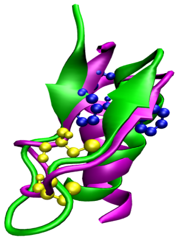
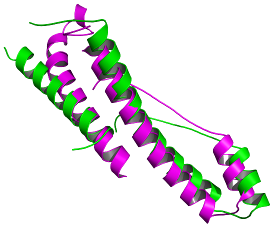
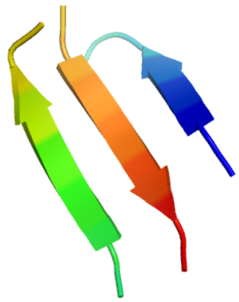
Protein folding, structure prediction and aggregation studies using a free-energy forcefield

DISSERTATION

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







vorgelegt von

Srinivasa Murthy Gopal
aus
Bengaluru, Indien

To My Parents and My Sister

Tag der mündlichen Prüfung	:	11. Juni 2007
Vorsitzender	:	Prof. Dr. Metin Tolan
1. Gutachter	:	Priv.-Doz. Dr. Wolfgang Wenzel
2. Gutachter	:	Prof. Dr. Joachim Stolze
Vertreter der promovierten wissenschaftlichen Mitarbeiter	:	Dr. Michael Vogel

Protein folding, structure prediction and aggregation studies using a free-energy forcefield

Abstract

Proteins are versatile molecules which perform multitude of functions in living organisms. The function of a protein depends on the precise three-dimensional structure which it attains under physiological conditions. Further, protein-protein interactions are responsible for many biochemical mechanisms in living organisms. Therefore the study of protein structure, protein folding and interactions is essential for understanding of biological processes. In recent years theoretical methods have increasingly complemented experiments in elucidating protein structure and function. This work pursues a de-novo protein modeling approach based on Anfinsen's thermodynamic hypothesis, which states that a protein in its native state is in thermodynamic equilibrium with its environment. The biologically active conformation thus corresponds to a global minimum of the free energy. We have developed a free-energy model for proteins (PFF02) in conjugation with efficient optimization methods for protein folding and structure prediction. With this approach a zinc finger motif was folded using an efficient evolutionary algorithm starting from an extended structure. In addition, we elucidated the folding characteristics of this protein by analyzing its energy landscape. We devised a de novo methodology for predicting the native structure of proteins, which was used to predict the structure of 27 targets in the CASP7 competition. Our method was quite successful for the free modeling targets. We investigated the aggregation of a fragment of amyloid beta protein, which is believed to play a key role in the aggregation of the full protein. A general computational scheme for protein-protein docking was developed and tested successfully for two protein dimers.

Proteinfaltung, Strukturvorhersage und Protein-Aggregation mit einem Kraftfeld für die freie Energie

Zusammenfassung

Proteine sind vielseitige Moleküle, die eine grosse Anzahl von Funktionen im lebenden Organismus erfüllen. Die Funktion eines Proteins wird von seiner genauen dreidimensionalen Struktur bestimmt, die unter physiologischen Bedingungen zumeist spontan angenommen wird. Darüber hinaus sind Protein-Proteinwechselwirkungen verantwortlich für viele biochemische Steuerungsmechanismen von Organismen. Aus diesem Grunde ist die Untersuchung von Proteinstrukturen, der Proteinfaltung und von Proteinwechselwirkungen wichtig für das Verständnis biologischer Vorgänge. In den vergangenen Jahren haben theoretische Methoden zunehmend experimentelle Untersuchungen zu Struktur und Funktion von Proteinen unterstützt. In dieser Arbeit wird ein Ansatz zur de-novo Proteinmodellierung verfolgt, der sich auf Anfinsens thermodynamische Hypothese stützt, nach der Proteine in ihrem nativen Zustand sich im thermodynamischen Gleichgewicht mit ihrer Umgebung befinden. Die biologisch aktive Konformation entspricht daher dem globalen Minimum der freien Energie. Wir entwickelten ein Modell für die freie Energie von Proteinen (PFF02) und effiziente Optimierungsverfahren für die Proteinfaltung und Strukturvorhersage. Mit diesem Ansatz wurde ein Zink-Finger Motiv aus der völlig entfalteten Struktur mittels eines effizienten evolutionären Algorithmus gefaltet.

Darüber hinaus konnten wir die Faltungscharakteristika dieses Proteins durch die Analyse seiner Energielandschaft beschreiben. Wir entwickelten einen de-novo Ansatz zur Proteinstrukturvorhersage, mit dem wir die Struktur von 27 Proteinen im CASP7 Wettbewerb vorhersagen konnten. Insbesondere für Proteine ohne Homologie zu bekannten Strukturen war unser Verfahren vergleichsweise erfolgreich. Schliesslich untersuchten wir die Aggregation eines Fragments des Beta-Amyloid Proteins, von dem man annimmt, dass es für die Aggregation des vollständigen Proteins eine entscheidende Rolle spielt. Darüber hinaus konnten wir ein Verfahren für das Protein-Docking entwickeln und an zwei Protein-Dimeren testen.

Contents

List of Figures

List of Tables

1

Preface

Proteins are versatile molecules which perform multitude of functions in living organisms. Structural proteins, such as actin, tubilin, form the building blocks of cell and tissues. Enzyme proteins such as oxidoreductases, hydrolases, isomerases catalyze a variety of biochemical reactions. Transport proteins, such as hemoglobin, are responsible for carrying small molecules or ions across cells. Proteins, such as flagellin or myosin, are also responsible for the mobility of the organism. Kinases function as molecular switches that control many cellular processes. Proteins, such as histones, bind DNA and regulate the genetic information inside the cellular nucleus. Protein-protein interactions are responsible for many mechanisms of cellular control, including protein localization, inhibition and gene regulation. All above mentioned functions and mechanisms depend on the precise structure which proteins attain under physiological conditions. The process by which a protein assumes this precise three dimensional structure is known as protein folding. The understanding of protein structure and of the folding mechanism is therefore essential for understanding the mechanisms of life.

Chemically proteins are polymers of amino acids. The composition of amino acids for a protein is encoded in the corresponding gene as a triplet DNA code. The process of protein synthesis in-vivo begins in the nucleus of the cell. Transcription is the first process in the synthesis. A mRNA (messenger RNA) template carrying the protein sequence information is produced in this process. The mRNA is transferred outside the cellular nucleus and its sequence is translated into a polypeptide chain with the help of the tRNA (transfer RNA) in the ribosome. The protein then folds after the translation process. Larger proteins are escorted to the endoplasmic reticulum or mitochondria to initiate folding. Molecular chaperones such as GroEL/GroES help in the folding of large and complex proteins. The folded protein undergoes a quality control check. Misfolded or incorrectly folded proteins are usually degraded by the enzyme proteasome and the degraded components are recycled. Occasionally the degraded components associate and form stable structures which lead to formation of fibrillar aggregates. These aggregates are the cause for several pathological conditions.

The amino acid sequence of the protein, as determined by the genetic code, can be elucidated by sequencing techniques. Entire genomes of several fungi, the fruit fly, the mouse and the human have been sequenced. This corresponds to over a million of protein sequences. In contrast there are only about 40,000 protein structures available in the protein database, which collects all publicly available information. The gap between the number of known sequence and the known structure is enormous.

Large-scale experiments, such as the structural genomic initiative, are aiming to structurally characterize all important proteins. Experimental methods, such as X-ray diffraction and NMR, are presently used to characterize the three dimensional structure of proteins. X-ray diffraction provides high resolution static structures of proteins. NMR methods can be used for either resolving the native structure or studying the protein folding mechanism. Other experimental methods, such as circular dichroism (CD), are used for characterizing the folding mechanism. Though the experiments provide the accurate answers, they have several drawbacks. Some proteins, which does not crystallize, are not amenable to X-ray diffraction. NMR studies are presently confined to small/medium size proteins and are difficult to use for multi-domain proteins. Additionally NMR and X-ray diffraction experiments are expensive and time consuming. CD spectroscopy provides accurate information about the secondary structure, but is not suitable for elucidating the three dimensional structure.

The work of Anfinsen and his colleagues contributed to the understanding of the protein folding process. Their work on the spontaneous folding of denatured ribonuclease and associated investigations established a general idea of protein folding: **the thermodynamic hypothesis** . In words of Anfinsen *the 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.) is the one in which the Gibbs free energy of the whole system is lowest; i.e. that the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence in a given environment.*

The *sequence determines structure* hypothesis led to a serious theoretical problem. Due to existence of astronomical number of possible conformations for a protein, a random search of all conformations is impossible. So the protein was thought to pass through well-defined partially structured states to reach its native structure. Formulated by Levinthal and co-workers, the Levinthal's paradox stipulated the existence of unique folding pathways. However, experimental protein folding studies suggested a contradicting idea. While folding of some proteins involved a well-defined intermediate, others folded in a two-state mechanism (only unfolded and native states were observed).

The folding funnel paradigm resolves the Levinthal's paradox. This paradigm suggests a globally funneled energy landscape which is to some extent rugged, i.e. contains traps in which the protein can transiently reside. But the important feature of the landscape is that it is biased towards the native state. The protein can reach its native state via multiple routes. The distinct pathways are only followed after the protein has attained an almost native conformation. The funnel paradigm is supported by computational simulations of proteins, as well as the experiments such as NMR spectroscopy, laser initiated folding and ultrafast mixing experiments of small proteins.

Theoretical and computational methods have been complementing the experiments in the understanding of protein structure and folding mechanism. Simplified protein representations, such as lattice models, were able to qualitatively demonstrate the general principles of protein structure, stability and folding kinetics. However these simplified models are not suitable for structure prediction. Another important methodology used for studying the folding mechanism is the *Gō* model. In this model the roughness of free-energy surface is minimized by biasing the landscape towards to native structure. Since a protein is foldable in this model by definition, the folding characteristics are conveniently extracted in such scheme. The main disadvantage of *Gō* model is that it neglects the actual interactions which give rise to the biased landscape. Also, these models cannot be used for the structure prediction.

The commonly used computational approach for characterizing the folding dynamics is the molecular dynamics (MD) methodology. MD has been successful in elucidating the folding pathway for small proteins of up to 36 amino acids. But this success comes at a huge computational cost. Folding of such small proteins can take several months on a supercomputer. Because of such problems MD is not suitable for structure prediction.

Knowledge based methods, such as comparative modeling are widely used for structure prediction. These methods are limited to known classes of proteins for which they can yield structures approaching experimental resolution. De novo protein modeling does not suffer such limitation. These methods utilize Anfinsen's thermodynamic hypothesis and identify the native structure by searching for the global minimum of a free energy model. The key ingredients of such a approach are a) an accurate energy function b) an efficient sampling technique. These two components are realized in different ways. The coarse-grained representation of proteins enable the efficient sampling of energy landscape, but lack an accurate energy function. Methods, such as Rosetta, perform rapid sampling by using preformed protein fragments, but lack an atomistic energy function, which is able to distinguish between native and non-native structures.

Our approach is a free-energy model for protein folding and structure prediction which is based on Anfinsen's thermodynamic hypothesis. The development of a reliable atomistic energy functions, such as PFF02, has facilitated studies of different aspects of protein structure: folding, structure prediction and protein-protein interactions. In this work, we address the computational approaches for de novo protein folding and protein structure prediction. In addition we will also demonstrate the application of a general computational methodology for protein-protein interactions, in particular for protein aggregation studies.

This work is structured as follows:

The first chapter introduces the features of proteins: composition, structure, properties and the aspects of protein folding.

The second chapter summarizes the currently available computational methods for studying protein folding and protein structure prediction. The methods discussed in this chapter include the molecular dynamics method, comparative modeling, G$\bar{o}$ models, de novo protein structure prediction using united residue models, fragment based models and free-energy models.

The third chapter focuses on the protein folding. We use the free-energy forcefield PFF02, in conjunction with a stochastic algorithm to study folding of a protein with an $\alpha\beta\beta$ fold. The protein of interest is a 29 amino-acid DNA-binding zinc finger motif. Starting from a completely unfolded conformation, we fold it to a resolution of 4.5 Å. We also elucidate the folding characteristics by analyzing the folding landscape for this protein.

In the fourth chapter, we turn our attention towards a related problem of finding the native structure, but without recourse to folding pathways. We devise a de-novo methodology for predicting the native structure of proteins using a combination of a heuristic method and a free-energy refinement protocol. The methodology is applied for 27 targets in a blind protein structure prediction exercise. The predictions for some of targets were within experimental resolution. Our method was quite successful in the difficult template free-modeling section, where targets were modeled only by de novo methods.

The fifth chapter concerns the study of aggregation of proteins. Protein aggregation is recognized to be associated with pathological conditions. The Alzheimer's disease is one such condition, caused by aggregation of $A\beta$ protein. We study the aggregation of the 16-22 amino acid segment of Aβ which is one of key element in full protein aggregation. In addition we also study the aggregation of a segment of chain A of insulin. We also describe a test case for protein-protein docking. Two dimeric protein systems were used for this study. We find that the docked structures agree with the experimental structures.

2

Introduction

This chapter gives a brief introduction of proteins: chemical composition, structure, properties and the aspects of protein folding.

Proteins are most versatile molecules which perform multitude of functions, such as catalysing various biochemical reactions, controlling cellular processes and carrying small molecules or ions across cells. The biological function of a protein depends on its three-dimensional structure and its interaction with components of the cellular environment. Proteins attain a specific three-dimensional structure (also known as native structure) under physiological conditions. This three-dimensional structure is able to perform specific biological functions due to its specific structural and chemical properties.Several of such single chain proteins can assemble to form complex functional units. We need to understand the protein structure and its formation mechanism for understanding the mechanism of life.

2.1 Amino acids

Amino acids are the building blocks of proteins. There are twenty amino acids. An amino acid contains of a central carbon atom (C_α), connected to an amino group (NH_2), a carboxylic group (COOH), a hydrogen atom and a distinctive functional group(also known as side-chain, R). Amino acids are *chiral* molecules [1]. Consequently, there are two mirror-image forms: L-isomer and D-isomer (Fig. 2.1). It is found that only L amino acids are part of natural proteins. These molecules are shown in the Fig. 2.2. Amino acids in solution at neutral pH exist predominately as dipolar ions (also know as zwitterions), i.e. with a protonated amine group (NH_3^+) and deprotonated carboxylic-group (COO^-).

Amino acids are commonly abbreviated by either a three-letter or one-letter code. Glycine(G) is the simplest amino acid consisting of a hydrogen atom as the side-chain, while Tryptophan(W) is the

[1] Glycine is *achiral*, because the C_α atom is attached to two H atoms.

bulkiest amino acid containing an indole ring linked with methylene group as the side-chain. Based on the different properties of side-chains the amino acids can be classified into

- Hydrophobic: Alanine (A), Phenylalanine (F), Isoleucine (I), Leucine (L), Methionine (M), Proline (P) and Valine (V).
- Charged: Aspartic acid (D), Glutamic acid (E), Lysine (K) and Arginine (R).
- Polar: Cysteine (C), Histidine (H), Asparagine (N), Glutamate (Q), Serine (S), Threonine (T), Tryptophan (W) and Tyrosine (Y).

Glycine (G) is a special case which does not fit into any category in the above classification.

2.2 Polypeptide chains

The amino acids are joined by a peptide bond. A condensation reaction between two amino acids lead to the formation of a peptide bond with elimination of a water molecule (Fig. 2.3). A series of amino acids joined by the peptide bonds form a polypeptide chain. The first residue in the chain is referred to as amino-terminal (N-terminal) and the last residue carboxy-terminal (C-terminal). A polypeptide chain has a repeating part, called the main-chain (C_α-H,NH_2 and C=O) and a variable part, comprising the distinct side-chains.

The peptide bond has considerable double bond character, which prevents rotation about this bond. So the peptide bond is essentially planar. Thus, for a pair of amino acids linked together by a peptide bond, six atoms (C_α of adjacent peptides, C=O and NH groups) lie in the same plane. Two configurations are possible for the planar peptide bond: a *trans* conformation, where two C_α are on the opposite side of the peptide bond and a *cis* conformation, where two C_α are on the same side of peptide bond (Fig. 2.3). The *trans* form is more stable compared to the *cis* form. There is a large energy difference between the *trans* and the *cis* forms. However, in the case of proline, these forms have similar energies. Thus *cis* forms found in polypeptides are mostly observed between a proline and its preceeding residues.

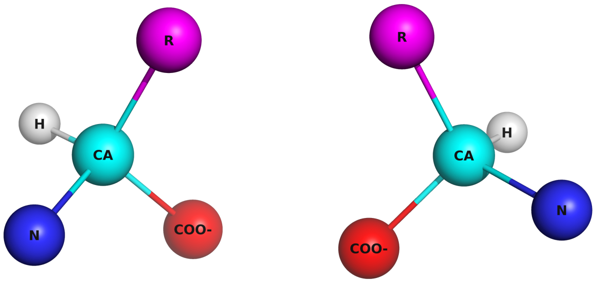
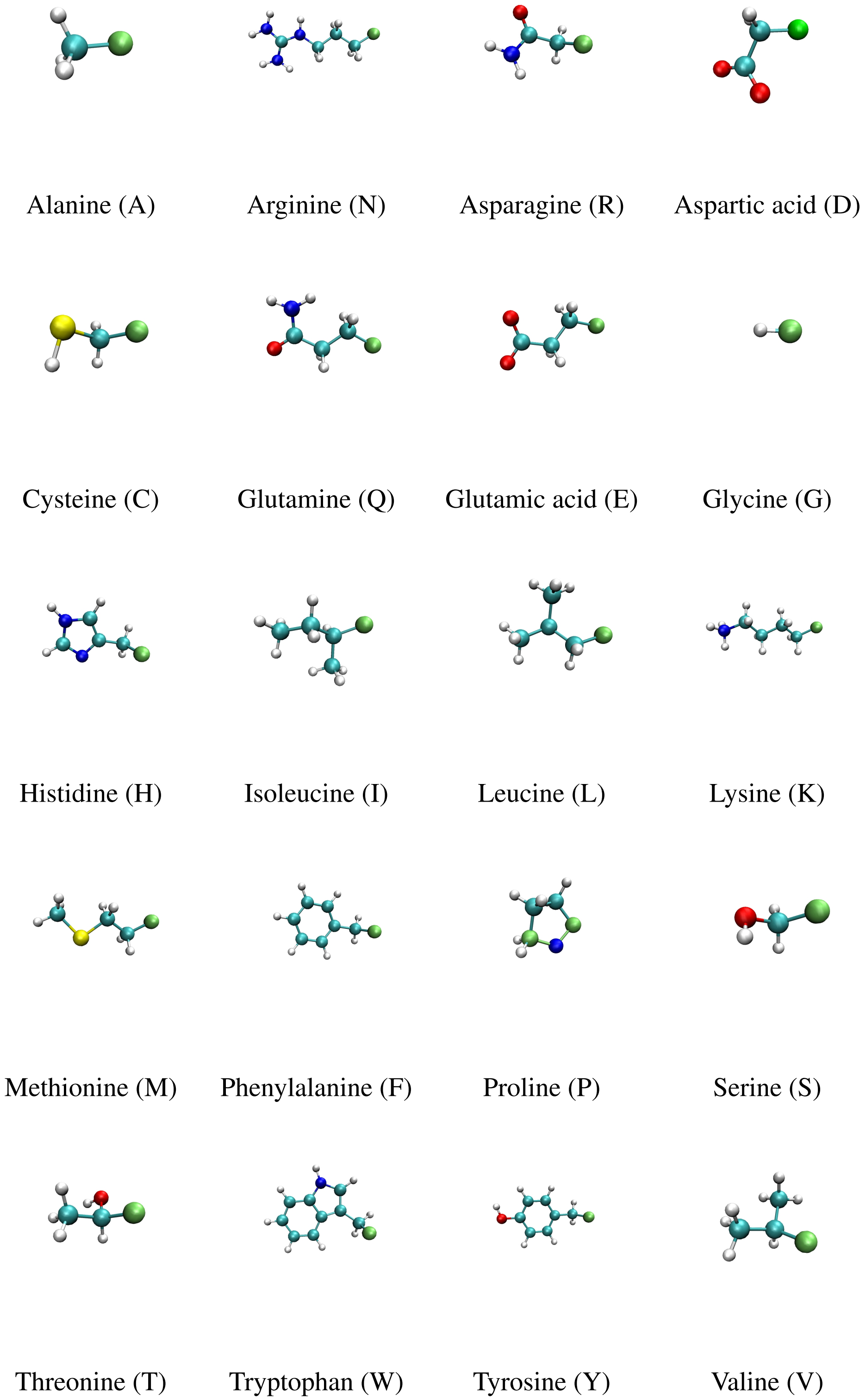


Figure 2.1: The D (right) and L (left) forms of amino acids. Only L forms are found in natural proteins.



