independent simulations of greedy basin hopping method with 100 cycles in PFF02. The simulations were started from completely extended conformation of the protein which had a RMSD of 10.5Å to the native state.

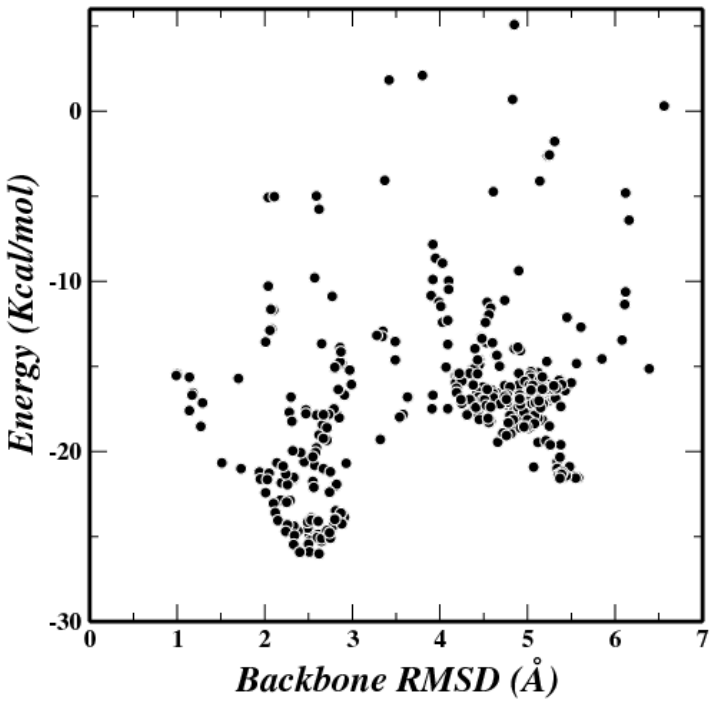
Eight out of the ten independent simulations after 100 cycles of greedy basin hopping find the β-sheet like conformations and converge to less than 3.0Å RMSD to the native conformation. The lowest energy conformation has an RMSD of 2.62 Å to the native conformation and had energy of -26.0 Kcal/mol.



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.12: 2EVQ: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

The scatter plot of the conformations visited during the simulations is shown in Figure 6.12(c). The scatter plot shows two funnels on the free energy surface for this hairpin. The metastable con-

Hydrogen bond	Native	Predicted
02 THR HN ⟶ 11 THR O	X	X
04 ASN HN ⟶ 09 LYS O	X	X
07 THR HN ⟶ 04 ASN O	X	
08 GLY HN ⟶ 04 ASN O	X	X
11 THR HN ⟶ 02 THR O	X	X

	Secondary Structure	RMSD (Å)
Native	CEEETTTTEEEC	-
Predicted	CEEETTTTEEEC	2.62

Table 6.4: 2EVQ: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

formations corresponding to the funnel at around 5.5 Å populated helical conformations, but is 7 Kcal/mol higher than the lowest energy conformation. This shows that the lowest energy conformation is native-like and significantly lower than other metastable conformations.

The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 6.12(a) and the corresponding C_β-C_β distance matrix is shown in Figure 6.12(c). Large black regions in the C_β-C_β overlay matrix indicates the agreement of native contacts between the two conformations.

Again we did the hydrogen bond analysis and four out of the five backbone hydrogen bonds of the native structure are predicted in the lowest energy structure found in the simulations. The pattern of backbone hydrogen bonds is shown in Table 6.4. The secondary structure of the predicted and native conformation is also shown in Table 6.4.

As eight of ten simulations converged to native-like conformation without any competing metastable conformations, the folding is concluded as reproducible and predictive.

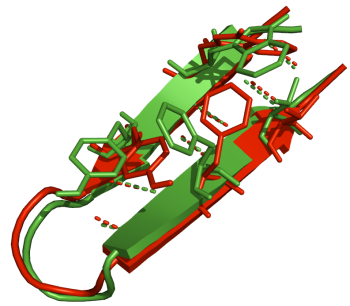
6.2.4 C terminal hairpin of the Protein G

The C terminal hairpin of protein G has been subjected to various scientific studies on beta sheet formations (Zhou et al., 2001; Islam et al., 2004; Nguyen et al., 2005; Nguyen, 2006) and is considered stable in isolation from the rest of the protein.

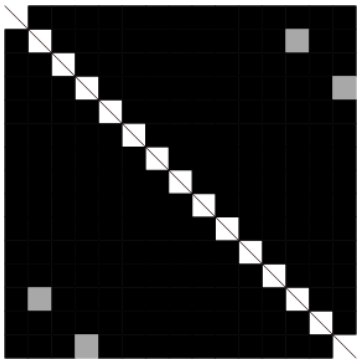
Here we study the folding of this hairpin domain in PFF02 with basin hopping simulations. We started ten independent basin hopping simulations for 100 cycles in PFF02. The staring conformation was completely extended and had RMSD of 15.8 Å to the native conformation.

We found that only one of ten simulations converged to a sheet-like conformation, while the remaining nine simulations are always stuck at the helical conformations. The lowest energy conformation is a beta hairpin and has only 1.27 Å RMSD to the native conformation with energy of -27.3 Kcal/mol. The energy of the lowest helical conformation is -26.9 and is thus only 0.4 Kcal/mol away.

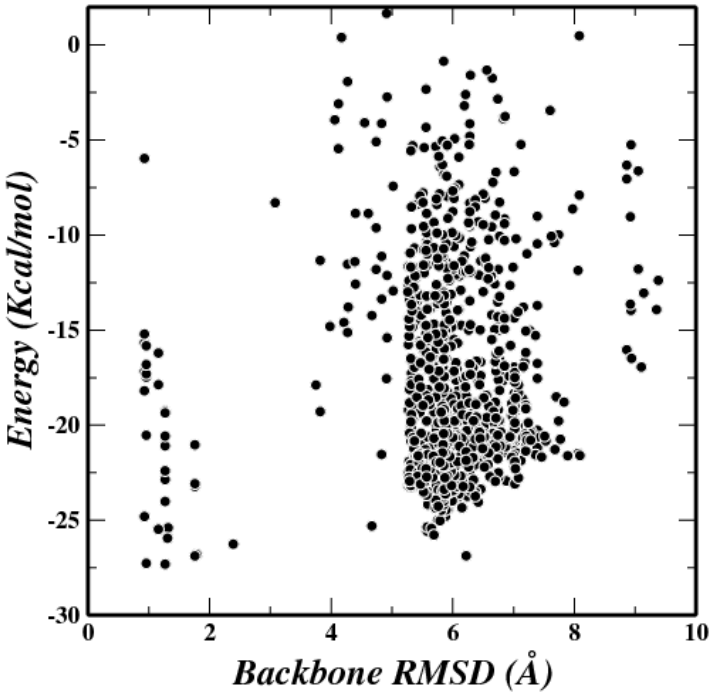
The scatter plot of all conformations visited during the simulations is shown in Figure 6.13(c). It



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

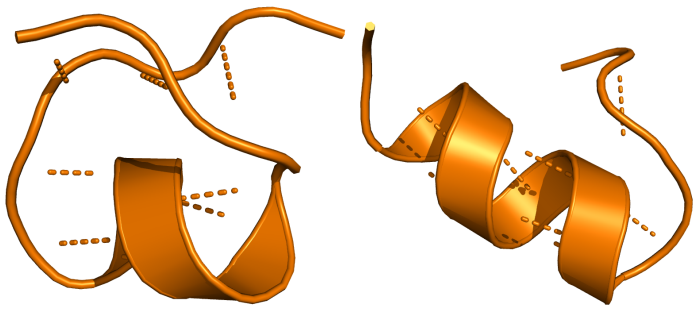
Figure 6.13: C terminal hairpin of protein G: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

can be easily seen that very few conformations are native like and there are many conformations at about 6 Å RMSD. There are almost no conformations in the region between 6Å and 1Å indicating the presence of a huge barrier between the helical conformation and the native conformation, which is not crossed by most of the simulations. The landscape for this hairpin appears to be very complex. Some of the misfolded helical conformations are shown in Figure 6.14.