Alanine (A)

Arginine (N)

Asparagine (R)

Aspartic acid (D)

Cysteine (C)

Glutamine (Q)

Glutamic acid (E)

Glycine (G)

Histidine (H)

Isoleucine (I)

Leucine (L)

Lysine (K)

Methionine (M)

Phenylalanine (F)

Proline (P)

Serine (S)

Threonine (T)

Tryptophan (W)

Tyrosine (Y)

Valine (V)

Figure 2.2: Side-chains of twenty naturally occurring amino acids. The main chain c_α atom is also shown(green). The carbon atoms are cyan, the oxygen red, the nitrogen blue, the sulfur yellow and hydrogen white.

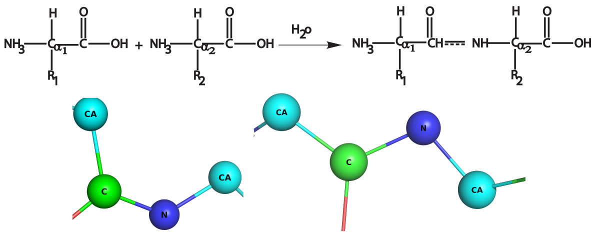


Figure 2.3: Top: Linking of two amino acids is accompanied by the loss of a water molecule
Bottom: The trans form of the peptide bond is strongly favored over the cis form to minimize steric clashes.

In contrast to the peptide bond, the bonds between the amino group and C$_\alpha$ atom and between the C$_\alpha$ atom and the carbonyl group are single bonds. The two adjacent rigid peptide planes can rotate about these bonds, taking on various orientations. This freedom of rotations about these single bonds allows proteins to exist in many different conformations. The rotation about the single bond is specified by a dihedral angle. The rotation about NH-C$_\alpha$ is called ϕ dihedral and about C$_\alpha$-C=O ψ dihedral. Not all the possible (ϕ, ψ) combinations are seen in the polypeptides. In fact, about three-quarters of the combinations are simply excluded due to local steric clashes (Berg et al., 2002). The allowed values can be visualized on a two-dimensional plot, called a Ramachandran plot (Ramachandran and Sasiskharan). There are two important allowed regions, one around ϕ = -57^o, ψ = -47^o and the other around ϕ = -120^o, ψ = 120^o. These regions are denoted by α_R and β in Fig. 2.4. The mirror image of α_R, called α_L is also allowed region, but observed rarely.

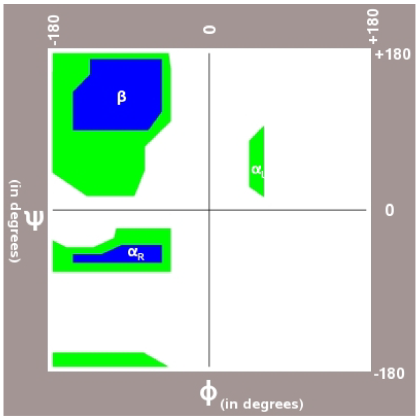


Figure 2.4: A Ramachandran plot showing the allowed combinations of ϕ and ψ dihedrals. The main allowed regions are concentrated around α_R (-57^o, ψ = -47^o) and β (ϕ = -120^o, ψ = 120^o).

2.3 Protein structure hierarchy

The rigidity of the peptide unit and the restricted set of allowed ϕ and ψ angles limits the number of favorable conformations of a protein. The structure of the protein is characterized by four hierarchical

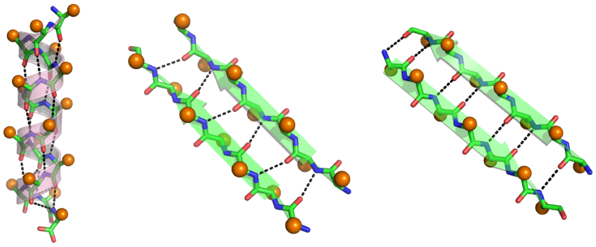


Figure 2.5: The α-Helix (left) , the parallel β sheet (middle) and the anti-parallel β sheet (right) are common secondary structure elements. The different hydrogen bonding pattern is illustrated for each of these structures. The side-chains are shown as orange spheres.

levels.

Primary structure

The primary structure is the sequence of amino acids numbered from the N-terminal to the C-terminal. It is also referred to as the linear structure since it contains no structural information about the protein. The primary structure is the link between the genetic information in DNA and the three-dimensional structure of a protein. The sequence information is expressed as either one-letter code or three-letter code.

Secondary structure

The polypeptide chains form distinct repeating units, called secondary structure. The commonly observed forms are the α-helix and the β-pleated sheet. Other categories of secondary structure of a protein are β-turns, ω loops, 3_{10} helices and polyproline helices. All these secondary structures are defined by the set of dihedrals and hydrogen bonding patterns. The secondary structure content of a protein can be obtained by circular dichroism experiments.

- α-helix: The α-helix is a rodlike structure where, the tightly coiled backbone form the inner part of the rod and the side-chains project outwards in a helical array (Fig. 2.5). The structure is stabilized by the hydrogen bond between the carbonyl group and the amine group of amino acid which is situated four residues ahead in the sequence. Except for the terminal residues in the helix, all the carbonyl and amine groups are hydrogen bonded. Each residue is separated by 1.5 Å along the helical axis and there are 3.6 amino acids per turn of helix. The α-helix can be right-handed or left-handed. The commonly observed α-helices in the proteins are right-handed.

  There are also other types of helices such as 3_{10}-helix, π-helix and polyproline II helix. All these helices are classified by their hydrogen bonding pattern. The polyproline π helix is left-handed, while other helices are the right-handed. The ideal parameters for the helices are given in Tab. 2.1.

Structure	ϕ	ψ	n	d (Å)	θ
α-helix	-57	-47	3.6	1.5	100^o
3_{10}-helix	-49	-26	3.0	2.0	120^o
π-helix	-57	-70	4.4	1.1	87^o
Polyproline II helix	-79	+149	3.0	3.1	-120^o

Table 2.1: Secondary structure is parameterized by the ϕ and ψ dihedrals, number of residues per turn (n), distance between successive residues along the helical axis and the rotation angle (θ).

- β sheet: The β-sheet structure differs from the rodlike α-helix. This secondary structure is made up of two (or more) extended chains, known as β strands. A β-strand is an extended structure where adjacent residues are 3.5 Å apart (compared to 1.5 Å in the α-helix). The side-chains of the adjacent residues point in opposite directions. A β-sheet formed by individual β strands can be either parallel or anti-parallel. In the parallel β-sheet the strands run in the same direction. Hydrogen bonds connect each amino acid on one strand with two different amino acids on the adjacent strand. In the anti-parallel orientation, hydrogen bonds each amino acid on one strand is paired to a single amino acid on adjacent strands. The hydrogen bonding pattern for parallel and anti-parallel β-sheets is shown in Fig. 2.5. The allowed regions in the Ramachandran plot for the β-sheets are concentrated around ϕ = -120^o and ψ = $+120^o$ respectively.

- Turns and Loops: The compact structure of globular proteins requires reversals in the direction of their polypeptide chains. These reversals are effected by a structural element called turn or loop. β-turn the hydrogen bond is between the amino acids π-turn between $i \rightarrow i \pm 5$ respectively. The longer reversal regions are referred to as ω-loops. Unlike the α-helix and β strands, turns and loops do not contain periodic structures.

Tertiary structure

The name tertiary structure refers to the three-dimensional structure of proteins. It is formed by an assembly of secondary structure elements such as α-helices, β sheets, loops and turns. The tertiary structure of globular water soluble proteins is characterized by a hydrophobic core and surface charged/polar residues. However in case of the membrane proteins, the core is formed by polar/charged residues.

The tertiary structure is also referred to as the native structure and is responsible for the biological function of the proteins. X-ray diffraction and NMR techniques are commonly used techniques for elucidating the tertiary structure of a protein. Many globular proteins can be well characterized by the currently available experimental techniques. However, very few structures have been obtained for membrane proteins (Out of 40,000 available three-dimensional structures, only 124 are membrane

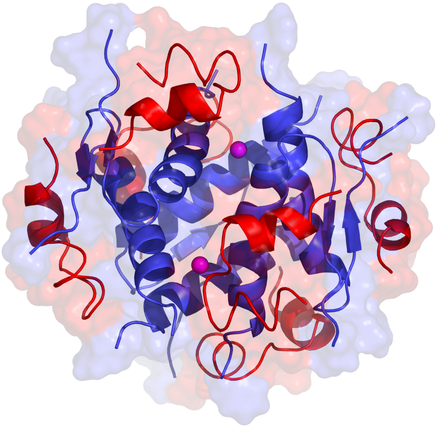


Figure 2.6: The quaternary structure of insulin complex contains six pairs of individual insulin dimers along with zinc ions

proteins).

Quaternary structure

Independently folding polypeptide chains (also known as domains) can assemble under physiological conditions to complexes that perform specific functions. Such spatial arrangement of domains, is known as a quaternary structure. The simplest possible quaternary structure is a dimer, i.e. two polypeptide chains. The quaternary structure of protein complexes are responsible for various biological activities. For example, the R6 human insulin which regulates the glucose content in the cell, contains six pairs of individual insulin dimers (Fig. 2.6).

2.4 In vivo protein folding

In-vivo protein folding starts with the transcription of genetic information contained in the DNA. The genetic code is composed of three nucleotides (also known as codons). Sixty out the sixty four possible triplets, code for the twenty naturally occurring amino acids, while remaining the four are signals for starting and ending protein synthesis. Methionine and tryptophan are the only amino acids encoded by just one triplet. In fact the start codon (AUG) is the codon for methionine.

The first process in protein synthesis is transcription. A mRNA template carrying the information of the protein sequence is produced in this process. The transcription occurs inside cellular nucleus in case of eukaryotes, while for prokaryotes this process takes place in the cytoplasm. The translation process involves the transfer of genetic information from the mRNA into a polypeptide, composed of amino acids. This process takes place with the help of transfer RNA (tRNA) in the ribosome.

Protein folding in certain cases is co-translational, i.e. it is initiated before the completion of protein synthesis when the nascent chain is still attached to the ribosome. Small proteins are formed in such folding processes. However larger proteins undergo a major part of their folding in the cytoplasm after the release from the ribosome, whereas yet others fold in specific compartments, such as mitochondria or the endoplasmic reticulum (ER), after trafficking and translocation through membranes. Folding of such proteins is often guided by molecular chaperones such as GroEL/GroES (Hartl and Hartl, 2002). Molecular chaperones do not themselves increase the rate of individual steps in protein folding, but contribute to efficiency of the folding process. These molecules (which are protein complexes) contain a cavity in which a incompletely folded polypeptide chain can enter and undergo some steps in formation to their native structure. There are intermediates during folding of such larger proteins (Roder and Colon, 1997). Fig. 2.7 shows the schematic representation of in vivo folding.

There is a "quality-check" control in the ER in which misfolded proteins are distinguished from native ones and degraded. Correctly folded proteins are released inside the cells or exported to the extracellular environment. However, failure of such system can have adverse effects. Proteins which are unable to fold correctly or to remain correctly folded, will give rise to malfunctioning of living systems and hence to disease. Some diseases, such as cyotic fibrosis result from incorrectly folded proteins (Thomas et al., 1995). In some cases, the misfolded proteins transform into insoluble aggregates within cells or in extracellular space. These aggregates are implicated in disorders, such as Alzheimer's disease, Creutzfeldt-Jakob disease and Type II diabetes (extracellular aggregates) and Parkinson's and Huntington's diseases (intracellular aggregates) (Dobson, 2001).

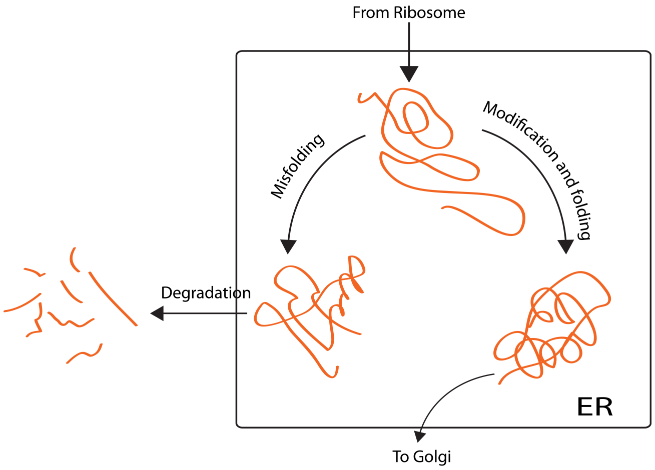


Figure 2.7: A schematic representation of in vivo folding. Many newly synthesized proteins are transported to ER, where they fold into a three-dimensional structure, assisted by molecular chaperones and folding catalysts. Incorrectly folded proteins are degraded by the proteasome system. However failure of such mechanism, lead to the formation of aggregates. This image is adapted from (Dobson, 2003)

2.5 Protein folding problem

Proteins are found to fold into a unique three-dimensional structure. Proteins can also fold outside the living environment, i.e. *in vitro* , under the suitable conditions. The folding times of proteins range

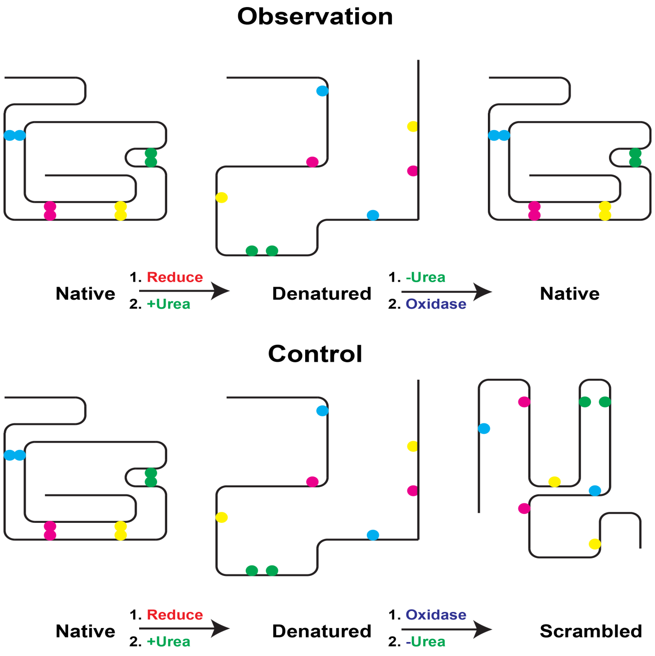


Figure 2.8: A Schematic diagram of Anfinsen's experiment is illustrated in two parts. Four pairs of the disulphide bridge forming cysteines are also shown.

from tens of microseconds to several seconds. The folding time for an average sized globular protein is in order of the order of several milliseconds.

The key questions concerning protein folding are:

How does a protein fold to a unique three-dimensional structure and how is the three-dimensional structure related to the amino acid sequence?

In the 1960's C. B. Anfinsen and his coworkers performed a series of seminal experiments in vitro that answered a key part of the protein folding problem. These works led Anfinsen to propose the "thermodynamic hypothesis". In words of Anfinsen: *This hypothesis states that the 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.) is the one in which the Gibbs free energy of the whole system in lowest; i.e. that the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence in a given environment*. (Anfinsen, 1973)

The crucial experiments influencing thermodynamic hypothesis is discussed briefly. These experiments were carried out on a 124 amino acid protein ribonuclease which contained four pair of disulphide bonds. The experiments are illustrated in two parts (Fig. 2.8).

- Observation: In the first phase, the disulphide bonds in the protein were reduced to eight -SH groups (using β-mercaptoethanol) and then treated with urea. Urea is known to effectively disrupt the non-covalent bonds, thus destroying the native structure of the protein. The denatured protein was not biologically active. In the second phase, the urea was removed from the system by dialysis and the -SH groups were oxidized back to the disulphide bonds. Now, the enzyme was able to regain about 90% of its initial activity. Thus the protein was found to spontaneously refold to its native structure under appropriate environmental conditions.

- Control: One of objections to the above result was that the protein was not completely unfolded by the addition of urea. To address this objection, the protein was reduced and denatured as above. But in the second phase the enzyme was oxidized first to form disulphide bonds and then urea was removed from the system. The activity of the enzyme was now about 1% of its original. The reason for such low activity was the incorrect disulphide bonds in the enzyme. The state of enzyme was described as ”scrambled”. There are 105 possible pairings for the eight cysteine molecules, but only one of them is found in the native structure.

  The scrambled protein was able to regain its activity, when treated with traces of β-mercaptoethanol. The added β-mercaptoethanol was able to catalyze the rearrangement of disulphide bonds until the native structure was regained in about ten hours. Thus,the native disulphide bonds contribute to the stabilization of thermodynamically preferred structure.

These experiments were basis for the key ideas of the protein folding problem. The native structure of the protein is solely determined by the sequence information: *sequence specifies structure*. In order for proteins to fold into the native state, the native state must be lower in free energy than the unfolded state. The native state must be the global minimum in the free energy that can be obtained without breaking of covalent bonds. The free energy difference (ΔG) between the folded and unfolded states is given by

$$\Delta G = \Delta H - T\Delta S \tag{2.1}$$

where ΔH, ΔS are enthalpic and entropic changes respectively. The enthalphic changes is associated with the atomic interactions, while the significant entropic changes comes from the configurational entropy and hydrophobic interactions. Because of the large competing enthalpic and entropic contributions, the free-energy difference is just few kcal/mol (Dill, 1990).

2.6 Dominant factors in protein folding

Proteins are marginally stable at room temperature. The native structure of protein is attained only within narrow ranges of conditions of solvent, pH and temperature. It is not enough for native conformation to be stable, but the protein must be able to find it in a short time, starting from a denatured state. The interactions of side chains and main chain, with one another, and with the solvent and with surrounding proteins or ligands, determine the energy of the conformation. In most cases, entire conformation (or at least significant part) is necessary for stability, because the protein is stabilized by both local and non local interactions. The dominant interactions which influence the structure of proteins are

- Chemical bonds: Proteins contain covalently bonded atoms in main-chain and side-chains. Some proteins contain a covalent bonds between atoms of different sidechain. One such interaction are disulphide bridges which commonly observed between the cysteine residues. Disulphide bridges increase the stability of the protein by topologically constraining different parts of polypeptide chain. Some proteins are stabilized by presence of ligands or metal ions. For example, zinc finger proteins contain zinc ion which is coordinated by side chains, while the hemoproteins are bonded to heme group.

- Hydrogen bonds: Hydrogen bonding plays a vital role in protein folding. When a protein is unfolded, the polar atoms of proteins form a hydrogen bond with the water molecules. In the folded state, the buried polar atoms lack these interactions. To compensate the energy loss, buried atoms form hydrogen bond with other protein atoms. The secondary structure elements of protein α-helices and β-sheets are stabilized by intermolecular hydrogen bonds.
- Hydrophobic effect: Certain amino acids have apolar sidechains which interact weakly with water molecules. These hydrophobic residues bury themselves in the interior of proteins, while the charged/polar residues come to the surface.
- Van der Waals interactions: Proteins are collapsed to form a compact structure. The compactness of the structure is influenced by both the attractive interactions which bring atoms together and also by the repulsive steric clashes between atoms.
- Configurational entropy: In an unfolded protein, the polypeptide can assume different conformations around (ϕ,ψ) dihedrals and side chains can adopt different roomers around χ angles. However, in the folded protein, (ϕ,ψ) are restricted to a single value, whereas χ dihedrals have very limited freedom. This loss in accessible conformational space translates into to loss of configurational entropy. Unlike the above mentioned interactions, configurational entropy opposes protein folding.

2.7 Levinthal's paradox

Anfinsen's thermodynamic hypothesis raised a fundamental question regarding the folding mechanism of proteins. Levinthal framed the question regarding the two major goals: finding the global minimum (known as thermodynamic control)and finding it quickly (kinetic control) (Levinthal, 1968; Dill, 1990). The protein conformation can be characterized by two dihedral angles (ϕ, ψ). The number of possible conformations is astronomically large (estimated $2^{100} \backsim 10^{30}$ conformations for an 100 amino acid protein). Even if the protein is able to sample 10^{11} conformations/second, it will take 10^{11} year to sample the entire conformational space. Thus a random search of conformational space cannot take place, for a protein which folds in millisecond time scale. Therefore folding was thought to be not under thermodynamic control. Levintal (Levinthal, 1968) suggested the existence of specific pathways for folding. These pathways are characterized by well-defined partially structured states, through which protein has to pass to attain its native state. This idea significantly reduced the conformation space and thus avoided astronomical folding times. But since the process was pathway dependent i.e. (final structure depend on initial conditions), proteins should reach only local minima. Thus the kinetic control dominated the protein folding.

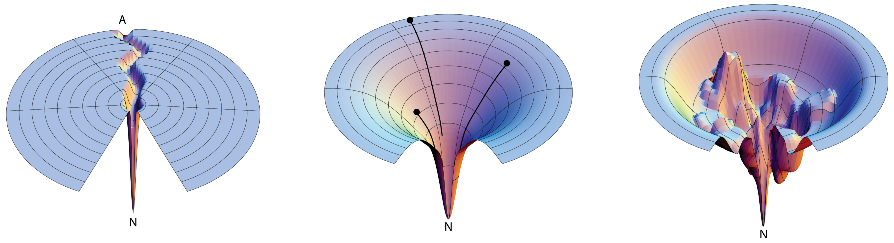


Figure 2.9: Left: Folding funnel corresponding to kinetic control of folding mechanism. Middle: An idealized folding funnel without kinetic traps.Right: A rugged energy landscape with kinetic traps, energy barriers and some narrow throughway paths to native. All these images are taken from (Dill and Chan, 1997)

2.8 Folding funnel

The idea of kinetic control of folding dynamics influenced the classical view, which describes the folding process through a sequence of discrete intermediate states. However, experimental results suggested a contradicting idea. While folding of some proteins involved a few well-defined intermediate, others folded in a two-state mechanism. A new view of protein folding replaced the idea of "folding pathways" with the broader notions of energy landscapes and folding funnels (Dill and Chan, 1997; Onuchic et al.; Brooks et al., 2001). Protein folding is not seen as a sequential event, but rather as a progressive organization of an ensemble of partially folded structures, through which the protein passes its way to the native state. Such mechanism are possible due the the globally funneled energy landscape.

The funnel paradigm resolves Levinthal's paradox. A kinetic view of folding funnel (Fig. 2.9) concurs to Levinthal's view by presuming a tunnel towards the global minimum of energy (thus eliminating the random search problem). But the physical basis for such specific sequences of events are unclear (Dill and Chan, 1997). A thermodynamic view of folding funnel scenario agrees with both Anfinsen's thermodynamic hypothesis as well as Levinthal's paradox. An idealized folding funnel without kinetic traps is shown in Fig. 2.9. When folding conditions are initiated, the protein can find the global minimum by many different routes, but yet do so in a directed and rapid way. The idea of sequential assembly of specific structures, is replaced by parallel processes and ensembles. The multi-state kinetics of folding is also realized by funneling. In such a scheme, protein has to pass through an ensemble of structures (intermediate states) to fold into a native state. In this case, the energy landscape is rugged due to energy barriers and kinetic traps (Fig. 2.9).

3

Computational methods for Protein Structure Prediction and Protein Folding

The understanding of protein structure and function is key for elucidating the biological processes. Computational methods should complement the experimental investigations in such endeavor. In the current chapter we present an overview of current state-of-art computational methods for protein folding and protein structure prediction.

There are two distinct theoretical and computational paths for studying protein structure. One class of method considers how does a protein attain a unique native structure by tracing its folding path (a time trajectory) from unfolded to the folded ensemble. The other class does not consider the how protein folds, but predict its native structure directly. Both of these ways have their own significance in the understanding of protein structure. Fig. 3.1 shows the distinction between protein structure prediction(PSP) and protein folding. We discuss some aspects of both PSP and protein folding. The simulation methods, such as molecular dynamics and *Gō* models are used for studying protein folding dynamics. Methods such as comparative modeling, ab-initio PSP methods, such as free energy models and Rosetta, are used for structure prediction. Ab-initio approaches such as free energy models or united residue models can be used for both structure prediction and folding dynamics.

3.1 Molecular dynamics

Molecular dynamics(MD) was first introduced for studying system of hard spheres. However they have been extended to several systems including Lennard-Jones systems, liquid water, molten salts and glasses, biomolecules such as proteins, lipids, nucleic acids etc (Allen and Tildesley, 1987).MD methods are intuitively appealing and involve mostly classical regime physics[1].

The MD simulations essentially solve the Newton's equations of motion for a system of interacting particles. The equations of motion are given by

[1] However the parameters for the potential are derived using both classical and quantum principles.

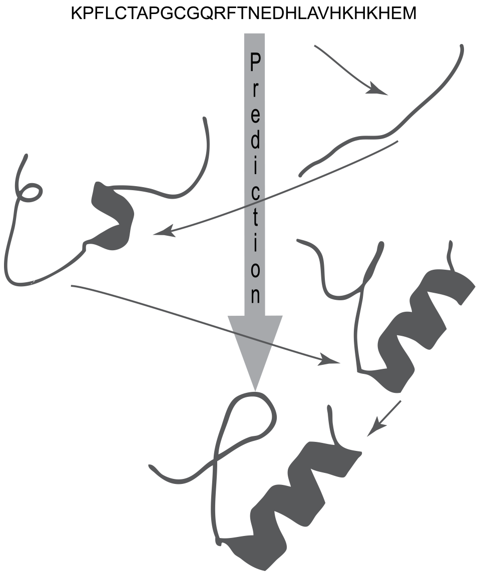


Figure 3.1: Protein Structure Prediction vs Protein Folding. The prediction does not give the information about the folding dynamics.

$$m_i \frac{\mathrm{d}^2 \overrightarrow{r}_i}{\mathrm{d}t^2} = \overrightarrow{F}_i \tag{3.1}$$

where m_i is mass of particle i, $\overrightarrow{r}_i$ is position vector of i^{th} particle,$\overrightarrow{F}_i$ is the force acting on i by surrounding particles. Force exerted by surrounding j particles on ith particle is given by

$$\overrightarrow{F}_i = -\sum_j \nabla \phi(r_{ij}) \tag{3.2}$$

where $\phi(r_{ij})$ is the corresponding pair wise potential function between particle i and j.

A MD trajectory involves three essential stages

- Initialization: The initial coordinates for the particle can be random or set on a lattice. For biomolecular simulations the initial co-ordinates are often taken from experimental structures. The initial velocity vector is adjusted such that kinetic energy of the system corresponds to expected value at target temperature of system. The individual particle velocities are drawn from a Gaussian distribution with zero mean and variance, $k_B T_o / m_i$ where k_B is the Boltzmann constant, m_i mass of the ith particle and T_o is the system temperature.

- Equilibration: Starting with assigned co-ordinates and velocities, the system is evolved for a certain *equilibration* time. During the equilibration, energy exchange between kinetic and potential components. The equations of motion are solved until the total energy converges.

- Production: The properties of the system are measured after the equilibration phase. The typical properties accessible via MD simulations are thermodynamic quantities such as energy, pressure, entropy, specific heat etc. structural and transport properties such as radial distribution functions, diffusion co-efficient, dynamic structure factors etc..

The application of MD for molecular system has been facilitated by the Born-Oppenheimer approximation to the Schrödinger equation. According to this approximation, the motions of the molecules in a system can be studied using a hierarchy of contributions. First, only the motion of electrons are considered and the nuclei is assumed to be fixed. This approximation helps to calculate the electronic energy levels as function of atomic coordinates. Second, the electronic ground state is used as the potential energy for the system.

In the MD simulations, the atoms of the biomolecular system are represented as a collection of point masses (centered at the nuclei) which are connected by springs. The molecule stretches, bends or rotates in response to the inter and intramolecular forces. Each atom has a partial charge, which reflects its molecular environment. A Lennard-Jones radius defines the spatial extent of each atom. The interactions of atoms is defined by a potential energy function, which is commonly referred to as forcefield. The forcefields can be purely ab-initio (from first principles), empirical or knowledge based (derived from a distribution). Many empirical forcefields are commonly used in MD.

3.1.1 Empirical forcefields

The potential energy function is the core component of a MD methodology. The empirical forcefield based MD studies are most commonly used for biomolecules including proteins, nucleic acids and lipids. The success of the empirical force field lies in the fact that they are able to reproduce experimentally accessible information, their simplistic functional form and the efficient algorithms which have enabled the microsecond/millisecond MD for small systems(about 5000 atoms) and nano second simulations for very large systems(about 100,000 atoms).

$$
\begin{aligned}
U(\overrightarrow{r}) &= \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
&+ \sum_{dihedrals} K_\chi[1+cos(n\chi-\delta)] \sum_{impropers} K_\zeta(\zeta-\zeta_0)^2 \\
&+ \sum_{nonbonded}\left(\varepsilon_{ij}\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon_D r_{ij}}\right)
\end{aligned} \tag{3.3}
$$

Equation 3.3 shows a potential energy function which is the most common form for protein forcefields (Mackerell, 2004). It is composed of simple functions which represent a minimal set of forces to describe a biomolecular structure:

Where b is the bond length, θ is the valence angle, χ is the dihedral angle ζ is the improper angle and r_{ij} is the distance between atoms i and j. The parameters which constitute the actual forcefield are bond force constant and equilibrium distance, K_b and b_0 respectively; the valence angle force constant and equilibrium angle, K_θ and θ_0 respectively; the dihedral force constant, multiplicity and phase angle, K_χ, n and phase angle δ respectively; K_ζ and ζ_0 represent the improper force constant and equilibrium improper angle respectively. All the above terms constitute the internal/intra molecular parameters. The non-bonded interactions are characterized by Lennard-Jones well-depth ε_{ij}, minimum interaction radius $R_{min_{ij}}$ for van der Waals interactions; the dielectric constant ε and partial charges of atom i and j, q_i and q_j for the coulombic interactions. The dielectric constant ε_D is usually set to 1, which corresponds to the permittivity of vacuum, in the calculations that incorporate explicit solvent representations. The alternative methods which treat the solvent environment implicitly have distance dependent dielectric constants. The terms contributing to the potential energy in Equation 3.3 are common to majority of popular empirical forcefields such as CHARMM (MacKerell et al., 1998), AMBER (Pearlman et al., 1995), GROMOS (Scott et al., 1999), and OPLS (Jorgensen et al., 1996).Some extended forcefield include higher order terms to treat the bond and valence angle terms, cross terms between bonds and valence angles or valence angles and dihedrals. They facilitate more accurate treatment of vibrational/rotational spectra. Other alternatives terms in the extended forcefield include Morse function for bonds, co-sine based angle terms and grid-based dihedral energy correction maps.

Among the non-bonded interactions, the van der Waal's interaction is sufficiently represented by Lennard-Jones 6-12 potential. On the other hand most current electrostatic implementation do not treat explicitly electronic polarizability. The polarizability is implicitly taken into account by choosing proper partial charges which overestimate the molecular dipoles. This overestimation mimics the condensed phase environment which occur in biomolecules. There are cases where an explicit inclusion of polarizability is being tested. The most common ways to include the electronic polarizability are induced dipole models, fluctuating charge models or their combinations (Halgren and Damm, 2001).

The solvation treatment is another difficult aspect of an empirical forcefield. Both explicit and implicit solvent treatments are being used. The popular water models used in the biomolecular simulations include the TIP3P, TIP4P and SPC models (Jorgensen et al., 1983). Most of the water models yield proper characteristics for bulk water at room temperature. The water models differ in the way the solvent molecule interaction sites are represented. Another important aspect is the fact that each of above water model is linked with a particular empirical forcefield, since forcefields are developed in conjunction with a specific water model. Thus AMBER, OPLS and CHARMM go with TIP3P, OPLS mostly uses TIP4P and GROMOS uses SPC. The implicit solvation (Orozco and Luque, 2000) models offer several advantages over the explicit treatment. The solvation treatment is approximately same as that of explicit models, while the computational costs are considerably lower. So these models are popular with simulations involving extensive conformational sampling. The different methods of implicit solvation treatment involve: 1)using a distant dependent dielectric constant; 2)Poisson-Boltzmann(PB) models; and 3) Generalized-Born(GB) models. The more accurate treatments involve coupling PB/GB methods with solvent accessibility, which treats the hydrophobic effect.

3.1.2 MD algorithms

Integration

Several numerical integrator are used to solved the equation of motion. The simplest one among them is Verlet integrator which is accurate to $\bigcirc(t^4)$

$$\overrightarrow{r}(t+\delta t) = 2\overrightarrow{r}(t) - \overrightarrow{r}(t-\delta t) + \delta t^2 \overrightarrow{a}(t) \tag{3.4}$$

A slightly modified version of Verlet algorithm, known as Verlet-leapfrog integrator is generally used to minimize the numerical errors associated with the Verlet algorithm (Allen and Tildesley, 1987).

$$\begin{aligned}\overrightarrow{r}(t+\delta t) &= \overrightarrow{r}(t)+\overrightarrow{v}(t+\frac{1}{2}\delta t)\delta t\\ \overrightarrow{v}(t+\frac{1}{2}\delta t) &= \overrightarrow{v}(t-\frac{1}{2}\delta t)+\overrightarrow{a}(t)\delta t\\ \overrightarrow{v}(t) &= \frac{1}{2}\left(\overrightarrow{v}(t+\frac{\delta t}{2})+\overrightarrow{v}(t-\frac{\delta t}{2})\right)\end{aligned} \tag{3.5}$$

The time step δt depends on properties which are to measured in the system. For example in the folding/unfolding studies a δt of 2fs is used, whereas δt is set to 1fs for calculating the spectral properties. When a large δt is used, the high frequency bond stretching is prevented by constrained dynamics. The algorithms SHAKE and RATTLE are commonly used for the rigid bonds (Adcock and McCammon, 2007).

Thermostats

The temperature of the system computed via the virial theorem fluctuates as the system is evolved. Therefore a temperature control is generally applied to the system. One of simplest method for this purpose is the temperature re-scaling (Allen and Tildesley, 1987). If T_A is instantaneous temperature of the system and T_o is the required temperature, the re-scaling factor $f_i(t)$ for velocities of particles is defined as

$$f_i(t) = \sqrt{\frac{3k_BT_o/m_i}{\sum_j v_i(t)^2/N_i}} \tag{3.6}$$

where $\sqrt{3k_BT_o/m_i}$ is the mean velocity of particle i with mass m_i, T_o is the desired temperature, N_i is number of particle of type i and $\sqrt{\sum_j v_i(t)^2/N_i}$ is the computed mean velocity of particle i for temperature T_A.

The temperature and pressure of the system can be monitored conveniently by using external weak coupling thermostats/barostats. These thermostats/barostats use a fictitious frictional coefficient γ which controls the relaxation rate of the coupling. The commonly used thermostats for constant temperature or pressure MD include Berendsen, Nose-Hoover, Anderson thermostats (Frenkel and Smit, 2001). An alternative method uses the elements of Langevin's dynamics for the thermostat.

Long range interactions

The evaluation of non-bonded interactions is of order $\bigcirc(N^2)$, where N is the number of atoms. To reduce the computational cost, these interactions are evaluated using three alternative schemes.

- Cutoff: The cutoff scheme involves a pre-defined bounding distance after which the long range interactions are no longer evaluated. Switching functions are used to alter original function to smoothly fade to zero at cutoff distance. This scheme is used for the van der Waal's interactions.

- Ewald summation: This is applicable only to systems which use periodical boundary conditions. The $\frac{1}{r}$ term in the electrostatics potential is decomposed to into a short range interaction in real space and long range interaction in reciprocal space (Frenkel and Smit, 2001). The electrostatics of the system is evaluated completely with such schemes. The variant of Ewald's summation, particle-mesh Ewald(PME)is used often in MD (Sagui and Daren, 1999). The computation of overall interactions is reduced to $\bigcirc(N \log N)$ with best implementation of PME.

- Multipole expansions: These schemes rely on a power-series expansion for the electrostatic interaction. The multipoles are expressed in terms of spherical co-ordinates (r, θ, α), since the expansion involve spherical harmonics $Y_{lm}(\theta, \alpha)$ (Frenkel and Smit, 2001). The multipole expansion scheme can be applied to both periodic and non-periodic systems. In MD, this scheme is realized by the fast multipole method(FMM) (Sagui and Daren, 1999). The computation of overall interactions is reduced to $\bigcirc(N)$ with the best implementation of FMM.

Enhanced sampling

The sampling of the phase-space by conventional MD is usually limited. In addition the system can get trapped in local minima. Several schemes are used to overcome these problems. One convenient scheme is replica exchange molecular dynamics(REMD) (Sugita and Okamoto, 1999). In this scheme identical non-interacting copies of the system are simulated at different temperatures. A swap of copies at different temperatures is attempted, such that all replica remain in thermal equilibrium.

$$\begin{aligned} W(m \to n) &= \begin{cases} 1 \\ exp(-\Delta) \end{cases} \\ \Delta &= [\beta_m - \beta_n](U_m - U_n) \\ \beta_m &= \frac{1}{K_B T_m} \end{aligned} \tag{3.7}$$

where $W(m \to n)$ is probability of swaps between temperature m and n respectively, U_m, U_n are potential energies. During the swap the high-temperature replicas escape from local traps or jump from one energy basin to another, whereas the low-temperature replicas explore a single region in energy landscape like conventional MD. In other methods such as Local Enhanced Sampling(LES) (Czerminski and Elber, 1991), a small fragment of the system (a sidechain or a ligand molecule) is duplicated to N non-interacting copies. The remainder of system is made to interact at reduced strength ($\frac{1}{N}$). This facilitates the significant increase of sampling for these fragments.