The overlay of the predicted (red) and native (green) conformation is shown in Figure 6.13(a) and the corresponding C_β-C_β distance matrix is shown in Figure 6.13(c). The C_β-C_β overlay matrix is completely black indicating complete agreement of native contacts between the two conformations.

Again we did the hydrogen bond analysis and all six backbone hydrogen bonds of the native conformation are predicted in the lowest energy conformation found in the simulations. The pattern of backbone hydrogen bonds is shown in Table 6.5. The secondary structure of the predicted, native and misfolded conformation is also shown in Table 6.5.

Although the lowset energy conformation has near perfect native contacts and backbone hydrogen bonding pattern, the simulation is neither predictive nor reproducible.



(a) (b)

Figure 6.14: C terminal hairpin of protein G: Misfolded structures with more backbone hydrogen bonds and more helical content.

Hydrogen bond	Native	Predicted
02 GLU HN ⟶ 15 THR O	X	X
04 THR HN ⟶ 13 THR O	X	X
06 ASP HN ⟶ 11 THR O	X	X
11 THR HN ⟶ 06 ASP O	X	X
13 THR HN ⟶ 04 THR O	X	X
11 THR HN ⟶ 02 GLU O	X	X

	Secondary Structure	RMSD (Å)
Native	CEEEEETTTTEEEEEC	-
Predicted	CEEEEETTTTEEEEEC	1.27
Misfolded	CEECHHHHHHHSEECC	6.22

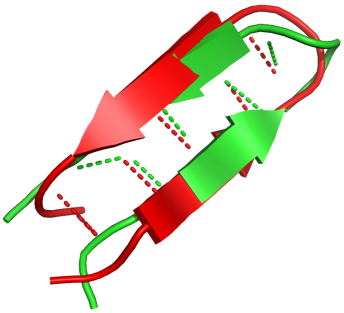
Table 6.5: C terminal hairpin of protein G: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

6.2.5 Designed stable β hairpin - 1J4M

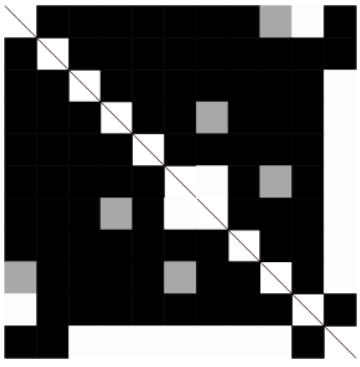
The hairpin 1J4M is a designed stable beta hairpin (Pastor et al., 2002). It is designed to be extremely stable in the β-sheet conformation.

Here we studied the folding of 1J4M with basin hopping simulations. We performed ten independent simulations with 100 basin hopping cycles in PFF02. The simulations were started from completely extended conformation of the protein which had a RMSD of 13.3Å to the native state.

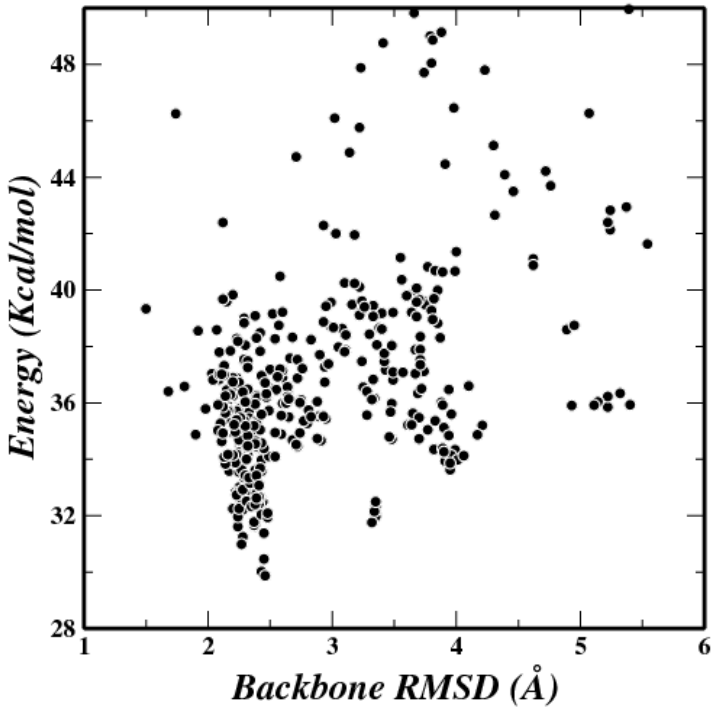
Nine out of ten independent simulations after 100 cycles of greedy basin hopping found the β-sheet like conformations and converge to less than 3.0Å RMSD to the native conformation. The lowest energy conformation has an RMSD of 2.46 Å to the native conformation and had energy of 29.9 Kcal/mol. The energies reported here are positive as the native conformation has some covalently bound atoms which are clashing in PFF02. As the bond distances are kept fixed in PFF02 from the starting conformation for all simulations, thus introducing a constant bias and keep the energies comparable.



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.15: 1J4M: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

The scatter plot of all conformations visited during the simulation are shown in Figure 6.15(c). The overlay of the predicted (red) and native (green) conformation is shown in Figure 6.15(a) and

Hydrogen bond	Native	Predicted
04 TRP HN ⟶ 11 TYR O	X	X
06 TYR HN ⟶ 09 ILE O	X	X
09 ILE HN ⟶ 06 TYR O	X	X
11 TYR HN ⟶ 04 TRP O	X	X
13 GLY HN ⟶ 11 TYR O	X	

	Secondary Structure	RMSD (Å)
Native	CCCEEETTEEECCC	-
Predicted	CCEEEETTEEEECC	2.46

Table 6.6: 1J4M: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.

the corresponding C_β-C_β distance matrix is shown in Figure 6.15(b). The C_β-C_β overlay matrix is completely black indicating complete agreement of native contacts between the two conformations.

Again we did the hydrogen bond analysis and four out of the five backbone hydrogen bonds of the native conformation are predicted in the lowest energy conformation found in the simulations. The pattern of backbone hydrogen bonds is shown in Table 6.6. The secondary structure of the predicted, native and misfolded conformation is also shown in Table 6.6.

As nine of ten simulations converged to native-like conformation without any competing metastable conformations, the folding is concluded as reproducible and predictive.

6.3 Three stranded sheet (GSGS Peptide)

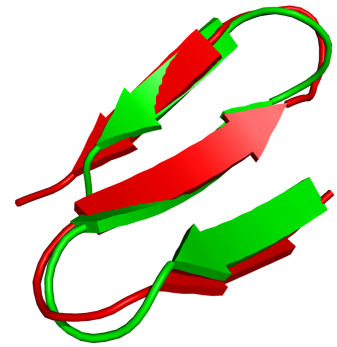
In this section we move our folding studies from simple two stranded β-hairpins to slightly more complicated β-sheet structures. The GSGS peptide is an antiparallel beta sheet with three strands (Alba et al., 1999) which was extensively investigated with phenomenological and all-atom molecular dynamics studies (Wang and S-Sung, 2000; Ferrara and Caflisch, 2000; Caflisch, 2006).

We studied the folding of this three stranded peptide with basin hopping technique in PFF02. We performed 200 cycles of basin hopping simulations for twenty independent simulations. The starting conformations were chosen randomly and had no secondary structure information.

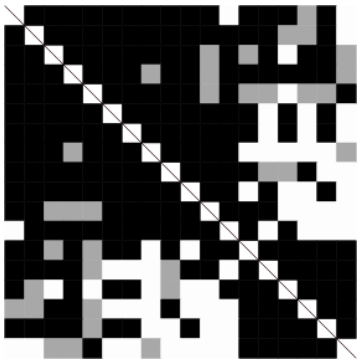
We found that three of four lowest energy trajectories converge to near-native conformations with a backbone root mean square deviation (RMSD) to the native conformation of 2.19,2.26 and 2.67 Å respectively.