3.1.3 Applications

A significant advantage with the MD method is the fact that it tries to capture properties of real proteins which are dynamical structures interacting continously with their environment. These observables include the molecular geometries and energies, the mean atomic fluctuations, conformational changes which are easily accessible by the MD. Since MD deals with the time evolution of the system, kinetic aspects are accessible through the trajectories. It is no surprise that MD is one of most widely used techniques for studying topics related to protein folding and dynamics.

The initial MD simulation of a complete protein involved a simulation of small bovine pancreatic trypsin inhibitor(BPTI, about 800 atoms) in vacuum for 9.2 picoseconds (McCammon et al., 1977). The recent MD simulations are several times larger in magnitude (both in time and system size). The first micro second simulation was performed on 36-residue villin headpiece in explicit water which was able to fold the protein to a native-like structure (Duan and Kollman, 1998). Recently the same system was studied for 500 micro second by distributed computing project, FoldingHome (Jayachandran et al., 2007). Other widely studied proteins on microsecond scale are trp-cage protein(20 residues) (Snow et al., 2002), C-terminal β hairpin of protein G (16 amino acids) (Zagrovic et al., 2001) and designed protein BBA5(23 residues) (Rhee et al., 2004).

Another application of MD is the study of transition states for two-state protein folding. The transition state is highest energy point in reaction pathway from which a protein can fold or unfold with equal probability. Proteins such as 63 residue chymotrypsin inhibitor 2 and 110 residue bacterial protein barnase were studied in conjuction with experiment (Fersht and Daggett, 2002). Other MD applications involve simulation of membrane channels (Roux and Schulten, 2004), an enzyme reaction (Neria and Karplus, 1997), protein aggregation studies (Wei et al., 2007; Klimov and Thirumalai, 2003), protein-ligand docking and protein design (Adcock and McCammon, 2007) and structure refinement.

3.2 *Gō* model

The *Gō* model (Go, 1983) is one of earlier models developed for the study of protein folding. The model employs a minimalistic protein representation (like the c_α trace). The *Gō* model minimize the roughness of the free energy surface by biasing the energy landscape to fold into given three dimensional structure. This native interaction biased free energy landscape avoids the problem of actual physical interactions responsible for the biased surface. The *Gō* interactions are generally sequence dependent. Advanced *Gō* models have included sequence independent terms and explicit treatment of solvent interactions (Cheung et al., 2002). Since a protein is foldable by definition in the *Gō* model, the folding characteristics can be extracted more comfortably compared to MD simulations. They reproduce qualitatively differences between folding kinetics of small and large proteins and explain folding rate and mechanism of folding (Head-Gordan and Brown, 2003).

3.3 Comparative Modeling

Comparative modeling is a knowledge-based method which predicts the tertiary structure of a given protein sequence (target) based on its alignment to one or more proteins of known structure (templates). The underlining philosophy of comparative modeling is based on the observation that evolutionarily related family of proteins have similar conformation. The similarity is detectable at the sequence level (ie amino acid sequence) and at 3D structure level. Even proteins that have no de-

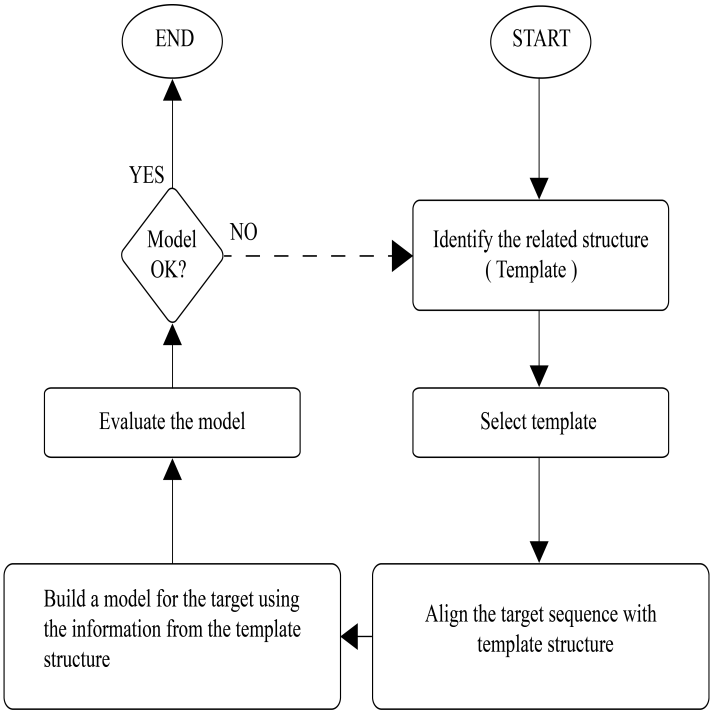

END
START
YES
Model OK?
NO
Identify the related structure (Template)
Select template
Evaluate the model
Build a model for the target using the information from the template structure
Align the target sequence with template structure

Figure 3.2: A flowchart illustrating the comparative modeling. This chart is adapted from (Marti-Renom et al., 2000)

tectable sequence similarity which perform related function can have similar structure. There are essentially four important sequential steps in comparative modeling (Fig. 3.2).

- Fold assignment and Template selection: The first step in comparative modeling is to identify all protein structures related to the target sequence and to then choose those which can be used as templates. The are several ways to identify a template. The simplest of these methods involve a pairwise sequence-sequence comparison of target with a database sequence. The programs such as FASTA (Pearson, 1998) and BLAST (Altschul et al., 1990) are used for this purpose. The second class of methods rely on multiple sequence comparison to improve the sensitivity of the search. The most widely used program is PSI-BLAST (Altschul et al., 1997) which iteratively expands the set of homologs of the target sequence. The third class of methods, so-called threading approach, depend on comparison of protein sequence and protein of known structure (Torda, 1997). Threading is useful when there are no sequences which are related to the target.

- Template-target alignment: In this important step of the modeling protocol, the target sequence is aligned to the template structure to produce the optimal alignment. For closely related proteins (with over 40% sequence identity) the alignment is usually correct, but the alignment becomes difficult in the *twilight zone* where the sequence identity is less than 30% (Rost, 1999). As the sequence similarity decreases, the alignment produces large number of gaps which have to modeled in concurrence with the template topology. Generally the alignment is done with multiple sequences. A frequently used program for multiple alignment is CLUSTAL (Jeanmougin et al., 1998).

- Model building: There are three popular ways to construct 3D models for the target from the alignment. The rigid-body assembly method builds a model by assembling a small number of rigid bodies obtained from alignment. This approach is based on the dissection of the protein folds into conserved core regions, loops and side chains. The second way to construct a model is by segment matching or coordinate reconstruction. The model is constructed using a subset of atomic positions (usually c_α) from the template as guiding positions to build up the 3D structure. The third class of methods build the model using constraints or restraints on the structure of target sequence based on the alignment. The model is derived by minimizing the violations of the restraints/constraints on distance geometries and dihedral angles.

  In a given protein fold family, structural variations arise from substitutions, deletions or insertions between the members of the family. These variations usually occur in the loop regions. The loop regions are important, since they are usually the active and binding sites. So considerable efforts are made to model the loop regions accurately. Loop modeling is difficult using homolog information since short residue fragments (up to 7) does not fold identically (Mezei, 1998). There are two widely used methods for modeling loop: ab-initio loop prediction based on conformational search in a given environment or a database approach to find a segment of main chain that fits the two stem regions of a loop. (Marti-Renom et al., 2000)

  A related problem in loop modeling is side-chain building. The side-chain angles are coupled to the backbone $\phi\,\psi$ dihedral angles. So the sidechains from the homologs are usually copied to the model. Methods which rebuild the sidechain from rotamer library sometimes fail for highly homologous sequences. Another effect which is usually neglected in the rotamer approach is the inclusion of the solvation term.

  The models can be further refined using physics based energy functions. Here protocols involving a molecular dynamics refinement or free-energy refinement are used.

- Model Evaluation: The quality of the predicted model determines the amount of information that can be extracted. The first step in a model evaluation is to check whether the predicted model and template share same fold(model-template alignment). The second step is to identify the conserved functional residues/segments in the model. Several statistical potentials of mean force and physics based energy functions are used to discriminate native from non-native models. If the model built is not satisfactory, the steps involving template selection/alignment and model building are repeated iteratively until a good model is obtained.

Comparative modeling is the first choice for modeling homologous proteins. Proteins with $\geq$50% sequence identity are modeled to 1-2 Å resolution. A recent study suggested that it would be possible to characterize structures of almost 90% proteins(including membrane proteins), if 16,000 carefully chosen homologous targets are resolved by experiments (Sali, 1998). Such a choice of targets is crucial for success of structural genomics initiative which aims for structural characterization of proteins (Baker and Sali, 2001). Other applications of comparative modeling include identifying, designing improving the binding sites of protein, protein-protein docking, refinement of poorly resolved X-ray/NMR models, designing mutants for modifying protein functions (Marti-Renom et al., 2000).

3.4 Ab-initio protein modeling and folding

Ab-initio methods predict the protein structure from physical principles without using any structural information from the protein database. This is not strictly accurate, since the potentials used in these methods are derived using the information from the protein database. But these methods differ from comparative modeling which uses the database explicitly (*i.e.* template for prediction). The advantage of these methods is that they are not limited to protein families of known structure like comparative modeling. Ab-initio methods are based on the assumption that the native state of protein is well defined, which is the global minimum of a certain free-energy function. The required ingredients of this approach are : an accurate free-energy function and an efficient optimization method for exploring the energy landscape of the protein. Ab-initio methods are also used for studying protein folding, though this field is currently dominated by the MD methods. Firstly we discuss coarse grained models and move to fragment based modeling. Finally we describe an all-atom free-energy model for PSP and protein folding.

3.4.1 Coarse grained models

Protein structure modeling can be performed at various levels of structural detail ranging from simplified two-dimensional or three-dimensional lattice models, continuous representations, united residue models or with all atom models. The reduced representations offer several advantages over all-atom models. The number of degrees of freedom in the system is reduced enabling more extensive sampling of the energy landscape. Folding studies of larger proteins, which are not accessible to all-atom MD simulations can be realized with these models. The important drawbacks of coarse grained models are the lack of accurate reduced representations and the difficulty to design adequate potentials to represent real proteins.

UNRES: United Residue Model

The UNRES model (Liwo et al., 1997a,b, 1998) is one of the forcefields for coarse grain protein modeling. It represents a polypeptide chain as sequence of c_α atoms linked by virtual bonds of length 3.8 Å. The interaction sites representing the peptide groups(PG) are located in middle of c_α virtual bonds. The side chains(SC) are represented as a single interaction site attached to the c_α and the c_α...SC bond lengths are fixed. The UNRES forcefield has been derived by averaging the system consisting of protein plus solvent with implementation of Kubo's theory of cluster cumulants. The energy function is parameterized to achieve a hierarchical structure of protein energy landscape.The forcefield contains terms which account for the interactions between PG-PG, SC-PG, SC-SC, local terms that account for rotation about c_α..c_α virtual bond axis, the bending of virtual valence angles and different rotameric states of SC and multibody correlation terms.

The global minimum is searched by using a hybrid optimization method, where a genetic algorithm generates the populations by a set of crossover operations, followed by a local minimization (Oldziej et al., 2005). The method is successful for small proteins (up to 80 residues). For larger proteins a complicated crossover operators involving non local pattern exchanges(eg nonlocal β sheets), dynamic formation/breaking of disulphide bonds is employed. The conformational space is searched by either starting with a random conformation or with conformation generated using secondary structure predictions.

The UNRES force field has been used for blind protein structure predictions (Oldziej et al., 2005). Other applications include protein folding studies using the molecular dynamics and Monte Carlo methods (Nanias et al., 2006).

3.4.2 Fragment based models

Conformation sampling is one of the most severe problems associated with ab-initio modeling. The size of conformational space increases drastically with the length of the protein chain. Efforts are made to accelerate the search by limiting the conformational space to include only the subset of structures compatible with local sequence structural propensities. These methods are based on experimental observation that local sequence preferences bias, but do not uniquely define the local structure of a protein. The native conformation of the protein is obtained by a interplay of local interactions with each another, to form a compact conformation with favorable non-local interactions, such as hydrophobic burial, disulphide bridge formation etc (Bonneau and Baker, 2001). We describe in brief one of most successful methodologies which uses fragment based modeling approach.

Rosetta

The Rosetta strategy is based on a picture of protein folding in which short fragments of the protein chain, consistent with local structure-sequence relation assemble to a native state where the interactions of such local segments are optimized to minimize the total free energy of the protein (Baker and Sali, 2001). Fig. 3.3 illustrates the conceptual basis of Rosetta methodology.

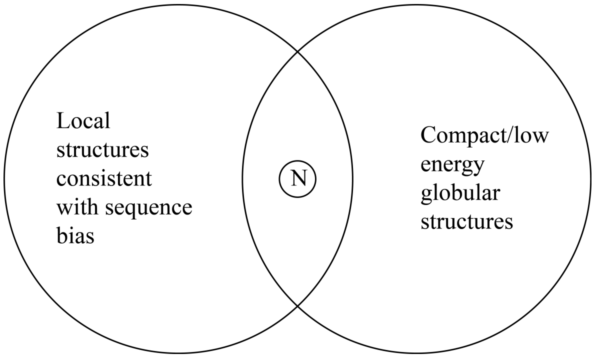


Figure 3.3: A Venn diagram illustrating the concept for the Rosetta methodology. The near-native structures are labeled N. This image is adapted from (Bonneau and Baker, 2001)

The Rosetta de-novo prediction protocol is explained briefly below. A detailed review can be seen in (Rohl et al., 2004b).

- Protein representation: Rosetta uses a torsion space representation for the protein. The bond angles and bond lengths are set to the ideal values. The side chain has two alternate representations; a reduced centroid representation or the full atom representations, depending on the potential terms and simulation phases. The sidechains are restricted only to a set of discrete conformations as described by the backbone structure (Dunbrack and Cohen, 1997).

- Scoring function: Rosetta scoring function is based on a Bayesian separation of the total energy into components that describe the likelihood of a particular structure, independent of sequence and those that describe the fitness of the sequence given a particular structure. (Simons et al., 1997).

  $$P(Structure \mid Sequence) = P(Structure) \times \frac{P(Sequence \mid Structure)}{P(Sequence)} \tag{3.8}$$

  where P(Structure|Sequence) is the probability that a structure exist for a given sequence , P(Structure) is the probability that a structure exist, P(Sequence) is the probability that given sequence exist and P(Sequence | Structure) is the probability that a sequence exists with a compatible structure.

  The term P(sequence) is constant, since different structures are compared for a same given sequence. The estimation of other two probabilities is explained briefly. A more detailed explanation is given in (Simons et al., 1997, 1999).

  - P(Sequence|Structure): The sequence dependent term $P(Sequence \mid Structure)$ in the Equation. 3.8 can be expressed as

    $$P(Sequence \mid Structure) = P(a_1, a_2, ...a_n | \overrightarrow{X}) \tag{3.9}$$

    where a sequence is written as a string of amino acids (a_i) of length n and the structure is described by a vector, $\overrightarrow{X} = \{x_1, x_2, ...x_n\}$. The occurrence of a particular amino acid at a given position is independent of the three-dimensional structure of protein and depends only on the local environment. The other factor involved is the pairing probability. The strongest interactions occur between the polar residues and the cysteine residues. Combining these two factors $P(Sequence \mid Structure)$ is written as

    $$P(Sequence \mid Structure) = P_{env} P_{pair} \tag{3.10}$$

    The P_{env} interactions can be thought to model the hydrophobic effect, whereas P_{pair} models the electrostatic and disulphide bonding.

  - P(Structure): is the sequence independent probability density which is designed to distinguish *protein-like* structures from random polymeric structures. There are several important factors in the protein-like structures. Compactness is modeled in two ways: the radius of gyration measure of all C_α atoms or a density measure defined by ratio of number of C_β atoms with 10 Å in real proteins to C_β atoms within 10 Å in random polypeptides. The most important metric which distinguishes proteins from random polymers is the existence of secondary structure elements and their packing in the native structure. The secondary structure packing is defined by a multiple probability density functions which describe: helix-helix packing, helix-strand packing, strand-strand packing, strand hydrogen bonding and strand assembly in sheets. The independent variable for these functions is the secondary structure vector. The other interactions include a van der Waal's interaction to avoid the atom-atom overlaps and sequence independent part of the packing orientation($P_{packing-struct}$).

- Fragment selection: The basic operation involved in the protein structure construction is the insertion of the fragments. For each fragment insertion, a consecutive window of three or nine residues is selected, and the torsional angles(ϕ, ψ) of these residues are replaced with the torsion angles obtained from a fragment of known protein structure. These fragments are chosen from customized fragment library built using a nonredundant database which composed of X-ray structures of $\leq$ 2.5 Å resolution and $\leq$ 50% sequence identity. The sequence profiles for the query sequence is initially built by using two rounds of PSIBLAST (Altschul et al., 1997) and a profile-profile similarity score is calculated. In addition, secondary structure predictions are performed on the given sequence using Psipred (Jones, 1999), SAM-T99 (Karplus et al., 2001) and JUFO (Meiler et al., 2001). A secondary structure similarity score is also calculated. The overall similarity score is defined as sum of sequence similarity score and half the secondary similarity score. A ranked list of the top fragments for each sequence window is assembled iteratively, adding the top scoring fragment according to each secondary structure prediction to the combined ranked list and eliminating the redundancies. The fragment selected according to the Psipred secondary structure predictions is included with threefold greater frequency than other predictions. A 200 nine-residue and 200 three-residue fragment library is constructed for every overlapping insertion window.

- Fragment assembly: The fragments are assembled into compact *protein-like* structures by a Monte-Carlo search. Initially a 9-residue window is randomly selected and a fragment corresponding to this window is inserted from the 25 top ranked fragments list. The energy of resulting conformation is evaluated. A conformation is accepted or rejected based on the Metropolis criterion. If no moves are accepted for several successive insertions, the acceptance probability is increased temporarily. After a successful move, the acceptance probability is reset to its original value. A total of 28,000 nine-residue insertions are attempted. Full scoring function is not used over the entire simulation. Initially only the steric terms are evaluated. During the course of simulation, other terms are progressively added to the total potential. After the assembly of the decoy structures with 9-residue fragments, each decoy is subjected to a short refinement of 8000 attempted 3-residue insertions using the complete scoring function. Several short simulations are carried out with different random seeds to generate an ensemble of decoys. This set is clustered by structural similarity to identify the native like structures. The fragment insertion protocol is improved for model refinement, loop modeling or protein design using operations such as random angle perturbations, the basin hopping technique (Wales and Doye, 1997), rapid optimization of side-chain rotamers etc.

The fragment assembly approach has several benefits: the use of preset library of low-energy local structures avoids calculation of actual local interactions; the need for calculating accurate positions of atomic co-ordinates is traded-off for the rapid search of large conformational space and finally the single-fragment insertion method significantly minimize the number of accessible conformers.