The scatter plot of all conformations visited during the simulation are shown in Figure 6.16(c). There are metastable conformations around 4.5 Å and have a random coil conformation. The overlay of the predicted (red) and native (green) conformation is shown in Figure 6.16(a) and the corresponding C_β-C_β distance matrix is shown in Figure 6.16(b). Many blocks in the C_β-C_β overlay matrix are black indicating good agreement of native contacts between the two conformations.

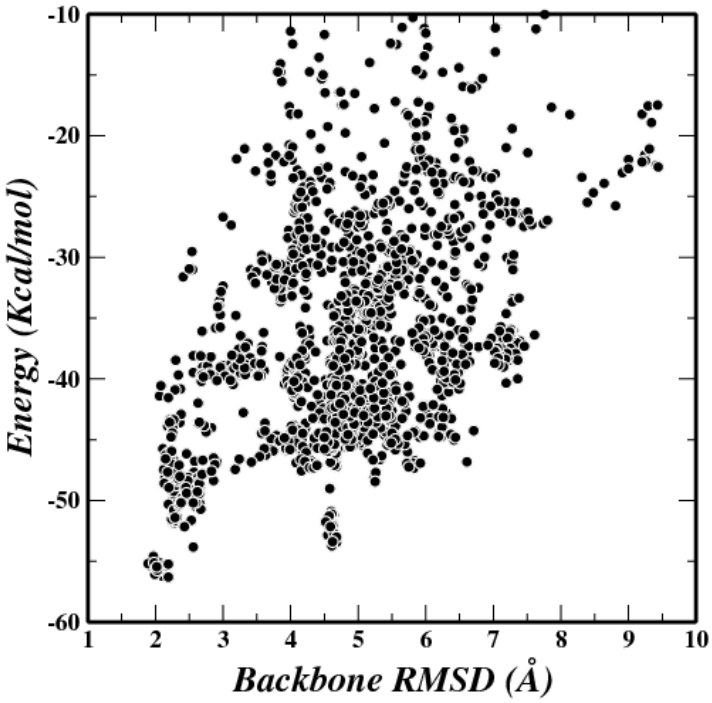
The folded conformatin of the GSGS shows a perfect alignment of the three secondary structure



(a) Overlay



(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.16: GSGS Peptide: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

elements and only small deviations in the loops connecting the defined secondary structure elements. We have performed twenty independent basin hopping simulations on the twenty amino acid GSGS peptide. Predictive reproducible folding of the mini-protein is thereby achieved.

Lined up independently the beta-sheet the regions from (2 to 5, 8 to 13, 16 to 19) agree to within 0.50, 0.55, 0.55 Å with the native conformation. The C_β-C_β distance difference matrix for the GSGS peptide indicates perfect alignment to within experimental resolution.

6.4 Mixed helix/sheet protein 1RIK

In an attempt to fold a mixed protein system which constitutes both helical and sheet elements we studied the folding of the E6-binding zinc finger domain (Liu et al., 2004). Zinc fingers are among the most abundant proteins in eukaryotic genomes and occur in many DNA-binding domains and transcription factors. 1RIK is a 29 amino acid protein with a helix and a small hairpin structure in a $\alpha\beta\beta$ architecture.

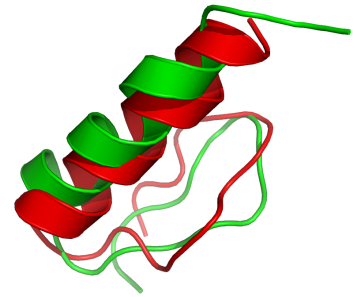
We studied the folding of this zinc finger domain in PFF02 with twenty independent basin hopping

runs with 500 cycles each. The starting conformation for these simulations was completelely extended with an RMSD of 23.5 Å to the native conformation.

Six out of twenty simulations converged to less than 4 Å RMSD to the native conformation. The lowest energy structure has an RMSD of 4.15 Å with the helical region predicted perfectly and certain differences in the beta sheet region. The beta sheet region in the native state is formed in the presence of zinc ion, which is not modelled in the PFF02, indicating the need for further factors to look for in the future.

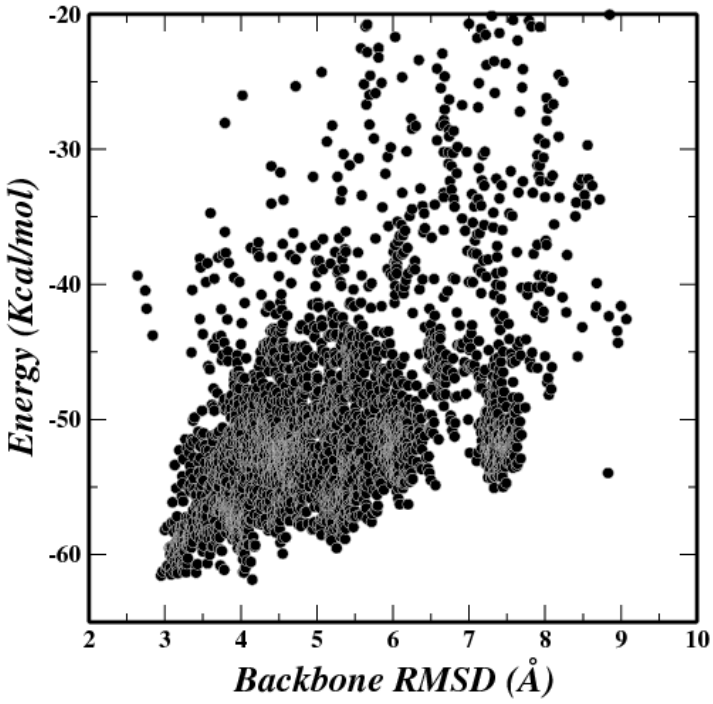
The scatter plot of all conformations visited during the simulation are shown in Figure 6.17(c). The plot indicates the existence of only one folding funnel for this protein. The overlay of the predicted (red) and native (green) conformation is shown in Figure 6.17(a) and the corresponding C_β-C_β distance matrix is shown in Figure 6.17(b). The helical region in the the C_β-C_β overlay matrix are black indicating the prediction of helix region of the protein. Independently the helix region has an RMSD of only 0.86 Å.

While 1RIK folds conformations around 4Å RMSD to the native, the beta sheet is not correctly predicted. This does not imply a failure of the PFF02 force field to predict mixed systems as another mixed protein (1BHI) has folded using the same force field recently (Gopal and Wenzel, 2006).



(a) Overlay

(b) C_β-C_β matrix



(c) Erms Plot

Figure 6.17: 1RIK: Overlay of predicted (red) structure to experimental (green) structure. The overlay of the C_β-C_β distance matrix and Energy vs. RMSD plot.

Discussion

We have located near-native conformations for as many as 13 proteins in PFF02 starting from extended conformations. We started with three helical proteins using the optimized basin hopping technique that were earlier folded in PFF01. The new forcefield correctly predicts the native-like states at lowest energies for these proteins. We then studied folding of larger helical proteins (50-60 amino acids) using evolutionary algorithm. In certain cases, we observe many metastable conformations which have same secondary structures with different arrangement. Mirror images also seem to have competing energies and make selection difficult.

Next we studied six hairpins in PFF02 to study to folding of proteins beta sheet secondary structure. All of the six hairpins fold into native-like conformations with correct pattern of backbone hydrogen bonds. The correct hydrogen bond pattern ensures that the hairpin has the correct bend and the side chains are also projected in the same directions. There are not many helical metastable conformations for these hairpins except the C terminal domain of protein G. This hairpin did not fold predictively and most simulations were stuck in higher energy helical like conformations.

We finally studied a three stranded beta sheet and a mixed protein which constitutes both helix and beta sheets with optimized basin hopping technique. Both these proteins were reproducibly and predictively folded in PFF02.

The overview of these folding simulations is given in Table 6.7.