The Rosetta ab-initio protocol has been applied successfully for protein structure prediction. It has consistently performed well over several CASP exercises (CASP5, 2003; CASP6, 2005; Bonneau et al., 2001). The Rosetta methodology is also used in protein design (Kuhlman et al., 2003), prediction of protein-protein interactions (Gray et al., 2003), structure determination from coarse experimental data (Rohl and Baker, 2002) and loop modeling (Rohl et al., 2004a).

3.4.3 Free-energy model methods

The free-energy model for protein structure prediction and folding is based on the Anfinsen's (Anfinsen, 1973) thermodynamic hypothesis which postulates that native state of protein is the global minimum of the free-energy surface. This model is physics based and rely on the design of suitable energy function. In the free-energy approach the native structure of a protein sequence is found by identifying the global minimum of the energy function.

PFF02

PFF02 is an all-atom (with exception of apolar CH_N groups) free-energy force field which identifies the native state of protein as its global minimum. The force field models the physical interactions of a protein in an implicit solvent(water) environment at a fixed temperature of 300K. The bond angles and bond lengths are set to standard values. Rotation about the peptide bond is forbidden. The degrees of freedom are the dihedral angles of backbone(ϕ , ψ) and sidechain dihedrals($\chi_1, \chi_2, ...$).

The force field PFF02 consists of six non-bonded interactions, including two interactions which were added to original forcefield PFF01 (Herges and Wenzel, 2004)

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

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

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 as a potential of mean force from experimental data by fitting short-range (2 Å-5 Å) radial distributions of a set of 138 different proteins which are believed span wide range of different folds(Avbelj and Moult, 1995).

- Electrostatics: The electrostatic interaction is modeled using the standard columbic potential. The contribution is split into main chain and side chain contributions.

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

  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.

- Hydrogen bonding: Hydrogen bonds play a vital role in the protein folding (Berg et al., 2002). The experimental estimate of the hydrogen bonding interaction ranges between -2.8 kcal/mol to +1.9 kcal/mol (Avbelj, 1992; McDonald and Thornton, 1994), which is much smaller than covalent bond interactions. Generally hydrogen bonding is not explicitly modeled, but its contributions are embedded partly in the electrostatics and Lennard-Jones. Since hydrogen bonding and solvent interaction are the two major contributions to protein folding, these interactions are specially emphasized and modeled by two contributions in PFF01/02.

  - Electrostatic interactions considering only the dipole-dipole interaction of the amino- and carboxyl-groups of the mainchain. The long range interactions are overemphasized due to the cooperative effects.

$$V_{hbdipole} = \frac{0.1064e^2}{4\pi\varepsilon\varepsilon_0}\left(\frac{1}{r_{C_iH_j}} - \frac{1}{r_{C_iN_j}} - \frac{1}{r_{O_iH_j}} + \frac{1}{r_{O_iN_j}}\right)$$

    (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).

  - An additional short-ranged term which corrects the hydrogen bonding by considering the alignment of the hydrogen bond with respect to the donor and acceptor groups (Sippl et al., 1984).

$$V_{hbcorr} = V_0 \sum_{ij} R(r_{H_iO_j})\Lambda(\alpha_{ij}, \beta_{ij})$$

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

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

The hydrogen bonding term is interpolates these contributions.

$$V_{hb} = \lambda V_{hbdipole} + (1-\lambda)V_{hbcorr}$$

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

- Solvation: The solvent energy and entropy influences the folding of a protein and contributes to the free-energy of the system. On the surface of a protein there are important solvent interactions: hydrophobicity, ie the entropy of water molecules, the conformational entropy of the protein sidechains[2] and the modulation of the solvation of charged side groups.

  The solvation effects are modeled in PFF02 via an implicit solvent model based on the Solvent Accessible Surface Area (SASA) of the protein (Lee and Richards, 1971). The SASA is calculated by rolling a water sphere of radius 1.4 Å over the protein surface which is defined by the Lennard-Jones radii.The solvation term is given by the relation

  $$V_{sol} = \sum_i \sigma_{PT(i)}A(i)$$

  where $PT(i)$ is the potential type of atom i, $\sigma_{PT(i)}$ describes the Atomic Solvent Parameter (ASP) according to the potential type and $A(i)$ is the SASA of the atom i. The parameters are derived by fitting first a SASA model (Eisenberg and McLachlan, 1986) to reproduce the enthalpies of solvation of tripeptide Gly-X-Gly (Sharp et al., 1991) and then adjusting these parameters to stabilize the native structure of villin headpiece (Herges and Wenzel, 2004).

- Local electrostatics correction: Aminoacids have a preference for secondary structure elements. For example tryptophan, threonine occur mostly in β sheet regions, whereas alanine prefers α-helical region. These preferences are influenced by different electrostatic interactions of the main chain dipoles in their local environment (Avbelj and Moult, 1995). The interaction E_{local} is 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

[2] The main chain is somewhat rigid and its entropic contribution is not significant

adjoining peptide groups. So NH_i interacts with CO_{i-2}, NH_{i-1}, CO_i & NH_{i+1} and CO_{i-1}, NH_i, CO_{i+1} & NH_{i+2} interact with CO_i.

$$V_{\text{local}} = \lambda_{\text{local}} \frac{332.15 \times \zeta}{2} \sum_j \sum_i \frac{q_i q_j}{r_{ij}}$$

where q_i is the charge on the atom and r_{ij} is distance between the atoms. The parameter ζ is amino acid specific parameter.

- Torsional energy: A weak dihedral angle dependent energy term is introduced to stabilize the residues in the beta sheet regions of Ramachandran plot. The interaction has a favorable energy contribution of 0.8 kcal/mol(maximum) for the residues forming beta sheet.

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

for all amino acids except proline and glycine. For proline and glycine $E_{\text{tor}} = 0$. ϕ_i and ψ_i are the backbone dihedral angles of amino acid i. We used $\phi_i = -110°$, $\psi_i = 130°$, $\gamma_\phi = 5 \times 10^{-3}$ deg^{-2} and $\gamma_\psi = 1.25 \times 10^{-3} \text{deg}^{-2}$.

Optimization methods

Efficient optimization methods (Schug et al., 2005a; Verma et al., 2006) are used to locate the global optimum in the PFF02 energy function for a particular sequence. These include

- Stochastic tunneling : Here a potential energy surface is transformed by using a non-linear transformation to suppress the barriers which are significantly above the present best energy estimate (Hamacher and Wenzel). The transformed energy surface which is used for exploration of global minimum is given by

$$E_{STUN} = ln(x + \sqrt{x^2 + 1}) \tag{3.11}$$

with $x = \gamma(E - E_0)$, where E is the present energy, E_0 is best estimation so far and γ the transformation parameter, which controls the rate of rise for the transformation.

- Parallel tempering: This method is Monte Carlo counterpart of the replica exchange molecular dynamics method described in the Sec. 3.1.2. A modified version of this method which uses an adaptive temperature control and replication step, is employed for exploration of the energy surface (Schug et al., 2005b).

- Basin hopping technique: In this scheme the original potential energy surface is simplified by replacing the energy of each conformation with the energy of a nearby local minimum (Wales and Doye, 1997). The minimization is carried out on the simplified potential with simulated annealing.

- Evolutionary strategy: This scheme is an extention of the above described BHT. Several concurrent simulations are carried out in parallel on a population. The population is evolved towards a global optimum of energy with set of rules which enforce energy improvement and population diversity. A detail view of this algorithm will be presented in the next chapter.

Protein folding with PFF02

Starting from an extended conformation, several helical (Schug et al., 2004, 2003; Schug and Wenzel), beta (Wenzel, 2006; Verma, 2007) and proteins of mixed folds (Gopal and Wenzel, 2006; Verma, 2007) were folded to near-native structures. The following table lists some of the proteins folded using PFF01/02.

PDBID	#AA	Topology	bRMSD (Å)
1L2Y	20	α	3.11
1WQE	23	α	2.33
1F4I	40	α	3.29
1ENH	54	α	3.40
1GYZ	60	α	4.30
2A3D	80	α	4.15
1K1V	42	α	4.40
1VII	35	α	3.56
1BDD	46	α	2.30
1LE0	12	β	1.50
1NIZ	14	β	2.04
GSGS	32	β	2.19
1RIK	29	α and β	4.15
1BHI	29	α and β	4.53

Table 3.1: The list of proteins folded in PFF01/02.
#AA is no. of amino acids, bRMSD is for the best prediction.

4

Folding Studies

As discussed in the previous chapter, there is a distinction between protein folding and protein structure prediction. In this chapter we focus on the folding problem. We use a stochastic algorithm in conjuction with the free energy forcefield PFF02 to study folding of a protein with an $\alpha\beta\beta$ fold.

We have introduced the free energy forcefield in the Sec. 3.4.3. PFF02 identifies the native structure of a protein sequence as the global minimum of the free-energy landscape. The task of finding the global minimum is complicated because the energy landscape of a protein is complex and rugged. Often there are several low-energy minima on this surface which differ only slightly in energy(few kcal/mol), but are far away from each other in the configurational space. Since analytical methods cannot be used to solve this optimization challenge, we focus our attention on problem-specific stochastical methods. Our group has adapted and developed several optimization methods, including stochastic tunneling (Hamacher and Wenzel), parallel tempering (Schug et al., 2004), basin hopping technique(BHT) (Verma et al., 2006) and evolutionary strategy(ES) (Schug and Wenzel).

4.1 Evolutionary Strategy

The popular BHT method (Nayeem et al., 1991; Wales and Doye, 1997) for global optimization eliminates high-energy potential-energy surface(PES) by replacing the energy of each conformation with the energy of a nearby local minimum. For protein folding we have replaced the original local minimization by simulated annealing(SA). In the course of our folding studies, we find that independent BHT simulations often find identical structures corresponding to same local(global) minimum. As a result, each independent simulation reconstructs the full folding path independently. It would be very desirable to develop methods, where several concurrent simulations exchange information to *learn* from each other. For a PES having many local minima, independent simulations limit the efficient exploration of the PES. Also, occasionally BHT simulations go astray, ending the search in a wrong energy basin of the PES. We have developed a *greedy* version of BHT (Wenzel, 2006) which overcome these problems to a certain extent.

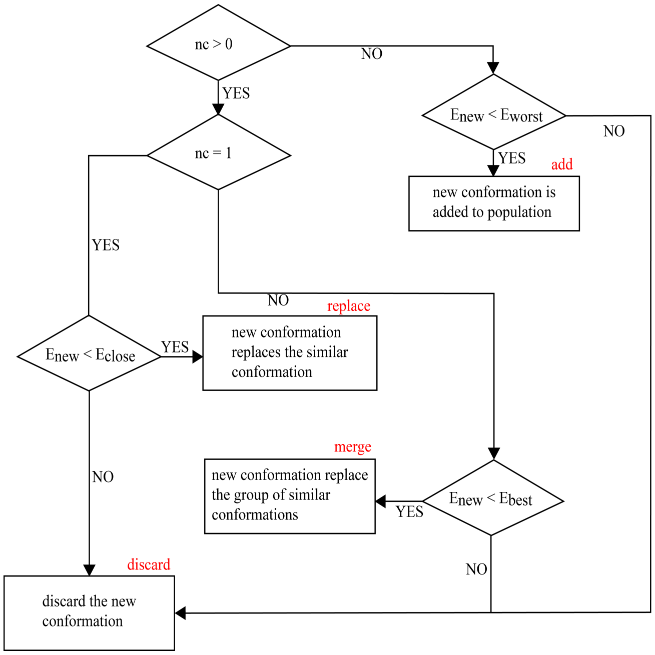

nc > 0
NO
YES
Enew < Eworst
NO
YES
add
new conformation is added to population
nc = 1
YES
NO
replace
Enew < Eclose
YES
new conformation replaces the similar conformation
merge
NO
new conformation replace the group of similar conformations
YES
Enew < Ebest
discard
NO
discard the new conformation

Figure 4.1: A flowchart illustrating the population update. See the text for an explanation

We have therefore generalized the BHT approach to a population of size N which is iteratively improved by P concurrent dynamical processes (Schug et al., 2005b). The population is evolved towards a optimum of the free energy surface with a ES that balances the energy improvement with population diversity. In the ES, conformations are drawn from the *active* population and subjected to an annealing cycle. At the end of each cycle the resulting conformation is either integrated into the active population or discarded. The algorithm was implemented as a master-client model in which idle clients request a task from the master. The master maintains the *active* conformation of the population and distributes the work to the clients. Each step in the algorithm has three phases:

1. Selection: A conformation is drawn randomly from the *active* population. We have used a uniform probability distribution with population of 20 conformers.

2. Annealing cycle: We use a simulated annealing schedule with T_{start} drawn from an exponential distribution and T_{end} fixed at 2K. The number of steps per cycle is increased as $10^5 \times \sqrt{cycle}$.

3. Population update: We have adjusted the acceptance criterion for newly generated conformations to balance the population diversity and energy enrichment. We define the two structures as *similar* if they have bRMSD less than 3 Å to each other. We define an *active* population as the pool containing mutually different lowest energy conformers. The master finds number of similar structures(*nc*) and then performs one of the following operations on complete population.

   (a) Add: If the new conformation is not *similar* to any structure(nc=0) in the population, we add to the population, provided its energy is less than the energy of conformation with highest energy(E_{worst})

(b) Replace: If the new conformation (with energy E_{new}) is *similar to one* existing structure in the population (with energy E_{old}), it replaces that structure provided $E_{\mathrm{new}} < E_{\mathrm{old}} + \Delta$ (see below).

(c) Merge: If the new conformation has *several similar* structures, it replaces this group of structures provided its energy is less than the best one of the group E_{best} plus an acceptance threshold Δ.

A flowchart illustrating the population update tasks of the master is shown in Fig. 4.1. In our first BHT/ES simulations we have used a fixed energy threshold (Δ) acceptance criterion. Here we have implemented a *variable* energy threshold which we define as $\Delta = A \times tanhD$, where

$$D = \frac{E_{\mathrm{new}} - E_{\mathrm{best}}}{A},$$

where A is the energy threshold (3kcal/mol), Enew is energy of the new structure, E_{best} is the lowest energy structure in the population. This choice of the energy criterion ensures that the conformation with the best energy is never replaced, while conformations higher in energy are more easily replaced in the secure knowledge that they are far from optimal. The rules for the *replace* and *merge* operations ensure the structural diversity of the population and its continued energetic improvement (on average).

We demonstrate the ES to fold a protein of mixed secondary structure.

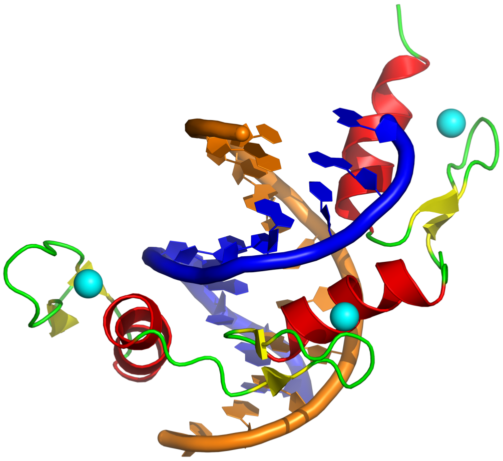


Figure 4.2: A classical Cys_2His_2 zinc finger motif with Zn-ion(orange) and DNA (magenta).

4.2 Folding of DNA-Binding Zinc Finger motif

Zinc fingers are among the most abundant proteins in eukaryotic genomes and occur in many DNA binding domains and transcription factors (Laity et al., 2001). They function in DNA recognition, RNA packaging, transcriptional activation protein folding and assembly and apoptosis. Many zinc fingers contain a Cys_2His_2 binding motif that coordinates the Zn-ion in $\alpha\beta\beta$ -framework (Lee et al., 1989; Pavletich and Pabo, 1991; Wolfe et al., 2000) and much effort is towards the engineering of novel zinc fingers (Urnov et al., 2005). A classical zinc finger motif binding DNA is illustrated in Fig. 4.2.

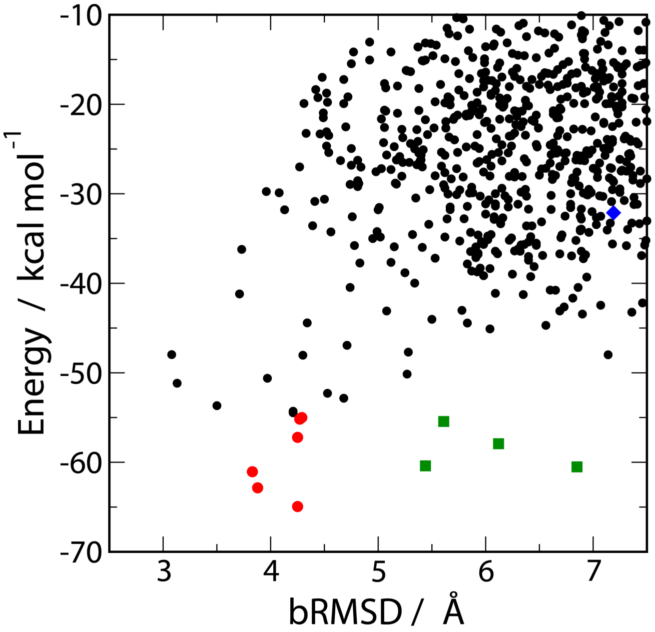


Figure 4.3: Free energy versus bRMSD of all accepted conformations in the simulation. The best 10 structures are highlighted as: red circles(native-like), green squares(non-native). The folding intermediate is denoted by blue diamond

The reproducible folding of such proteins with mixed secondary structure, however, remains a significant challenge to the accuracy of the all-atom forcefield and the simulation method (Abagyan and Totrov, 1999). We use the all-atom free-energy forcefield PFF02 to predictively fold the 23-51 amino-acid segment of the N-terminal sub-domain of ATF-2 (PDBID 1BHI) (Nagadoi et al., 1999), a 29 amino acid peptide that contains the basic leucine zipper motif . 1BHI folds into the classical TFIIIa conformation found in many zinc-finger like sub-domains. The fragment contains all the conserved hydrophobic residues (PHE25, PHE36, LEU42) of the classical zinc finger motif and the CYS27, CYS32, HIS45, HIS49 zinc binding pattern. The sequence is illustrated below.

LYS-PRO-**PHE25**-LEU-**CYS27**-THR-ALA-PRO-GLY-**CYS32**-GLY-GLN-ARG-
PHE36-THR-ASN-GLU-ASP-HIS-**LEU42**-ALA-VAL-**HIS45**-LYS-HIS-LYS-**HIS49**- GLU-MET51

Starting from a completely unfolded conformation with no secondary structure (16 Å backbone

RMSD (bRMSD) to native) we performed 200 cycles of the evolutionary algorithm. The distribution of bRMSD versus energy of all accepted conformations during the simulation (Fig. 4.3) demonstrates that the simulation explores a wide variety of conformations, with regard to their free-energy and their deviation from the native conformation.

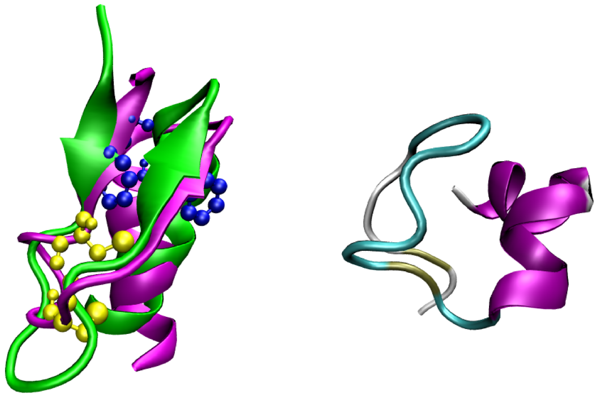


Figure 4.4: Left: Overlay of the native(green) and folded(magenta) conformations. The conserved hydrophobic residues are shown in blue and Zn binding cysteines are shown in yellow. Right: The intermediate conformation with partially formed helix and β sheet.

#	Energy kcal/mol	bRMSD Å	Secondary Structure
E01	-64.94	4.25	CC**EE**CTTTTSCC**EE**SSC**HHHHHHHHHHHH**C
E02	-62.84	3.88	CC**EE**CTTTTSCC**EE**SSC**HHHHHHHHH**STTC
E03	-61.05	3.83	CC**EE**CTTTTCCC**EE**SSC**HHHHHHHHH**STTC
E04	-60.51	6.85	CC**EE**CTTTTSCC**EE**CSC**HHHHHHH**SCCCCC
E05	-60.40	5.44	CCBBCTTTTCCCBCCSC**HHHHHHHH**CCCBC
E06	-57.93	6.12	CC**EE**CTTTTSCC**EE**CSC**HHHHHHH**SCCCCC
E07	-56.21	4.25	CC**EEEE**CSSSSC**EEEE**SC**HHHHHHHHHHH**C
E08	-55.44	5.61	CCSSSCSSCCSSCCCSC**HHHHHHHHH**TTTC
E09	-55.18	4.27	CCCC**EE**CTTSSC**EE**CS**HHHHHHHHHH**CSCC
E10	-55.02	-4.29	CCCCBTTTTBTTCCCSS**HHHHHHHHHHHH**C

Table 4.1: Energy, bRMSD and secondary structures of best 10 lowest energy structures

Among the ten energetically lowest conformations (see Tab. 4.1) six fold into near-native conformations with bRMSDs of 3.68-4.28 Å, while four fold to conformations with a larger bRMSD. The three energetically best conformations are all near-native in character, an overlay with the experimental conformation (left panel of Fig. 4.4) illustrates that the helix, beta-sheet and both turns are correctly formed. The hydrophobic residues, which determine the packing of the beta-sheet against the helix, are illustrated in blue in the figure.

The helical section (GLU39-GLU50) and the beta-sheet (PHE25-LEU26 and ARG35-PHE36) deviate individually by 1.6 Å and 2.4 Å bRMSD from their experimental counterparts, respectively. The overall deviation between the experimental and the folded conformations stems from the relative arrangement of the beta-sheet with respect to the helix, which are dominated by unspecific hydrophobic interactions. All conserved hydrophobic sidechains are also buried in the folded structure. The zinc-coordinating cysteine residues (CYS27,CYS32) are within 2 Å of their native positions and available association with the Zn-ion.

The partially folded low-energy conformations among the ten energetically best (green symbols in Fig. 4.3) also have significant native content. All of these conformations exhibit a fully formed helical section in the right region of the amino acid sequence, a turn between helix and beta-sheet regions. The right panel of Fig. 4.4 shows the conformation furthest from native with a partially unzipped beta-sheet. The hydrogen bonds near the turn region are still present, while the native hbonds at the end of the zipper have not yet formed.

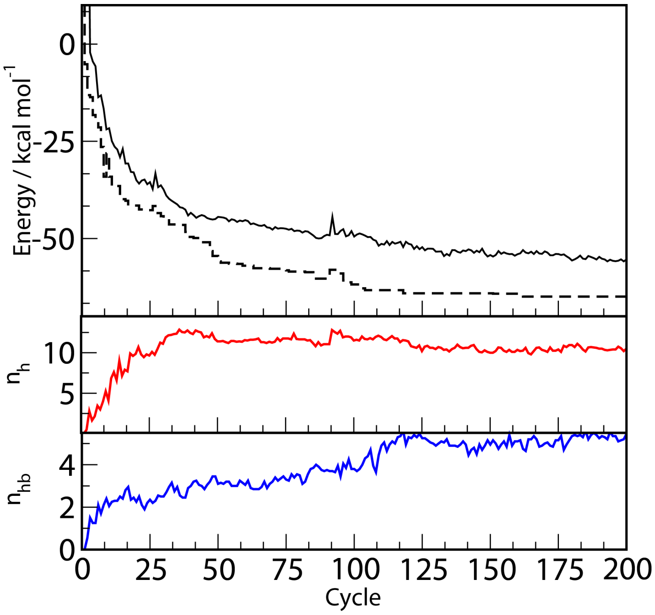


Figure 4.5: Top: Average (solid line) and best (dashed line) energies, Middle: number of amino acids (n_h) in a helical conformation (as computed by DSSP) and Bottom: number of hydrogen bones (n_{hb}) as function of the ES cycle number

Fig. 4.5 (bottom panel) shows the convergence of the energy. After about 120 attempted updates per population member (3.5×10^8 function evaluations) the population converged to the native ensemble. According to the funnel paradigm for protein folding (Onuchic et al.), tertiary structure forms as the protein slides downhill on the free-energy surface from the unfolded ensemble towards the native conformation. Each annealing cycle generates a small perturbation on the existing conformation, which averages to a 0.5 Å bRMSD change (max 3 Å initially). As new low-energy conformations replace old conformations, the population slides as a whole down the funnel of the free energy landscape.

Ensemble averages as a function of time over the moving population are thus associated with different stages of the structure formation process. In the lower panels of Fig. 4.5, we plot the average helical content and the number of beta-sheet H-bonds as a function of the cycle number. Following a rapid collapse to a compact conformation, the helix forms first, followed by the formation of the beta sheet. An analysis of the folding funnel upwards in energy illustrates that the lowest energy metastable conformations correspond to a partial unzipping of amino acids PHE25-ARG35, while the conserved cysteine residues are still buried. Even much higher on the free energy funnel (blue diamond in Fig. 4.3), we find many structures that have much residual structure, but essentially not long-range native contacts. The preformed sheet-region is stabilized by hbonds (LEU26-CYS27, ARG35) and packs at a right angle to the helix, the hydrophobic residues are only partially buried. This conformational freedom may be relevant in DNA binding, where the helical part of the zinc finger packs into the major groove of the DNA.

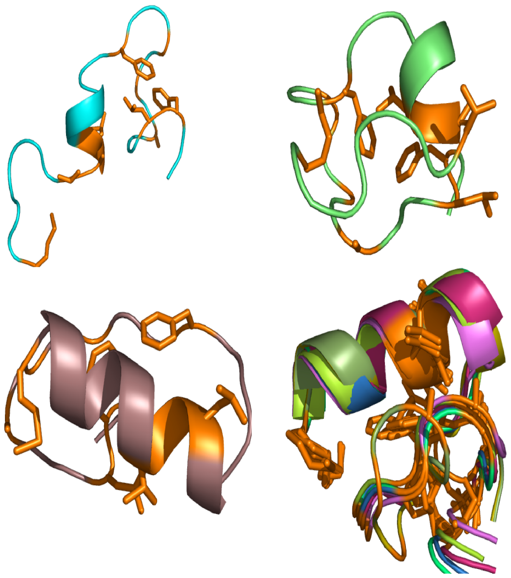


Figure 4.6: Key events in the folding: helix nucleation (top left), collapsed globular conformation (top right), fully formed helix(bottom left), partially formed beta-sheets using the helix as a template(bottom right).

De novo folding of the zinc finger domain permits a direct sampling of the relevant low-energy portion of the free-energy surface of the molecule as a first step towards the elucidation of the structural mechanisms involved in DNA binding (Laity et al., 2000). We find that much of the structure of the zinc finger is formed even in the absence of the metal ion that is ultimately required for the stabilization of the native conformation. Because the algorithm tracks the development of the population it is possible to reconstruct a folding pathway by reconstructing the sequence of events starting with converged conformation and moving backwards to the completely unfolded conformation.

Crucial steps along the continuous folding pathway are illustrated in Fig. 4.6 (note that there is no quantitative mapping onto the time axis). The early folding process is characterized by helix nucleation and concurrent collapse into a globular conformation with a radius of gyration that is comparable to that of the native conformation. The simulation then explores conformations of the same spatial extent with increasing helical, but no beta-sheet content. Lower in free-energy the simulation samples conformations in which partially formed beta-sheets pack against the helix. On the basis of the free-energy estimate to conformations without the helix (8 kcal/mol) such conformations can be explored in DNA binding and transcription. Our simulation approach permits a rapid exploration of this free energy region and thus characterizes the biologically active ensemble. MET51 packs in all low-energy conformations against the combined scaffold and acts as a closure of the DNA binding motif. It may thus provide an enthalpic contribution to a non-standard helix-capping motif that differs from the TGEKP linker sequence observed in multi-finger domains (Laity et al., 2001) in several zinc fingers.

We have thus demonstrated predictive all-atom folding of the DNA binding zinc-finger motif in a free-energy forcefield PFF02. This investigation offers the first unbiased characterization of the low-energy free-energy surface of the zinc finger motif, which is unattainable in coarse-grained, knowledge-based models. We find that the helix forms first along the folding path and acts as a template against which a variety of near-native beta-sheet backbone arrangements can pack. There are many zinc fingers with bRMSD of less than 2 Å to 1BHI (Nagadoi et al., 1999), this investigation provides thus one important step in the theoretical understanding of zinc-finger formation and function.

5

Prediction Studies

We have discussed the all atom folding of a zinc finger protein in the previous chapter. Now we turn our attention towards the related problem of finding the native structure of protein, but without recourse to folding pathways.

There is a growing gap between the number of protein sequences and structures. The structural genomics project (Burley et al., 1999) aims to narrow this gap and determine many of protein structures by careful combination of experiments. Computational methods for protein structure prediction (PSP) play an important role in such projects (Sali, 1998). The development of the reliable methods is essential for such task. In this chapter we discuss one of the prediction exercises for protein structure prediction.

5.1 Critical Assessment of Structure Prediction(CASP)

CASP is a large-scale community exercise for *blind* protein structure prediction held once two years. The main goal of CASP is to obtain an in-depth and objective assessment of the prediction methods. The organizers of the CASP exercise collect information for soon-to-be released experimental structures(targets) and distribute the sequence to the predictors. The predictors fall into two categories: participants who devote considerable time(about three weeks) for prediction(human groups) and automated servers which return predictions within 48 hours without human intervention(servers). The targets are divided according to prediction difficulty, such as modeling based on clear sequence relations(comparative modeling), modeling based on more distant evolutionary relations or on non-homologous folds(fold-recognition) and template-free modeling(new folds). The predictions are submitted in several formats including complete 3D coordinates, residue-residue contacts and 3D domain assignments. The quality of the models are evaluated by several measures.For example, the 3D coordinates are evaluated by C$_\alpha$RMSD (Root-mean-square deviation of C_α atoms) and the global distance test score(GDT_TS).

A decade of CASP experiments (Moult, 2005; Kryshtafovych et al., 2005) have monitored the progress in the prediction methods. Comparative modeling with accurate alignment have resulted in models within experimental resolution. There has been improvement in modeling the loop regions absent in the template. But this success is only achieved for targets with good homologous structures with at-least 30% sequence identity. Nevertheless, comparative modeling is very much useful for finding members of protein family performing similar function.

For the targets in mid-range of difficulty, alignment accuracy has improved by sophisticated methods which use multiple sequences for finding good templates. Another reason contributing to increased success is the use of meta servers (Rychlewski and Fischer, 2005; Ginalski et al., 2003) for choosing a consensus template. These methods have several drawbacks, which are mainly related to template selection and alignment and fail for template-free targets. In spite of the limitations, these methods provide an idea of overall structure or give information about molecular function based on template.

The new-fold target section is usually the most difficult section in a CASP exercise. Traditionally, physics-based approaches such as lattice models, coarse grain models which dominated this section have taken backseat to newer fragment based methods. There has been a steady progress for new-fold targets. In the CASP1, there were no successful predictions in this category. The quality of models has improved for at-least smaller targets. In CASP6, good models emerged for some of non-homologous targets,one within 1.5Å (Bradley et al., 2005) to the experimental structure. Nevertheless, these methods fail for targets with more than 100 residues. The bottle-neck for these methods is incomplete sampling and absence of reliable scoring functions for discriminating between more/less accurate models (Kryshtafovych et al., 2005).

The seventh instalment of CASP was held recently between May-August 2006. The exercise involved 100 targets, which were divided into 124 domains. Since targets consist of two or more structural domains, this division helps to capture the predictions accurately. The domain structures are judged by the assessors. The predictions were received in the following formats: 3D co-ordinates, alignments, residue-residue contacts, domain assignments, disordered regions, function prediction, quality assessment and model refinement. 207 human groups and 98 servers submitted 63717 models for above the sections [1]. The CASP targets were divided into mainly three categories based on sequence and structure similarity criteria:

- HA-TBM (High accuracy template based modeling): Target domains for which a suitable template was available and the best prediction had atleast GDT_TS of 80.

- TBM (Template based modeling): Target domains for which at least one structurally similar template was available and this template had been used in at least one prediction.

- FM (Free modeling): Target domains for which no structurally similar templates were identified or no template-based predictions were submitted.

[1] http://predictioncenter.genomecenter.ucdavis.edu/casp7/

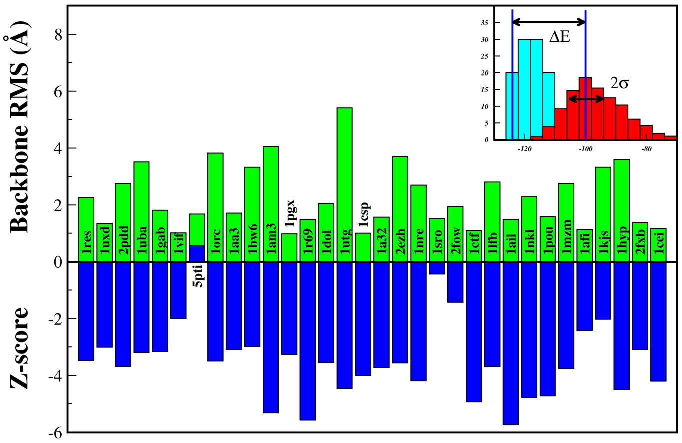


Figure 5.1: C_{α}RMSDs and Z-scores for proteins of Rosetta decoy set. The inset illustrates the concept of Z-score(see appendix B).This figure was taken from (Verma and Wenzel, 2007)

5.2 POEM-REFINE in CASP7

We participated in the CASP competition for the first time. The models were submitted under the group name POEM-REFINE. We chose to model all the medium sized targets with less than 150 amino acids. We submitted 3D models for 27 targets (out of which 2 targets were canceled), 4 of those were HA-TBM, 12 TBM and 9 FM targets respectively. Since CASP prediction had be submitted within three weeks after the sequence release, de-novo structure protein folding (starting from extended conformations) was not an viable option. The prediction protocol must simplify and enhance the conformation sampling. Heuristic methods such as Rosetta, can generate a large number of conformations with less computational effort. However Rosetta lacks an atomistic energy function which is able to distinguish between native and non-native structures (Bonneau et al., 2001).

We had earlier investigated a low cost free-energy refinement protocol with PFF02 for the decoys generated by Rosetta. The Rosetta decoy set consisted of 78 different proteins of various folds and presented a benchmark for the evaluation of the scoring functions (Tsai et al., 2003). We investigated, whether our free-energy function can distinguish the native structure from non-native conformations in this decoy set. However the decoys generated by one method (with *energy function A*) cannot be trivially ranked by another method(with *energy function B*). In order to obtain a meaningful estimate of the energy of a decoy, it must be relaxed to its nearby local minimum in B. After relaxation, the decoys can be ranked, since they will have proper energy estimates in B. The relaxation protocol used for the Rosetta decoy set was a single annealing run for 50,000 steps, with $T_{start} = 200K, T_{final} = 2K$ which was sufficient enough to push the decoys to their local minimum in the PFF02 force field. The relaxed decoys did not differ significantly from their initial structure.

The best energy found during the relaxation depends stochastically on the trajectories. One way to minimize this fluctuation is to generate longer or several independent trajectories. Since several decoys were available, the clustering of low-energy decoys was the most reliable option to balance selectivity and computational effort. Several ab-initio structure prediction methods use clustering techniques to distinguish the native and non-native structures (Moult, 2005). If a cluster contained at least 60% of all low-energy structures, then its average structure was used for the prediction.

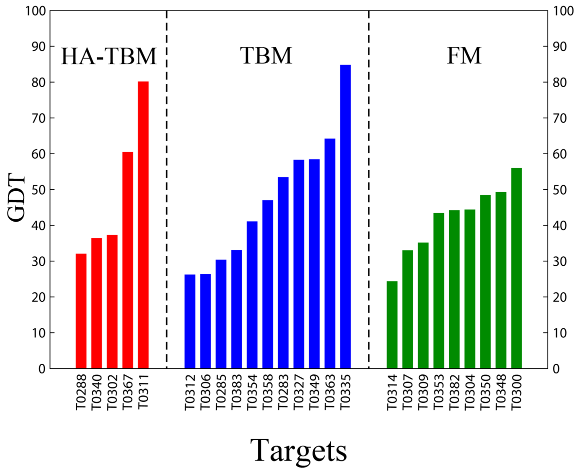


Figure 5.2: The best predictions(in GDT_TS measure) of POEM-REFINE for all CASP targets.

A subset of the Rosetta decoy set was chosen for this study, which consisted of 32 monomeric proteins of different folds (Verma and Wenzel, 2007). The decoy sets were categorized as: high-quality decoy(10% native-like) or a low-quality decoy set(less than 10% native-like).

The results for the decoy set is shown in Fig. 5.1. We were able to obtain conclusive results for the all but one protein of high-quality sets, where the prediction had an average C_αRMSD of 3 Å. For the low-quality sets the average C_αRMSD was slightly higher(6 Å). This study demonstrated that for given a good decoy set, a low-cost(20-50 CPU days) refinement method was capable to select the near-native decoys i.e. *picking the needle from haystack*. This study also highlighted the selectivity of force field PFF02 because we were able to obtain a high average value of the Z-score(-3.03) for all decoy sets.

5.2.1 Methodology

Motivated by these results, we applied a similar decoy refinement approach for our CASP7 predictions. The semi-automated prediction protocol consisted of three stages.

- Decoy set generation: The decoy set for a target sequence was generated using Rosetta. The details of Rosseta protocol was described in Sec. 3.4.2. Fragments for most of targets used three secondary structure predictors: Psipred (Jones, 1999), Sam99 (Karplus et al., 2001) and Jufo (Meiler et al., 2001). 10000-20000 decoys were generated for each target.

- Refinement: It would have been ideal to score all the 10000 decoys. But due to computational and time constraints, we used only 500-1000 decoys for refinement. These decoys were chosen from the most populous clusters. Clusters which contributed less than 50 members were not represented in refinement. The refinement protocol was identical to one described above.

- Selection: The 50 lowest energy decoys are clustered using a hierarchical clustering algorithm. The predictions were chosen from largest cluster. In absence of dominant cluster, predictions were picked from the large clusters by visual inspection.