PDB ID	N	Topology	RMSD (Å)
1L2Y	20	α	3.11
1WQE	23	αα	2.33
1F4I	40	ααα	3.29
1ENH	54	ααα	3.40
1EDK	56	ααα	4.05
1LE0	12	ββ	1.50
1NIZ	14	ββ	2.04
1U6U	17	ββ	4.57
2E4Q	12	ββ	2.62
G Cterm	16	ββ	1.67
1J4M	14	ββ	2.46
GSGS	20	βββ	2.19
1RIK	29	αββ	4.15

Table 6.7: Overview of folding studies in PFF02, G Cterm is the C-terminal hairpin of protein G and GSGS is the synthetic three stranded β-peptide, N indicates the number of amino acids in the protein.

We have therefore studied proteins spanning both helix and sheet secondary structural elements. Five helical, six hairpins, one three stranded β-sheet and one mixed protein were folded in PFF02 using stochastic optimization methods. The average RMSD for the lowest energy structures to their

respective native conformation for these 13 proteins is only 2.87 Å. The study included both helical and sheet like proteins along with a mixed system varying from 12 to 56 amino acids. PFF02 is thus able to predict the native state of a wide range of proteins at the global minimum of their free energy surface and the basin hopping technique and evolutionary algorithm were able to locate this free energy surface.

7

Summary

Proteins are the molecular machines of cellular life. As linear biopolymers they assume a unique 3D conformation, which is encoded in their amino acid sequence. Many disorders and diseases originate primarily due to misfolding of proteins. This misfolding can arise from slight change in environment or mutations in the protein sequence. Therefore it is important to understand protein folding mechanism and the kinetics and dynamics of proteins. Much work has been invested, both experimentally and theoretically, to understand how it is possible that such complex molecules perform their auto-induced folding reaction. Better understanding of intermediate state ensemble and transition states of proteins can help understand their folding pathways. Various methods have been developed to understand protein folding and dynamics. Lattice models were among the first methods and allowed quick and effective sampling of the conformational space of protein molecules. Gō models that incorporated favorable native contact interactions were used to understand some aspects of folding. Presently, molecular dynamics is widely used to study the dynamical behavior of proteins, but this method is limited to time scales of a few hundred nanoseconds (computationally demanding microsecond simulations have been reported for some proteins). As protein folding occurs in the millisecond time, a single simulation study of protein folding with molecular dynamics is not feasible, therefore replica exchange MD has been used to fold some small proteins.

It in addition the number of presently available protein sequences outnumbers the available structures in the Protein Data Bank by a large margin. It is therefore important to develop feasible methods for *de novo* protein structure prediction on the basis of the amino acid sequence alone. Protein structure prediction methods are particularly important for protein families that cannot be addressed experimentally, such as transmembrane proteins. Such methods, even if they are not completely successful for large proteins can help to resolve structures where only insufficient experimental information is available.

From this observation two theoretical challenges arise that are addressed in this thesis:

- Which modles can predict the native state of a protein from sequence information alone?
- Which simulation methods would be needed to achieve this goal?

We approached the folding/structure prediction problem on the basis of Anfinsen's thermodynamic hypothesis that under physiological conditions most proteins are in thermodynamic equilibrium

with their environment. The native three-dimensional conformation then has the lowest free energy of the system. In order to determine the native state of a protein, it is first necessary to have an accurate model for the free energy of competing protein conformations. Secondly, methods are needed to locate the global minimum of this complex free energy reliably and predictively.

We started with a free energy forcefield, which was biased toward helices in protein structure. The goal of this work was to improve the free energy force field PFF01. Secondly we need to develop methods which can effectively locate this minimum starting from completely extended conformations of proteins, *i.e.*, sequence information alone. To this end we optimized and implemented efficient algorithms for locating the free energy minimum.

We succeeded to modify the protein force field PFF01 to obtain a more universal free energy force field, which stabilized a wide range of protein structures. This was achieved by the inclusion of a local electrostatic correction differentiating α-helices from β-sheets. The local electrostatic contribution takes into account the dipole arrangement of NH and CO groups of every residue with its adjoining residues in the protein structure. This differentiates α-helices from β-sheet secondary structures as the dipoles are aligned parallel in α-helices whereas they are antiparallel in β-sheets. This electrostatic correction did not prove enough for folding of several β-hairpins, where the lowest energy conformations were still helical indicating a bias towards helical conformations. In order to reduce this bias, we introduced a weak torsional potential for dihedral angles favoring β-sheets. This torsional potential gives a small contribution to the energy when the dihedral angles of any amino acid (except proline and glycine) that are located around the beta sheet region in the Ramachandran plot. A combination of local electrostatic correction and torsional potential succeeded to fold three β-hairpin proteins (1E0Q, 1A2P and 1K43).

Protein folding with free energy methods is much faster than the direct simulation of the folding pathway by kinetic methods such as molecular dynamics. Using just standard PCs we can fold a simple hairpin with fifteen to twenty amino acids in a matter of hours, at most in a day. Unfortunately even for free energy methods the computational cost rises steeply with the system size and for this reason it is impossible to test the full range of applicability of PFF02 for large family of proteins in a direct folding study. There is, however, an indirect way to test the viability of the free-energy forcefield using a large database (decoy set) of possible conformations for a given protein, including some near-native conformations.

We therefore studied the selectivity of PFF02 by ranking conformations from decoy sets. We used two sets of decoy libraries for our study, PFF01 decoy sets and the Rosetta decoy sets. PFF01 decoy set included decoys generated in the folding studies using PFF01 and we included 32 proteins from the Rosetta decoy set. One criterion to quantify the selectivity of a decoy set is a Z-score, which measures the energy difference between native conformations to the average conformations of a decoy set in units of its standard deviation. We calculated the Z-scores for all the proteins in the decoy sets. For the calculation of energies for the native conformation, we performed relaxation simulations starting from the native structure in PFF02. The average Z-score for the decoy sets was -2.74 and -3.26 for PFF01 and Rosetta decoy set respectively which shows the selectivity of PFF02 to differentiate native-like decoys from the non-native counterparts.

The second ingredient in protein folding studies, aside from the force field, are the simulation

protocols, which ultimately determine whether the global optimum of the forcefield is determined accurately and reliably. We have therefore attempted to develop and adopt such methods, e.g. the stochastic tunneling or the basin hopping technique, which had proven successful in early folding studies for small proteins, in order to find a particularly efficient algorithm. We experimented with all parameters of these methods that included the number of steps and starting and final temperatures.

The basin hopping technique was modified by increasing the number of steps with every basin hopping cycle and the starting temperatures for annealing were taken from an exponential distribution. This protocol increased the convergence of the basin hopping simulations. This protocol was further modified to a "greedy" version, which always retains the best energy conformation found so far. These improvements together increased the speed and reliability of the simulations and resulted in lower final energies, which is the goal for these optimization problems.

One of the key limitations of these methods is that they maps the global optimization problem onto a single fictitious dynamical process. In this type of simulation protocol, the molecule constructs one trajectory starting somewhere in the unfolded ensemble, which hopefully converges towards the native conformation. Even with standard basin hopping simulations, several simulations are necessary to obtain a predictive and reproducible result. In the standard protocol the simulations are completely independent of one another. This raises the obvious question, whether an improved convergence can be obtained by coupling a number of concurrent dynamical processes. The second, related question concerns the largest number of concurrent processes that can be coupled together to speed the overall search. In this respect optimization based methods have a significant advantage over traditional kinetic methods, because the latter must ultimately strive to construct one single consecutive trajectory. The only option to speed the simulation for a single trajectory is the parallelization of the energy and force evaluation, which requires a large amount of data transfer. The optimization methods using a large number of concurrent dynamical processes, on the other hand, are able to use coarse-grained strategies in which a single processor performs one of many largely independent simulations.

We have implemented an evolutionary algorithm on massively parallel architectures such as the BlueGene computer. The algorithm is implemented in a master-client model which keeps a diverse population on the master and the clients sample the protein landscape simultaneously and return to the master. The algorithm scales very well with the number of processors used (up to 4096 tested on the IBM BlueGene). Using this algorithm we folded various proteins such as 40 amino acid HIV accessory protein (1F4I) and 54 amino acid engrailed homeodomain protein (1ENH) in a single day. The folding of the engrailed homeodomain protein was carried out in a single day using 512 processors on the Barcelona Mare Nostrum Supercomputer, the current largest supercomputer in Europe. This is a great achievement as the folding of a protein of comparable size required about 4 months using 50 processors in earlier studies. The folding of the tryptophan zipper proteins (1LE0) was possible in only 14 minutes using 128 processors.

Using PFF02 along with modified versions of the basin hopping technique, we could fold several protein structures starting from completely extended conformations. These include various helical proteins, the tryptophan cage protein (1L2Y), the HIV-accessory protein (1F4I) and a potassium channel blocker protein (1WQC) which were earlier folded in PFF01. The tryptophan cage protein is a widely studied model for protein folding both theoretically and experimentally. The HIV acces-

sory protein and potassium channel blockers are biologically important proteins. The HIV accessory protein destroys the host cell's ability to survive by binding to a host receptor and restricting an important enzyme to activate the cell's immune system. Potassium channel blockers are toxic venom peptides involved in blocking of potassium channel in cells. We also folded much larger and widely studied model proteins (both experimentally and theoretically) like the engrailed homeodomain protein (1ENH) and E-domain of the of Staphylococcal Protein A (1EDK), which were folded with the evolutionary algorithm.

We then investigated the folding of various β-hairpins in PFF02. These hairpins included tryptophan zipper protein (1LE0), HIV-1 V3 loops (1NIZ, 1U6U), designed stable beta proteins (2EDK, 1J4M) and the C terminal hairpin of G protein to experimental resolution. The tryptophan zipper protein and C terminal hairpin of protein G have been subjected to many theoretical and experimental studies. The HIV-1 V3 loops are highly homologous loops which have a different hydrogen bonding pattern responsible for co-receptor selectivity by the virus. The loop conformation is responsible for selecting infection of T-cells or macrophages. The folding of these loops (1NIZ and 1U6U) is particularly encouraging because PFF02 can distinguish these very similar sequences and correctly predicts a one residue shift in backbone hydrogen bonding pattern resulting in different side chains orientation responsible for co-receptor selectivity of the virus protein. The experimentally stable hairpins serve as good model systems for studies on beta sheet formation and folding.

Apart from two stranded β-hairpins, we also studied the folding of the three stranded GSGS peptide. The GSGS peptide is a designed stable three stranded beta sheet with glycine-serine (GS) bends and has been a model system for three stranded beta sheet formation. We finally studied the folding of αββ zinc finger domain protein 1RIK. Zinc fingers are among the most abundant proteins in eukaryotic genomes and occur in many DNA-binding domains and transcription factors.

In this thesis, we succeeded to fold near-native conformations for 13 proteins spanning both helix and sheet secondary structural elements in PFF02 starting from extended conformations. In certain cases, we observe many metastable conformations which have same secondary structures with different tertiary arrangements. Mirror image conformations often have comparable energies and make selection difficult. For beta sheets we located the near-native decoys with correct backbone hydrogen bonding pattern which ensures that the hairpin has the correct bend and the side chains project in the same directions as in the native state. The GSGS peptide and mixed protein 1RIK both reproducibly and predictively folded in PFF02. The average RMSD for the lowest energy structures to their respective native conformation for these 13 proteins is only 2.87 Å. PFF02 is thus able to predict the native state of a wide range of proteins at the global minimum of their free energy surface and the basin hopping technique and evolutionary algorithm were able to efficiently locate their global minima on a complex free energy surface.

Outlook

In this thesis we developed methods to find the native state of various proteins by locating the global minimum of the free energy surface. There are, however, a large number of questions that remain to be addressed. Fortunately there are complementary methods , which in combination with the free-energy

methodology developed here, can address these problems. For example, we have neglected the details of the kinetics of protein folding in our approach. As stated earlier, its important to study kinetics of folding to understand protein folding mechanism and to predict folding rates. Because free-energy methods sample exhaustively the low-energy conformations of the protein that are accessible under physiological conditions it may be possible to reconstruct the folding kinetics on the basis of that ensemble of conformations. This can be achieved by a dynamical analysis of the low energy region by using master equations assuming diffusive processes between similar conformations.

A related interesting aspect of protein folding is the study of transition states. Transition states are the saddle points of the free energy surface that connect the unfolded state and the folded state. Computationally transition states can be determined by a so-called p-fold analysis, *i.e.* searching for protein conformations that fold or unfold with the equal probability at some finite temperature. Experimentally, transition-state analysis is carried out by mutating the sequence of the protein and measuring the changes in kinetics and equilibria of protein folding (psi value analysis). This raises the question of protein stability under mutations. The latter question can be addressed by computing the free-energy difference between the folded and unfolded ensemble for a variety of mutations.

Also, further developments could be made in the direction of protein-protein interactions. These studies can help understand protein aggregation which are responsible for various diseases, such as Alzeimer or Parkinson's disease. We have already implemented modules in our simulation package that can treat protein-protein interactions and the first studies regarding protein-protein docking are presently under way.

The protein force field PFF02 still does not incorporate the formation of disulphide bridges. These disulphide bridges provide stabilizing to many protein structures with the formation of covalent bonds. More studies are needed in this direction to understand wider range of protein structures.

Finally we must address the question how we can fold even larger proteins, with more than one hundred amino acids. We have encountered the problem of freezing when studying large proteins. Once the protein collapses, it is difficult to generate no clashing Monte Carlo moves, which leads to poor acceptance ratio. As a result the protein cannot explore the conformational space. Further development of methods which are faster in locating the global minimum is still required to study all atom folding of proteins over hundred amino acids. One possible solution to this is by splitting the protein into fragments and later joining them to obtain tertiary structure. This method can generate native like conformations for the protein which can be further relaxed and identified. Such methods have been very useful in the field of protein structure prediction.

With the development of the all-atom protein forcefield (PFF02) we have made a significant step towards a universal free-energy approach to protein folding and structure prediction. The massively parallel simulation methods developed in the last few years now permit the protein folding of medium-size proteins from random initial conformations. This work thus lays the foundations to further explore the mechanism of protein folding, to understand protein stability and ultimately develop methods for *de novo* protein structure prediction.

Appendix A

Programs and definitions

RMSD

Mathematically the root mean square deviation can be defined as

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\overrightarrow{A_i} - \overrightarrow{B_i}\right)^2}$$

where A and B are two sets of N points and $\overrightarrow{A_i}$ and $\overrightarrow{B_i}$ are position vectors containing x, y and z components. It is the most common definition used to compare two different protein structures. RMSD between two proteins can be calculated considering all atoms of the protein, backbone of the protein or only C_α atoms of the protein structures. In this thesis, the RMSD's are always calculated for the complete backbone, *i.e.*, N, C_α, C and O atoms of the backbone.

Before the calculation of RMSD, the two proteins must be aligned together, *i.e.*, one of the conformation must be rotated and translated to give the best possible fit. Thus the RMSD of two protein conformations refers to the lowest possible RMSD possible after all translations and rotations.

C_β-C_β overlay matrix

A C_β-C_β overlay matrix allows a quick optical comparison for the overlay of two structures. To generate it the relative distances of all C_β atoms of the two conformations are calculated as a C_β distance matrix.Then the difference between each entry in these two distance matrices is the C_β-C_β matrix. It is defined as

$$CBCB(A,B)_{ij} = \left|\Delta C_\beta(A_i,A_j) - \Delta C_\beta(B_i,B_j)\right| \tag{A.1}$$

where $CBCB(A,B)_{ij}$ is i,j entry of the matrix and $\Delta C_\beta(A_i,A_j)$ is the relative difference of C_β for i^{th} and j^{th} atom.