5.2.2 Results

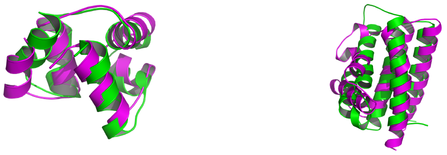


Figure 5.3: Overlay of predicted structures (magenta) with the native structure (green) for targets T0311 (left) and T0367 (right)

The results for the predictions are summarized in the Fig. 5.2. For the HA-TBM category, our average GDT score is 52.17 and best prediction is T0311 (80.08). For the TBM category the average is 46.64 and best prediction T0335 (84.72). Finally in the FM category we have an average GDT score of 41.94 and the best prediction is T0300 (55.89). The results for the individual sections are described briefly below.

Prediction for HA-TBM targets

The targets T0288, T0311, T0340 and T0367 belonged to the HA-TBM category. Our predictions for model T0311 and T0327 were quite accurate. The best prediction for T0311 had a C_αRMSD of 2.74 Å. The best prediction for T0367 differed from the native structure by a C_αRMSD of 4.66 Å. We were unable to predict good models for the targets T0288 and T0340. Both of these targets were predicted to high accuracy owing to highly homologous templates by other groups. The failure of our predictions was associated with the initial decoy set and also with the relaxation of limited number of decoys. The overlay for the successful predictions are shown in Fig. 5.3

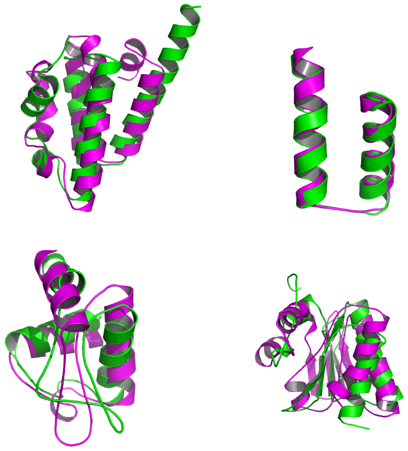


Figure 5.4: Overlay of predicted structure (magenta) with the native structures (green) for targets T0283 (left) and T0335 (right) [top panel], T0327 (left) and T0354(right) [bottom panel]

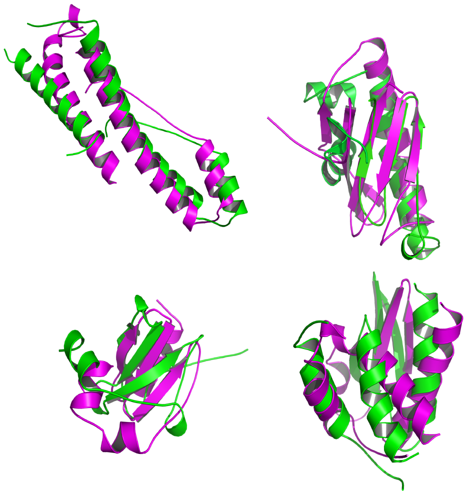


Figure 5.5: Overlay of predicted structures (magenta) with the native structure (green) for targets T0300 (left) and T0304 (right) [top panel], T0348 (left) and T0350(right) [bottom panel]

Prediction for TBM targets

The targets T0283, T0285, T0302, T0306, T0312, T0327, T0335, T0349, T0354, T0358, T0363, and T0383 belonged to the TBM category. We have predicted good models for the targets T0283, T0327, T0335, T0349 and T0363. Our model for target T0335 was one among the best of all submitted models. The best model for this target was predicted with experimental resolution of 1.6 Å C_αRMSD. Other good predictions in this category were models for target T0283 with 5.76 Å, T0327 5.11 Å and T0354 6.81 Å. The overlay for some of these predictions are shown in Fig. 5.4

#	Target	#*AA*	#*DAA*	POEM-BEST	
				GDT_TS	C_αRMSD Å
1	T0283	112	97	53.35	5.76
2	T0285	125	99	30.30	10.49
3	T0288	94	86	31.98	13.00
4	T0300	102	89	55.89	5.51
5	T0302	132	102	37.21	11.28
6	T0304	122	101	44.31	8.14
7	T0306	95	95	26.31	12.30
8	T0307	133	123	32.92	13.25
9	T0309	76	62	35.08	13.11
10	T0311	97	64	80.08	2.74
11	T0312	132	132	26.14	15.54
12	T0314	106	103	24.27	13.65
13	T0327	102	73	58.22	5.11
14	T0335	85	36	84.72	1.61
15	T0340	90	82	36.28	10.79
16	T0348	68	61	49.18	6.52
17	T0349	75	57	58.34	5.34
18	T0350	117	91	48.35	5.37
19	T0353	85	83	43.38	12.58
20	T0354	130	122	40.98	6.81
21	T0358	87	65	46.92	7.75
22	T0363	97	46	64.13	3.49
23	T0367	125	125	60.37	4.66
24	T0382	123	119	44.12	8.13
25	T0383	127	125	33.00	14.77

Table 5.1: Summary of the POEM-REFINE CASP predictions
#*AA* : No. of amino acids in the target sequence
#*DAA* : No. of amino acids in the CASP domain definition

Prediction for FM targets

The targets T0300, T0304, T0307, T0309, T0314, T0348, T0350, T0353 and T0382 belonged to this category. Three FM targets(T0304, T0348 and T0382) were also assessed in the TBM category. The prediction for targets of this category was hard due to the absence of homologous templates or lack of good fold recognition. Our prediction for T0300 was best submitted model with 5.51 Å C_αRMSD to the native conformation. It should be noted that T0300 has a long unstructured region, which was correctly identified by our model. The other targets where we succeeded in good prediction were the targets T0348 (6.52 Å C_αRMSD), T0350 (5.37 Å C_αRMSD) and target T0304 (8.14 Å C_αRMSD). The overlays of the predictions with experimental predictions are shown in Fig. 5.5. Tab. 5.1 summarizes the CASP targets which were predicted by our group.

We were able to identify strengths and the weakness in our methodology from this exercise. One of important factors contributing to failure of prediction is the quality of the initial decoy set. Our previous work on Rosetta decoy set (Verma and Wenzel, 2007) demanded the native content to be at least 10% for accurate prediction. Such high native content was not obtained in our CASP7 decoy sets generated by Rosetta. Fig. 5.6 shows the decoy distribution for four predictions: T0311, T0335 (best) and T0283 (average) and T0288 (worst). The decoy set for T0288 lacked native decoys, whereas T0283 had very few native decoys. The best prediction for these targets differed from the corresponding native structures by 2.74 Å, 1.61 Å, 5.76 Å and 13.00 Å respectively (Tab. 5.1). Thus we find that the quality of the decoy set correlates with the accuracy of the prediction.

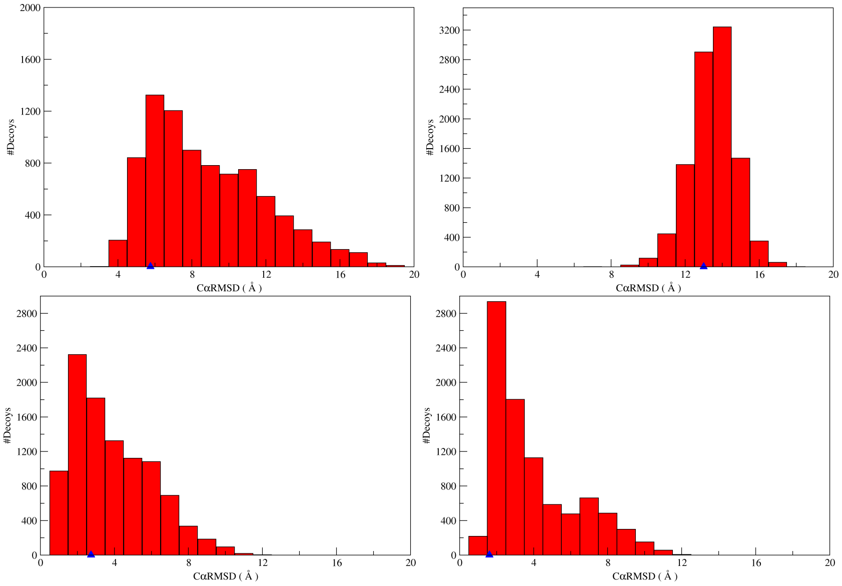


Figure 5.6: The distribution of decoys for targets T0283 (left) T0288 (right) in the top panel and T0311(left), T0335(right) in the bottom panel. The prediction for each target is marked by a blue triangle.

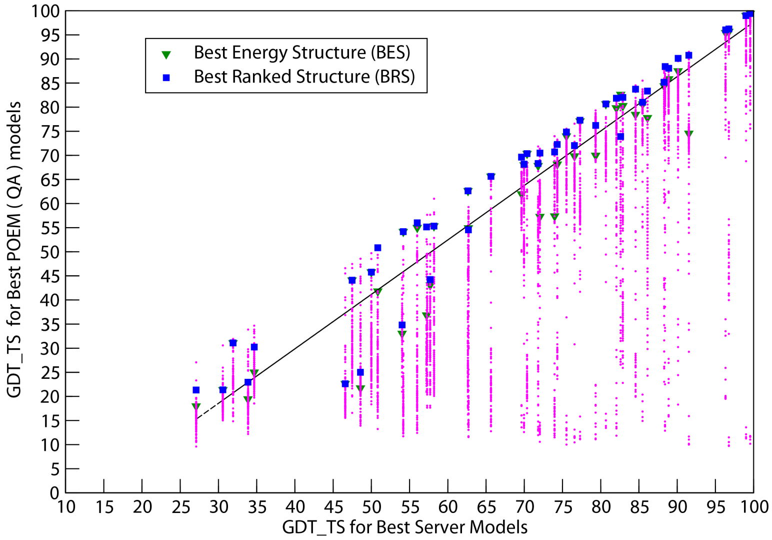


Figure 5.7: The Correlation plot for the Quality Assessment of server models

5.3 Quality assessment of CASP7 server models

Quality assessment(QA) was a new category in the CASP7 experiment. The task of this exercise is to chose the best model from the automated server predictions. The official experiment requires the participants to set a score for each of server model. During the CASP7 exercise, we lacked the computational resources to participate in the assessment exercise. However we performed our own QA on models for the CASP7 targets. The models were chosen from CASP7 targets which were less than 200 amino acids. Only models with complete prediction (without gaps) were used. There were on an average 125 structures for each target. We used the free-energy refinement protocol to rank the models of the targets. We define two parameters for the purpose of ranking:

- Best Energy Structure (BES) is the lowest energy structure for a given target.
- Best Ranked Structure (BRS) is the best structure (by GDT_TS) within the top 5% lowest energy structures.

There is an excellent correlation between the best server models and PFF02 ranked BES/BRS as demonstrated in the Fig. 5.7. Each discrete set of points correspond to server models for a particular target. The best server prediction occupy the top position for each target. The BRS/BES are found to be close to the best server prediction in most of cases(40 out of 48 targets)

We were unable to identify the best models for at least 8 targets. Four of such targets (T0361, T0299, T0306 and T0354) turned out to be oligomeric proteins which cannot be treated adequately with our approach. The server models for rest of targets (T0314, T0351, T0354 and T0358) were refined further. The results from extended simulations improved the BES/BRS.

The method of QA presented is a promising candidate for consensus protein structure prediction which is presently dominated by the meta-predictors such as Robetta (Chivian et al., 2005) and 3Djury (Ginalski et al., 2003). These servers are becoming significant in PSP, where their results are used in template selection/fold recognition/quality assessment purposes. The basic assumption used by meta-predictors is the fact that highly reliable models have less ambiguities in alignment and found more often by the servers. Thus a pair-wise simple similarity score (using C_α alignment) would suffice for the identification of native structure. In contrast our refinement protocol select models by energy criterion. The refinement protocol is effective for all categories of PSP, in particular for the free-modeling section where consensus server prediction fails.

6

Aggregation Studies

We have discussed protein folding and its structure prediction in preceeding chapters. Biological function of a protein strongly correlates with structure. In certain cases proteins fail to fold correctly or to remain correctly folded. Some misfolded proteins can form aggregates, which are recognized to be the origin of several pathological conditions. In this chapter we study the aggregation behavior of two such proteins.

Proteins in most living systems fold to unique 3D structure. The biochemical synthesis of protein contains a quality check mechanism which prevents the proteins from being misfolded (Sec. 2.4). Failure of such mechanisms leads to the formation of misfolded proteins or protein aggregates.

6.1 Amyloid diseases

More than twenty diseases are now found to be associated with the protein aggregates in variety of organs including liver, heart and brain. These diseases are termed as *amyloidoses* because the aggregated material stains dyes such as Congo red in manner similar to starch (amylose). The diseases include Alzheimer's disease, Creutzfeldt-Jakob disease and Type II diabetes (Dobson, 2001). In addition to these diseases, Parkinson's and Huntington's disease appear to involve similar aggregates, but the aggregates are intracellular unlike those in amyloidoses. The striking feature of amyloid diseases is that the associated fibrils are very similar in their overall properties and appearance (Sunde and Blake, 1997). This feature has enabled the in depth study of one model system from which one can try to extract the mechanism and properties of others. Another important property of the proteins involved in these diseases is that fibrillar forms are generated easily *in-vitro*, which has enabled multitude of experimental studies. The experimental methods for structure characterization of aggregates are X-ray diffraction, electron microscopy and solid state NMR. Several structures have been proposed for the amyloid fibrils formed by the SH3 domain, Aβ(1-40,1-42,11-25), sup35p, insulin, lysosome (Makin and Serpell, 2005) proteins. Some current therapeutic strategies for these diseases involve design of mutant proteins which stabilize the native structures or molecules which inhibit the amyloid formation.

6.2 Amyloid β Aggregation

The 39-40 residue Aβ peptide is part of the β-amyloid precursor protein (APP) which is found to form extracellular aggregates in Alzheimer's disease patients. The aggregates are found to form fibrils which are about 100 Å in diameter. The accumulation of these fibrils leads to formation of so called *senile plaques* in the diseased brain. These plaques are suspected to be one of causes for the cell death and tissue loss in the patients.

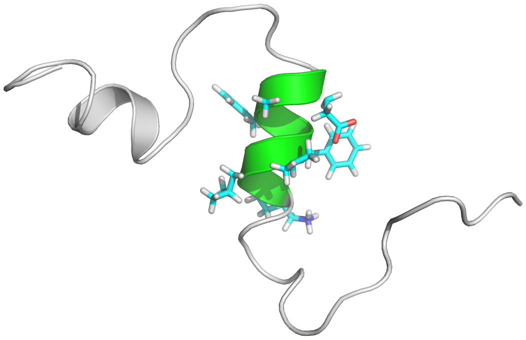


Figure 6.1: The experimental structure of the (1-40)βAmyloid protein (PDBID:1AML) has partial helix structure. The green region shows the (16-22) segment used for the present aggregation study.

The Aβ peptide is derived from proteolysis of APP and occurs mainly as either 40 (1-40) or 42 (1-42) aminoacid variants. The experimental structure of the monomeric protein is shown in the Fig. 6.1. The monomer contains 65% disordered regions and occurs as soluble component in the body fluids. However the oligomeric forms of this peptide form insoluble fibrils. There are several experimental studies on the fibrillar structures of this protein containing the (1-42), (1-40), (11-25), (16-22) and (34-42) amino acid segments. These studies suggest stacked parallel in-register βsheet structures for the (1-42) segment (Lührs et al., 2005) and antiparallel arrangements for (34-42) and (16-22) segments (Balbach et al., 2000) respectively. There are extensive computational studies on the Aβ fibrils including lattice model simulations (Jang et al., 2004; Dima and Thirumalai, 2002) and atomistic simulations (Klimov and Thirumalai, 2003; Wei et al., 2007; Favrin et al., 2004; Meinke and Hansmann, 2007; Jang and Shin, 2006; Cecchini et al., 2006).

The seven amino acid segment of Aβ(16-22) contains five hydrophobic aminoacids sandwiched between the polar residues: LYS^{16} **-LEU-VAL-PHE-PHE-ALA-** GLU^{22}. The five hydrophobic residues constitute a central are identified as key element in aggregation (Tjernberg et al., 1996). Since this short fragment is known to be essential for full-length peptide fibrillation, it is well suited for probing mechanisms of aggregation and fibril formation. The study of this fragment is also moti-

Type	Non-rigid body moves			Rigid body moves		
Backbone	LIB	18.67%	**56%**			
	REL	18.67%				
	PIV	18.67%				
Sidechain	REL		**24%**			
Interchain				TRN	10%	**20%**
				ROT	10 %	

Table 6.1: The probabilities for different types of moves are shown. LIB: library move, REL: relative move, PIV: pivot move, ROT: rigid body rotational move and TRN: rigid body translational move

vated by the existence of several important mutations of Aβ which occur in the region (21-22). Thus a comparative study of the aggregation of Wild-type and Mutants are possible for this segment.

We have recently adapted our protein simulation package POEM to treat oligomers. Here we study the aggregation of Aβ(16-22) using the free-energy model PFF02 (Sec. 3.4.3).

6.2.1 Protocol for study of oligomers

We made certain changes in the simulation protocol for simulating the oligomers:

- **Energy function**: The PFF02 energy function used for earlier folding and prediction studies was used for oligomeric studies, without any change of parameters. However a simple pseudopotential was added to constrain the oligomers to a certain distance. This potential contributed an energy penalty if the oligomers are separated by a certain predefined distance. The maximum separation allowed between a pair of monomers without energy penalty is 10 Å, but there is no force guiding the monomers together.

- **MC moves**: For the simulation of monomers a conformational change is induced by single dihedral angle change of sidechain(30% probability) and backbone (70% probability). Half the backbone moves were generated from an equidistributed interval with maximum change of 5^o, while the other half was drawn from a library comprising of dihedral angles occurring in proteins. We have included rigid body translational and rotational moves to treat aggregates. Translational moves are drawn from an equidistributed interval with maximum change of 1 Å, whereas the rotational moves are drawn from an equidistributed interval with maximum change of 0.1 Rad. An additional move, called pivot move, which induces conformational change by simultaneously rotating six backbone dihedrals was also used. The probabilities of the moves are given in Tab. 6.1

- **Optimization**: We have used the optimization methods : gBHT (Wenzel, 2006) and ES (Sec. 4.1)

Monomer Simulations

We started gBHT simulations on extended structures of the Aβ(16-22) fragment. We performed 100 cycles of the gBHT on ten identical copies. The starting temperatures were chosen from an exponential distribution which has an average value of 750K. The number of steps increased with the gBHT cooling cycle by $10^4 \times \sqrt{m}$ where m is the number of minimization cycle. The energy threshold used for acceptance after each gBHT cycle is dynamically adjusted during the course of simulation (Wenzel, 2006).

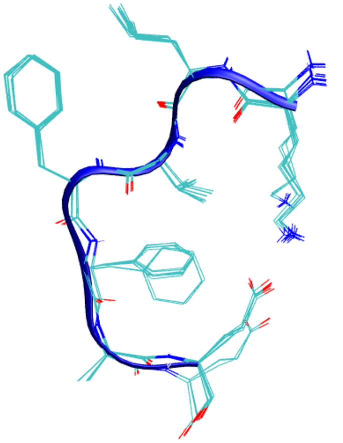


Figure 6.2: The conformations of Aβ(16-22) sampled by the ten gBHT simulations. The segment is always unstructured. The sidechain conformations for all the ten final structure are also shown.