To get a visual representation of this matrix it is color coded depending upon the corresponding difference. If this difference is less than 1.5 Å the according pixel on the C_β-C_β overlay matrix is

colored black, for differences between 2.25 Å and 1.5 Å it is colored grey and white for the distances greater than 2.25 Å.
This C_β-matrix gives a good overview about secondary structure and the alignment of the two conformations. The local regions along the diagonal correspond to the secondary structure since the position of the C_β is same if the secondary structure is the same. In the regions further away from the diagonal, it indicates the arrangement with regions further away in the sequence, thus indicating the overall tertiary agreement.

DSSP

Dictionary of Secondary Structure of Proteins (DSSP) (Kabsch and Sander, 1983) was used to characterize the secondary structure of proteins. The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling et al. in 1951. The different assignments and the corresponding letters used are:

- G = 3-turn helix (3_10 helix). Min length 3 residues.
- H = 4-turn helix (alpha helix). Min length 4 residues.
- I = 5-turn helix (pi helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = beta sheet in parallel and/or anti-parallel sheet conformation (extended strand). Min length 2 residues.
- B = residue in isolated beta-bridge (single pair beta-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment)

Although the DSSP formula is a relatively crude approximation of the physical hydrogen bond energy, it is generally accepted as one of the standard tool for defining secondary structure of proteins.
`http://swift.cmbi.ru.nl/gv/dssp/`

Molmol

Molmol (Koradi et al., 1996) was used to locate the backbone hydrogen bonds within a protein conformation using the default distance and angle cutoff defined in the program. The default valued used in this thesis for distance cutoff and maximum angle are 2.4 Å and 35°respectively.
`http://hugin.ethz.ch/wuthrich/software/molmol/`

Pymol

Pymol (DeLano, 2002) was used for visual representation of protein structures and the generation of overlay pictures for the thesis.
`http://pymol.sourceforge.net/`

Protein Data Bank

All experimentally resolved biomolecular structures are deposited in a database called Protein Data Bank (PDB). The proteins in the database are identified by a unique four-letter identifier called PDB-ID. There are currently over 40,000 entries in PDB and is growing. The database can be accessed freely under at http://www.pdb.org.

The experimental conformations for the proteins under study in this thesis (except GSGS peptide) were taken from this database. There are multiple models in case of structures resolved using NMR. In these cases, for sake of uniformity, we have always calculated RMSD's with respect to the first model in the PDB file.
`http://www.pdb.org`

Generation of extended conformations

The extended conformations were generated by setting dihedral angles of all amino acids (except proline) to 180°. These were generated using either Pymol (DeLano, 2002) or Molden (Schaftenaar and Noordik, 2000).

POEM

For all the simulations reported in the thesis, the simulation package POEM was used. It is written using C and allows for calculation of energies using PFF02 and perform various optimization methods. The evolutionary algorithm was implemented in the same package and used on multi-teraflop architectures. It was compiled using GNU C Compiler and used under Linux.

Xmgrace

All the graphs in this thesis were generated using Xmgrace which runs on unix-like systems.
`http://plasma-gate.weizmann.ac.il/Grace/`

LaTeX

The thesis was typesetted using LaTeXusing the Kile-1.8 integrated environment for LaTeX.
`http://www.latex-project.org/ http://kile.sourceforge.net/`

Bibliography

R. A. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002, 1994.

R. A. Abagyan and M. Totrov. Ab initio folding of peptides by the optimal bias monte carlo minimization procedure. *J. Comp. Phys.*, 151:402–21, 1999.

E. De Alba, J. Santoro, M. Rico, and M. A. Jimenez. De novo design of a monomeric three-stranded antiparallel β-sheet. *Prot. Sci.*, 8:854–65, 1999.

N. H. Andersen, K. A. Olsen, R. M. Fesinmeyer, X. Tan, F. M. Hudson, L. A. Eidenschink, and S. R. Farazi. Minimization and optimization of designed β-hairpin folds. *J. Am. Chem. Soc.*, 128: 6101–10, 2006.

C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–30, 1973.

F. Avbelj. Use of a potential of mean force to analyze free energy contributions in protein folding. *Biochemistry*, 31:6290–6297, 1992.

F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34:755–764, 1995.

R. Baldwin. Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci. USA*, 83:8069–72, 1986.

O.M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106:1495–1517, 1997.

J. Berg, J. Tymoczko, and L. Stryer. *Biochemistry*. Michelle Julet, 2001.

R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker. Rosetta in CASP4: progress in ab-initio protein structure prediction. *Proteins*, 45:119–126, 2001.

C. Branden and J. Tooze. *Introduction to protein structure.* Routledge, 1999.

A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr. Op. Struct. Biol.*, 16: 71–8, 2006.

G. Casari and M. J. Sippl. Structure derived hydrophobic potentials. a hydrophobic potential derived from x ray structures of globular proteins is able to identify native folds. *J. Molec. Biol.*, 224: 725–732, 1992.

B. Chagot, C. Pimentel, L. Dai, J. Pil, J. Tytgat, T. Nakajima, G. Corzo, H. Darbon, and G. Ferrat. An unusual fold for potassium channel blockers: NMR structure of three toxins from the scorpion opisthacanthus madagascariensis. *Biochem. J.*, 388:263–71, 2005.

H. S. Chan and K. A. Dill. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins: Struc. Func. and Gen.*, 30:2–33, 1998.

M. E. Clamp, P. J. Baker, C. J. Sterling, and A. Brass. Hybrid Monte Carlo: An effective algorithm for condensed matter simulations. *J. Chem. Phys*, 15:838–46, 1994.

J. B. Clarge, T. Romo, B. K. Andrews, B. M. Pettitt, and Jr. G. N. Philipps. A sampling problem in molecular dynamics simulations of macromolecules. *Proc. Natl. Acad. Sci. USA*, 92:3288–92, 1995.

N. D. Clarke, C. R. Kissinger, J. Desjarlais, G. L. Gilliland, and C. O. Pabo. Structural studies of the engrailed homeodomain. *Protein Sci.*, 3:1779–87, 1994.

A. G. Cochran, N. J. Skelton, and M. A. Starovasnik. Tryptophan zippers: stable, monomeric β-hairpins. *Proc. Natl. Acad. Sci. USA*, 98:5578–83, 2001.

V. Daggett and A. Fersht. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell. Biol.*, 4:497–502, 2003.

W. L. DeLano. *The PyMOL User's Manual*. DeLano Scientific, San Carlos, CA, USA, 2002.

K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–55, 1990.

K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding–a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

K.A. Dill and H.S. Chan. From levinthal to pathways to funnels: The "new view" of protein folding kinetics. *Nature Structural Biology*, 4:10–19, 1997.

F. Ding, S. V. Buldyrev, and N. V. Dokholyan. Folding trp-cage to nmr resolution native structure using a coarse-grained protein model. *Biophys. J.*, 88:147–55, 2005.

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

D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319: 199–203, 1986.

J.L. Fauchere and V. Pliska. Hydrophobic parameters pi of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. med. Chem.-Chim. ther.*, 18:369–375, 1983.

P. Ferrara and A. Caflisch. Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA*, 97(20):10780–10785, 2000.

S. M. Gopal and W. Wenzel. De novo folding of the DNA-binding ATF-2 zinc finger motif in an all-atom free-energy forcefield. *Angew. Chem. Int. Ed.*, 45:7726–8, 2006.

A. T. Hagler and C. S. Ewig. On the use of quantum energy surfaces in the derivation of molecular force fields. *Comp. Phys. Comm.*, 84:131–55, 1994.

U. H. E. Hansmann. Global optimization by energy landscape paving. *Phys. Rev. Letters*, 88:068105, 2002.

T. Herges. *Entwicklung eine Kraftfelds zur Strukturvorhersage von Helixproteinen.* PhD thesis, Dept. of Physics. University of Dortmund, 2003.

T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.*, 87:3100–9, 2004.

T. Herges and W. Wenzel. Reproducible in-silico folding of a three-helix protein and characterization of its free energy landscape in a transferable all-atom forcefield. *Phys. Rev. Lett.*, 94:018101, 2005a.

T. Herges and W. Wenzel. Characterization of the free energy landscape of the villin headpiece in an all-atom forcefield. *Structure*, 13:661–668, 2005b.

M. Hollecker and T. E. Creighton. Effect on protein stability of reversing the charge on amino groups. *J. Mol. Biol.*, 701:395–404, 1982.

S. A. Islam, M. Karplus, and D. L. Weaver. The role of sequence and structure in protein folding kinetics: The diffusion-collision model applied to proteins L and G. *Structure*, 12:1833–45, 2004.

W. L. Jorgensen and J. Tirado-Rives. Monte carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.*, 100:14508–13, 1996.

W. L. Jorgeson. Quantum and statistical mechanical studies of liquids. 11. Transferable intermolecular potential functions. application to liquid methanol including internal rotation. *J. Am. Chem. Soc.*, 103:341, 1981.

J. Juraszek and P. G. Bolhuis. Sampling the multiple folding mechanisms of trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. USA*, 103:15859–64, 2006.

W. Kabsch and C. Sander. A dictionary of protein secondary structure. *Biopolymers*, 22:2577–2637, 1983.

S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220: 671–80, 1983.

R. Koradi, M. Billeter, and K. Wuetrich. Molmol: A program for display and analysis of macromolecular structures. *J. Mol. Graph.*, 14:51–5, 1996.

M. Krishnan, A. Verma, and S. Balasubramanian. Computer simulation study of water using a fluctuating charge model. *Proc. Ind. Acad. Sci (Chem. Sci)*, 113:579–90, 2001.

L. J. LaBerge and J. C. Tully. A rigourous procedure for combining molecular dynamics and Monte Carlo simulation algorithms. *Chem. Phys.*, 260:183–91, 2000.

T. Lazaridis and M. Karplus. "new view" of protein folding reconciled with the oldthrough multiple unfolding simulations. *Science*, 278:1928–1931, 1997.

A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd., 2001.

B. K. Lee and F. M Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Molec. Biol.*, 79:379–400, 1971.

P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA*, 89:8721–25, 1992.

A. M. Lesk. *Introduction to Protein Architecture*. Oxford University Press, 2001.

C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65:44, 1968.

M. Levitt. Protein conformation, dynamics and folding by computer simulation. *Ann. Rev. Biophys. Bioeng.*, 11:251–71, 1982.

A. Linhananta, J. Boer, and I. MacKay. The equilibrium properties and folding kinetics of an all-atom go- model of the trp-cage. *J. Chem. Phys*, 122:1–15, 2005.

Y. Liu, Z. Liu, E. Androphy, J. Chen, and J. D. Baleja. Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. *Biochemistry*, 43:7421–31, 2004.

A. D. Mackerell. Atomistic models and forcefields. In M. Watanabe, B. Roux, A. MacKerell, and O. Becker, editors, *Computational Biochemistry and Biophysics*, pages 7–38. Marcel Dekker, 2000.

A. D. MacKerell, B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. Charmm: The energy function and its parameterization with an overview of the program. *The Encyclopedia of Computational Chemistry*, 1:271–277, 1998.

G. I. Makhatadze and P. L. Privalov. On the entropy of protein folding. *Protein Sci.*, 5:501–510, 1996.

S. A. Mason. Origins of biomolecular handedness. *Nature*, 311:19–23, 1984.

B. W. Matthews. Genetic and structural analysis of the protein stability problem. *Biochemistry*, 26: 6885–8, 1987.

Y. Mauguen, R. W. Hartley, E. J. Dodson, G. G. Dodson, G. Bricogne, C. Chothia, and A. Jack. Molecular structure of a new family of ribonucleases. *Nature*, 297:162–64, 1982.

U. Mayor, N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato S, G. S. Jas, S. M. Freund, D. O. Alonso, V. Daggett, and A. R. Fersht. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*, 421:863–7, 2003.

I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potentials in protein folding. *J. Molec. Biol.*, 238:777–793, 1994.

P. N. Mortenson and D. J. Wales. Energy landscapes, global optimisation and dynamics of the polyalanine Ac(ala)8 NHMe. *J. Chem. Phys.*, 114:6443–54, 2001.

P. N. Mortenson, D. A. Evans, and D. J. Wales. Energy landscapes of model polyalanines. *J. Chem. Phys.*, 117:1363–76, 2002.

A. Nayeem, J. Vila, and H. A. Scheraga. A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. *J. Comp. Chem.*, 12:594–605, 1991.

J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. Designing a 20-residue protein. *Nat. Struct. Biol.*, 9:425–30, 2002.

P. H. Nguyen. Complexity of free energy landscapes of peptides revealed by nonlinear principal component analysis. *Proteins: Struc. Func. and Gen.*, 65:893–913, 2006.

P. H. Nguyen, G. Stock E. Mittag, C. K. Hu, and M. S. Li. Free energy landscape and folding mechanism of a β-hairpin in explicit water: A replica exchange molecular dynamics study. *Proteins: Struc. Func. and Gen.*, 61:705–808, 2005.

J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Ann. Rev. Phys. Chem*, 408:545:600, 1997.

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*, 5:1093–108, 1997.

B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Molec. Biol.*, 258:367, 1996.

M. T. Pastor, M. Lopez de la Paz, E. Lacroix, L. Serrano, and E. Perez-Paya. Combinatorial approaches: a new tool to search for highly structured β-hairpin peptides. *Proc. Natl. Acad. Sci. USA*, 99:614–9, 2002.

S. Patel and C.L. Brooks III. Fluctuating charge force fields: Recent developments and applications from small molecules to macromolecular biological systems. *Molecular Simulation*, 32:231–49, 2006.

D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. Amber, a package of computer programs for applying molecular

mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, 91:1–41, 1995.

P. L. Privalov. Stability of proteins: small globular proteins. *Adv. Prot. Chem.*, 33:167–241, 1979.

P. L. Privalov and S. J. Gill. Stability of protein structure and hydrophobic interaction. *Adv. Prot. Chem.*, 39:191–234, 1988.

G. N. Ramachandran and A. K. Mitra. An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. *J. Mol. Biol.*, 107:85–92, 1976.

G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–9, 1963.

S. W. Rick, S. J. Stuart, and B. J. Berne. Dynamical fluctuating charge force fields: Application to liquid water. *J. Chem. Phys.*, 101:6141–56, 1994.

D. R. Ripoll, J. A. Vila, and . A. Scheraga. On the orientation of the backbone dipoles in native folds. *Proc. Natl. Acad. Sci. USA*, 102:7559–7564, 2005.

O. Rosen, J. Chill, M. Sharon, N. Kessler, B. Mester, S. Zolla-Pazner, and J. Anglister. Induced fit in HIV-neutralizing antibody complexes: Evidence for alternative conformations of the gp120 V3 loop and the molecular basis for broad neutralization. *Biochemistry*, 44:7250–8, 2005.

G. Schaftenaar and J.H. Noordik. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Design*, 14:123–34, 2000.

T. Schlick. *Molecular Modeling and Simulation: An interdiscilinary guide*. Springer-Verlag New York, 2002.

J. Schneider, I. Morgenstern, and J. M. Singer. Bouncing towards the optimum: Improving the results of monte carlo optimization algorithms. *Phys. Rev. E*, 58:5085–95, 1998.

A. Schug. *Free-energy simulations using stochastic optimization methods for protein structure prediction*. PhD thesis, Dept. of Physics. University of Dortmund, 2005.

A. Schug and W. Wenzel. Reproducible folding of a four helix protein in an all-atom forcefield. *J. Am. Chem. Soc.*, 126(51):16736–16737, 2004.

A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Letters*, 91:158102, 2003a.

A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.*, 91:1581021–4, 2003b.

A. Schug, T. Herges, and W. Wenzel. All-atom folding of the three-helix hiv accessory protein with an adaptive parallel tempering method. *Proteins*, 57:792–8, 2004a.

A. Schug, T. Herges, and W. Wenzel. All atom folding of the three helix hiv accessory protein with an adaptive parallel tempering method. *Proteins*, 57(4):792–798, 2004b.

A. Schug, B. Fischer, A. Verma, W. Wenzel, and G. Schoen. Biomolecular structure prediction with stochastic optimization methods. *Adv. Eng. Mat.*, 7:1005–9, 2005a.

A. Schug, W. Wenzel, and U. H. E. Hansmann. nergy landscape paving simulations of the trp-cage protein. *J. Chem Phys.*, 122:1–7, 2005b.

A. Schug, T. Herges, A. Verma, K. H. Lee, and W. Wenzel. Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. *Chemphyschem*, 6:2640–6, 2006.

W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren. The gromos biomolecular simulation program package. *J. Phys. Chem.*, 103:3596–3607, 1999.

M. Sharon, N. Kessler, R. Levy, S. Zolla-Pazner, M. Gorlach, and J. Anglister. Alternative conformations of HIV-1 V3 loops mimic β-hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. *Structure*, 11:225–236, 2003.

K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig. Extracting hydrophobic free energies from experimental data:relationship to protein folding and theoretical models. *Biochemistry*, 30:9686–9697, 1991.

M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials from crstal data. 6. determination ofemperical potentials for o-h$\cdots$o=c hydrogen bonds from packing configurations. *J. Phys. Chem.*, 88:6231–6233, 1984.

J. Skolnick and A. Kolinski. Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.*, 40:207–35, 1989.

C. D. Snow, B. Zagrovic, and V. S. Pande. The trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.*, 124:14548–9, 2002.

M. A. Starovasnik, N. J. Skelton, M. P. O'Connell, R. F. Kelley, D. Reilly, and W. J. Fairbrother. Solution structure of the E-domain of staphylococcal protein A. *Biochemistry*, 35:15558–69, 1996.

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*, 53:76–87, 2003.

J. P. Ulmschneider, M. B. Ulmschneider, and A. De Nola. Monte carlo vs molecular dynamics for all-atom polypeptide folding simulations. *J. Phys. Chem. B*, 110:16733–42, 2006.

A. Verma and W. Wenzel. Predictive and reproducible de-novo all-atom folding of a β-hairpin loop in an improved free energy forcefield. *J. Phys. Cond. Matt*, 19:285213, 2007a.

A. Verma and W. Wenzel. All-atom protein folding in a single day. submitted, 2006a.

A. Verma and W. Wenzel. Protein structure prediction by all-atom free-energy refinement. *BMC Structural Biology*, 7:12, 2007b.

A. Verma and W. Wenzel. De-novo all atom folding of a HIV-1 V3 hairpin loop in an imporved free energy forcefield. Submitted, 2006b.

A. Verma and W. Wenzel. Towards an all-atom free-energy forcefield for protein folding. in preparation, 2006c.

A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *J. Chem. Phys.*, 124:044515, 2006.

A. Verma, S. M. Gopal, J. Oh, K. H. Lee, and W. Wenzel. All atom de-novo protein folding with a scalable evolutionary algorithm. *J. Comput. Chem.*, in press, 2007.

D. J. Wales and P. E. J. Dewsbury. Effect of salt bridges on the energy landscape of a model protein. *J. Chem. Phys.*, 121:10284–90, 2004.

D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A.*, 101:5111–6, 1997.

H. Wang and S. S-Sung. Molecular dynamics simulations of three-strand β-sheet folding. *J. Am. Chem. Soc.*, 122:1999–2009, 2000.

D. W. Weatherford and F. R. Salemme. Conformations of twisted parallel β-sheets and the origin of chirality in protein structures. *Proc. Natl. Acad. Sci. USA*, 76:19–23, 1979.

W. Wenzel. Predictive folding of a β-hairpin in an all-atom free-energy model. *Europhys. Lett.*, 76: 156–162, 2006.

E. S. Withers-Ward, T. D. Mueller, I. S. Chen, and J. Feigon. Biochemical and structural analysis of the interaction between the UBA(2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. *Biochemistry*, 39:14103–12, 2000.

R. Zerella, P. Y. Chen, P. A. Evans, A. Raine, and D. H. Williams. Structural characterization of a mutant peptide derived from ubiquitin: implications for protein folding. *Protein Sci.*, 119:2142–50, 2000.

R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA*, 98:14931–36, 2001.

Acknowledgments

It is a pleasure to thank many people who made this thesis possible.

It is difficult to overstate my gratitude to my Ph.D. supervisor, Dr. Wolfgang Wenzel. With his enthusiasm, his inspiration, and his great efforts to explain things clearly and simply, he guided me with my thesis on a highly interdisciplinary field of research. I would also like to thank Dr. Frank Schmitz at Institut für Wissenschaftliches Rechnen for all the help and support during my thesis.

I would like to thank my teachers who have taught me physics (especially Prof. Sharashchandra Patil and Prof. Yogendra Gambhir) for their kind support, wise advice and motivation. I would also like to thank Prof. Liisa Holm at University of Helsinki for opening up the world of fascinating proteins. Additionally I would like to thank my colleagues at Indian Institute of Technology Bombay for providing and fun environment and motivating me towards scientific research.

I am indebted to my colleagues for providing a stimulating and fun environment to learn. I am especially grateful to Dr. Holger Merlitz, Dr. Konstantin Klenin, Dr. Alexander Schug, Aina Quintilla, Srinivasa M. Gopal and Bernhard Fischer who were always helpful and congenial. In addition, I would like to thank Institute of Nanotechnology to provide an excellent working environment during my stay. I am also grateful to the entire cricket team at Karlsruhe who made my stay fun, enjoyable and exciting.

I would also like to thank my friends Harshad Joshi and Prasad Phatak for helping me get through the difficult times, and for all the emotional support, camaraderie, entertainment, and caring they provided. Harshad Joshi specially helped me learn some basics of molecular dynamics simulations.

I am grateful to the Christine Batsch and Erika Schütze for assisting me in various different tasks during my stay at Institute of Nanotechnology. I am also thankful to Institut für Wissenschaftliches Rechnen and Forschungszentrum Karlsruhe for providing excellent infrastructure and funding during my thesis. I would also like to thank Dr. K H Lee for providing with the computational resources at the KIST supercomputer and the Barcelona Supercomputing Center for access to the MareNostrum supercomputing facility.

I wish to thank my entire family for providing a loving environment for me especially my sister and brother-in-law.

Lastly, and most importantly, I wish to thank my parents. They raised me, supported me, taught me, and loved me. To them I dedicate this thesis.

List of Publications

- A. Verma & W. Wenzel, *Towards a uniform free energy approach for all atom protein folding.* (2007) (*submitted*)
- A. Verma & W. Wenzel. *De-novo all atom folding of a HIV-1 V3 hairpin loop in an imporved free energy forcefield.* (2007) (*submitted*)
- A. Verma, S.M. Gopal, J.S. Oh, K.H. Lee & W. Wenzel, *All-atom de novo protein folding with a scalable evolutionary algorithm*, **J. Comp. Chem** (2007) (*in press*)
- A. Verma & W. Wenzel, *Protein structure prediction by all-atom free-energy refinement*, **BMC Structural Biology**, **7**:12 (2007)
- A. Verma & W. Wenzel, *Reproducible and predictive de novo all atom folding of a* β*-hairpin loop in an improved free energy forcefield*, **J. Phys.: Condens. Matter**, **19**:285213 (2007)
- A. Verma, A. Schug, K. H. Lee & W. Wenzel, *Basin hopping simulations for all-atom protein folding*, **J. Chem. Phys.**, **124**:044515 (2006)
- A. Schug, B. Fischer, A. Verma, H. Merlitz, W. Wenzel & G. Schoen, *Biomolecular structure prediction with stochastic optimization methods.*, **Adv. Eng. Mat.**, **7**:1005 (2005)
- A. Schug, T. Herges, A. Verma, K. H. Lee & W. Wenzel, *Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein.*, **ChemPhysChem**, **6**:2640 (2005)
- A. Schug, T. Herges, A. Verma & W. Wenzel, *Investigation of the parallel tempering method for protein folding*, **J. Phys. Condens. Matter**, **17**:S1641 (2005)