All gBHT trajectories converged to very similar structures, as illustrated in Fig. 6.2. The predicted structure is a random coil. The native structure of this segment in the full protein is a helix Fig. 6.1. Earlier works on monomers indicated that they exist mostly as random coils at room temperature (Favrin et al., 2004). The strand conformation is also found to exist with smaller propensity (Klimov and Thirumalai, 2003)[1], while a recent study also indicated a helix conformation (Meinke and Hansmann, 2007). It should also be noted that each of these investigations used different definitions for secondary structure elements. Our results suggest that the Aβ(16-22) lacks a regular secondary structure. Neither strand or helix conformation are observed in the low-energy structures.

[1] This study reported their trajectories for Aβ(16-22) had 68% random coil and 29% strand conformations

6.2.2 Dimer Simulations

We have studied the interactions of two chains(dimers) using the gBHT method. First we discuss the aggregation of the Wild type(WT) dimers.

WT Dimers

Two extended chains of the WT are the starting structure for the gBHT simulation. The chains are initially separated by 10 Å. Ten independent gBHT simulations are performed on these dimers. The lowest energy structure is an anti-parallel beta sheet conformation which has an energy of -19.68 kcal/mol.

A significant percentage of the accepted conformations consisted of anti-parallel conformations. Fig. 6.3 shows the distribution of all the accepted structures during gBHT simulation. Also shown are the three different alignments of the beta strands. High energy structures with parallel and perpendicular orientations were also found with lower probability. Our results for the dimers agree with the earlier work (Wei et al., 2007) which indicated higher propensity for anti-parallel conformations. These conformations are stabilized by the a favorable side chain interactions and backbone hydrogen bonding.

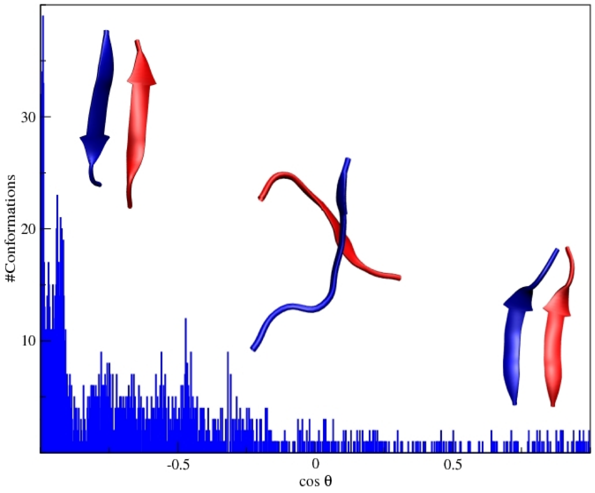


Figure 6.3: The distribution of the conformations as function of cosθ, where θ is angle between the end-to-end unit vectors of individual chains. The majority of conformations have anti-parallel orientation. Also shown are representative structures of the distribution. The lowest energy structure (left) has an energy of -19.68 kcal/mol.

Mutants

We have also examined three mutants of the Aβ. The mutants are found to be associated with familial Alzheimer's disease. The mutations are concentrated between the segments (21-23) of APP. These

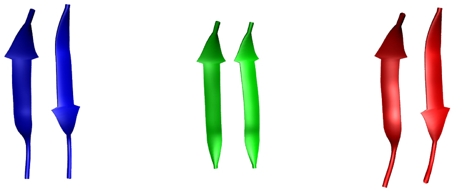


Figure 6.4: The lowest energy structures of Aβ mutants: Flemish (A21G), Italian (E22K) and Dutch (E22Q).

mutants are named Flemish (A21G), Arctic (E22G), Dutch (E22Q), Italian (E22K), and Iowa (D23N). We have performed the gBHT simulations for dimers of the Flemish, Dutch and Italian mutants. The lowest energy structures for these mutants are shown in the Fig. 6.4. The lowest energy structure for the Flemish and Dutch mutants are anti-parallel beta conformations. In contrast, the lowest energy structure for Italian mutant is a parallel beta conformation.

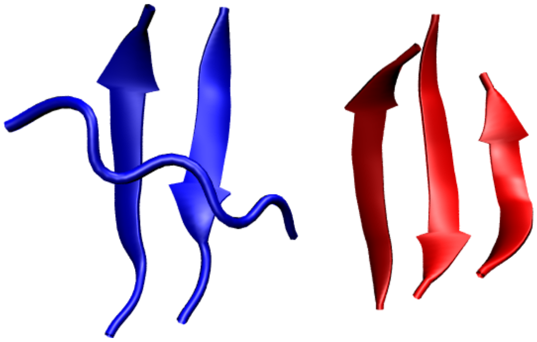


Figure 6.5: The lowest energy structures for the 3mers of WT found during the gbht simulation (left) and in the ES (right). The three strand antiparallel conformation is lower in energy by 7.16 kcal/mol compared to two stranded structure

6.2.3 Trimer Simulations

The 3mers for the WT are studied by both the gBHT and the ES. The protocol for gBHT method consisted of 200 cycles of twenty independent simulations. The lowest energy conformation found in these simulations is shown in the Fig. 6.5 and has an energy of -39.55 kcal/mol. The structure has two strands in anti-parallel orientation. The third partially formed strand occur at a right angle to this β sheet region. Most of the low energy structures are similar the above structure, differing only in orientation of third chain. The higher energy structures with parallel orientation occurs to a lesser extent.

We have also explored the same 3mer system with the ES. The starting structures used for this simulation is identical to that used before. The simulation is performed for 200 cycles. The lowest energy structure found by the simulation was an anti-parallel beta sheet conformation with an energy of -46.71 kcal/mol. This structure is shown in Fig. 6.5. There are other low energy conformations which resemble the ones found in previous gBHT simulation. The energy of the best structure is plotted as function of ES cycle in Fig. 6.6. The lowest energy structure found after 145 cycles is preserved until the end of the simulation. The presence of anti-parallel beta sheets is attributed to formation of salt bridges (Klimov and Thirumalai, 2003), though other work (Favrin et al., 2004) which neglected the side-chain Coloumbic interactions between charged residues yielded same result. The anti-parallel beta sheets are favored mostly due to a optimized hydrophobic side chain packing, though the presence of salt bridge tends to increase the stability.

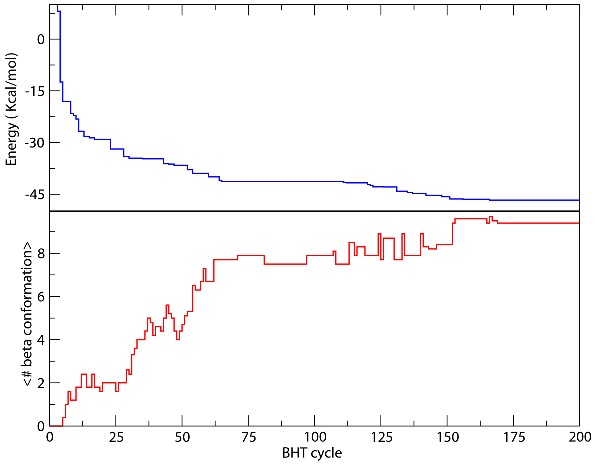


Figure 6.6: The lowest energy of each ES cycle is plotted as function of ES cycle. Also shown is the average strand conformation of the active population.

One of the earlier studies on 3mers indicated the existence of an obligatory α helical intermediate (Klimov and Thirumalai, 2003). To search for presence of such intermediate, we have analyzed all the accepted structures during the simulation. We found no structure in the helical conformation. Several other works (Favrin et al., 2004) also found no sign of an α helix intermediate. Fig. 6.5 shows the average beta content [2] for these active population as function of ES cycle. We find that average beta content correlates well with the best energy.

6.3 Insulin fibrils

Insulin is a naturally occurring hormone which is used for regulation of the glucose content in the cell. The protein consists of two small chains A (21 residues) and chain B (30 residues), linked by two in-

[2] Number of residues in strand conformation

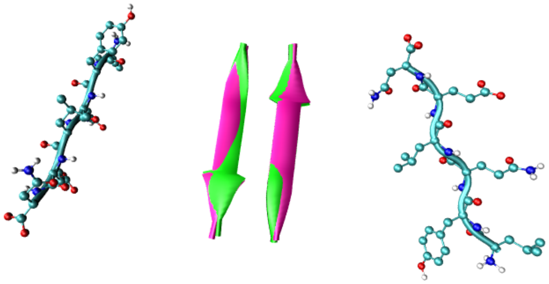


Figure 6.7: The lowest energy structure(magenta) of the insulin fibril has an bRMSD of 0.54 Å with respect to the experimental structure(green). Also shown is the starting conformations for the gBHT simulation.

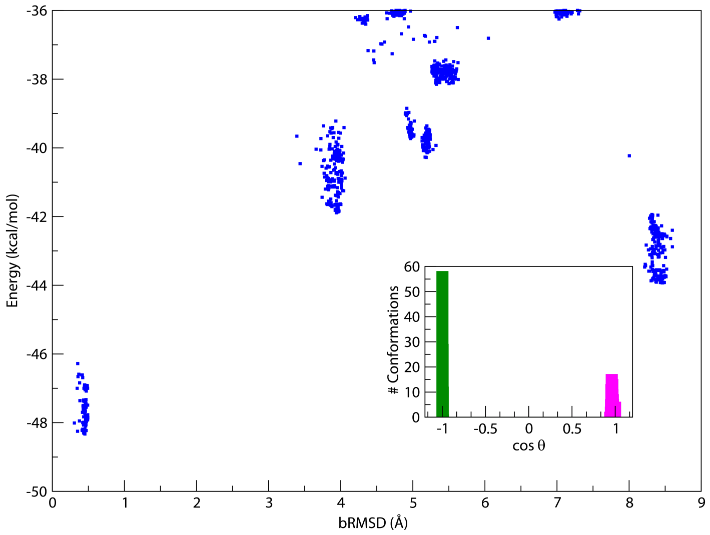


Figure 6.8: The energy of the decoys is plotted as function of their bRMSD. The inset shows the distribution of top 100 decoys as function of cosθ, where θ is angle between the end-to-end unit vectors.

terchain disulphide bonds. In solution, insulin can exist as monomers, dimers or tetramers depending on the pH, solvent composition and presence of metal ions. It is also known that insulin is prone to aggregation under slightly destabilizing conditions. The insulin fibrillation is proposed to occur through the disassociation of the oligomeric units into the monomer, followed by a conformational change. (Brange et al., 1997). A recent investigation on individual chains found that independent A and B chains can undergo fibrillation, though the rate of formation depends on the factors such as pH and concentration. (Hong et al., 2006).

A three dimensional structure of insulin aggregate was not available for a long time. Recently experimentalists were able to resolve the fibrillar structures for segments of chain A and B. A well characterized 3D structure is now available for the segment $LEU^{13} - TRY - GLN - LEU - GLU - ASN^{18}$ of chain A (PDBID:2OMP) and $VAL^{12} - GLU - ALA - LEU - TRY - LEU^{17}$ of chain B (PDBID:2OMQ). We have explored the fibrillar formation of chain A.

The starting structure for this study are two extended chains of LYQLEN separated by 10 Å as shown in Fig. 6.7. We have performed twenty independent gBHT simulations for 200 cycles. The lowest energy structure is close to the experimental structure with 0.54 Å bRMSD (-48.31 kcal/mol). Only two of ten gBHT simulations found the native structure. The other simulations terminated in high-energy non-native ensembles.

We have analyzed the accepted conformations during the simulation. A plot of energy as function of their bRMSD with respect to experimental structure (Fig. 6.8) reveals the existence of very few accessible states. The native ensemble has an average energy of -47.619 kcal/mol. The next lower energy ensembles contains parallel beta sheet conformation(average energy -43.192 kcal/mol) and partially formed anti-parallel beta sheet ensemble (average energy -41.126 kcal/mol). We find that the anti-parallel orientation is found to be preferred over parallel orientation for the low energy structures(Fig. 6.8).

7

Conclusions and Outlook

Proteins are the workhorses of the cell. The amino acid sequence encodes the precise three-dimensional structure which is essential for their function. Protein-protein interactions are responsible for many biochemical mechanisms in living organisms. In certain cases a protein fails to fold or to remain correctly folded. Such failures are often causes of pathological conditions. Therefore the study of protein structure, folding and interactions is essential for understanding of many biological processes. Considerable experimental and theoretical efforts have addressed these problems. The experimental methods, such as X-ray diffraction, have contributed high resolution static structures. NMR experiments have helped in resolving the structure of the native state ensemble and also in understanding the folding process. Theoretical methods have also investigated structure prediction, folding kinetics and protein-protein interactions. Knowledge-based methods are already able to predict experimental resolution structures for certain protein families within experimental accuracy. Simplified protein models, such as lattice models and G$\bar{o}$ models are able to qualitatively explain the folding process. MD methods have been used widely in understanding the folding process, aggregation studies and enzymatic reactions for small peptide systems. De-novo protein modeling is one of the promising approaches for protein structure prediction. Despite such progress, many problems remains to be addressed:

- Protein structure prediction: Knowledge based methods are limited to known protein classes, whereas *de novo* methods are often plagued by the lack of an accurate energy function and in efficient sampling.

- Protein folding: An average size globular protein domain (about 150 amino-acids) folds in millisecond time. The conventional MD methods are able to simulate the complete folding dynamics for some peptides. However these methods have been limited to the study of small peptides (about 30 amino-acids), though at huge computational cost.

- Protein-protein interactions: Generally, protein-protein docking is performed by matching the individual protein shapes with a pseudo-potential. Though these methods are fast, they neglect the actual physical interactions which establish the conformation of the complex.

Using an all-atom free-energy forcefield and efficient optimization methods we are able to address some aspects of the folding, structure prediction and interaction problems discussed above.

To start with this work, we had an atomistic free-energy function PFF01/02 which has been capable of identifying the native structure of large family of proteins. We were interested in applying PFF02 to various aspects of protein science: protein folding, structure prediction and protein-protein interactions. So we have devised new methods for the above mentioned aspects of the biomolecular structure problem.

Our first goal was to improve the existing optimization methods. Several methods, such as basin hopping technique (BHT), were successful in our earlier studies on small proteins. The independent BHT simulations, however offer only limited sampling of a very complex protein energy landscape. First the BHT method was modified to overcome such problems. This modified method, greedy BHT enhanced the sampling of lower energy region by preserving the best energy structures, which in case of the original BHT were sometimes lost in the simulation. The greedy BHT (gBHT) scheme was demonstrated to be a very efficient optimization technique for small proteins.

Both BHT and gBHT involve several independent simulations which explore the free energy landscape. It would be desirable to develop a simulation protocol, where several concurrent simulations exchange the information to learn about their respective sampling regions. The gBHT simulations exchange information about the current lowest energy, which enables the simulation to preserve the best energy structure, but often the independent BHT and gBHT trajectories converge to identical conformations. Fortunately we could develop a scheme to generalize the method to multiple interdependent trajectories. In the evolutionary strategy (ES) a population of size N which is iteratively improved by P concurrent process. The algorithm is implemented as a master-client model, where the master keeps a copy of population and clients perform local short minimizations on a chosen member of population, which is then returned to the master. The master decides about the fate of new structure based on a set of rules favoring a diverse population. We have improved this method by modifying the rules involved in population update, which is the crucial stage of their ES. The acceptance criterion was adjusted to enhance population diversity and energy enrichment. The algorithm was implemented on massively parallel architectures and found to scale up to 4096 processors on the IBM BlueGene supercomputer.

One of advantages of ES over BHT is that the evolution of members of the population can be monitored (i.e. each member can be traced back to its ancestor). Therefore it is possible construct a folding pathway by reconstructing the sequence of events starting with converged conformation and moving backwards to the completely unfolded conformation. This pathway characterizes the crucial states involved in the folding process. However, unlike MD there is no time evolution in the trajectory. Nevertheless the folding path helps us to qualitatively describe the folding process.

Having improved the optimization technique we focused our attention towards protein folding. The protein which we selected for this purpose was a 29 amino-acid zinc finger motif with an $\alpha\beta\beta$ fold. No protein with mixed secondary structure was folded from an extended structure in our previous simulations. Starting from a completely unfolded conformation with no secondary structure (16 Å bRMSD) we performed 200 cycles of ES. Among the ten energetically lowest conformations, six folded into near-native conformation with bRMSDs of 3.68-4.28 Å. The conserved hydrophobic side chains were buried in the folded structures, with the zinc-coordinating cysteine residues within 2 Å of their native positions and available for association with zinc ion.

With this simulation we could illustrate the crucial steps involved in the folding process: The folding process was characterized by helix nucleation and concurrent collapse into globular conformations. These structures further evolve by first forming the complete helix and then beta sheets later in the process. Our work was one of the first unbiased characterization of the free energy surface of the zinc-finger motif. Since there are several zinc fingers with bRMSD of less than 2 Å to the protein we studied, this investigation provides one important step in theoretical understanding of zinc-finger formation and function.

Our earlier work on the Rosetta decoy set demonstrated the transferability and selectivity of forcefield PFF01/02. This study illustrated the success of the methodology in identifying the native structure, when a good decoy set(with at least 10% native structure) was available. This result inspired us to design a free-energy refinement protocol for protein structure prediction. Heuristic methods, such as Rosetta, are able to generate large number of conformations with little computational effort. However, Rosetta lacks an atomistic energy function, which is able to distinguish between the native and non-native structures. Since PFF02 has demonstrated high selectivity for the Rosetta decoy set, we decided to combine these two approaches. We devised a scheme for structure prediction, where we generate thousands of decoys for a given protein sequence with a heuristic approach, relax these decoys with our free-energy refinement protocol and chose the native structure from the cluster of lowest energy structures. Initial tests on several proteins including few CASP6 targets, were promising.

We participated for the first time in the 7th CASP blind protein structure prediction exercise. CASP organizers collect the information for soon-to-be released experimental structures (targets) and distribute only the sequence information for the predictors. The predictors had three weeks time for submitting the prediction. We applied the free-energy refinement protocol for 27 targets and submitted predictions. Our predictions in the free modeling category were satisfactory. Our model for T0300 (102 amino-acid, 5.51 Å bRMSD) was the best submitted prediction among 300 prediction groups. Other good models for this category includes T0304 (122 amino-acid 8.14 Å bRMSD), T0348(68 amino-acid, 6.52 Å bRMSD). In the TBM category our successful predictions included T0283 (112 amino-acid, 6.51 Å bRMSD), T0327 (73 amino-acid, 5.11 Å bRMSD), T0335(36 amino-acid, 1.61 Å bRMSD) and (46 amino-acid, 3.49 Å bRMSD). It should be mentioned that knowledge based methods yielded better models compared to our predictions in the TBM category. The HA-TBM category included targets with a high degree of homology. So our good predictions for targets T0311 (64 amino-acid, 2.74 Å bRMSD and T0367(125 amino-acid, 4.66 Å bRMSD) were surpassed by experimental resolution structures from knowledge based methods.

We were unable to participate in the official quality assessment exercise due to lack of computational resources. After the CASP exercise, we ranked the automated server models for 48 targets which contained less than 200 amino acids. Our free-energy refinement methodology was able to pick the best available structure in 40 cases. This result is encouraging and suggests a promising method for consensus protein structure prediction, where the native structure is determined from a set of models from automated servers.

Our protein simulation software POEM was improved to treat multiple protein chains. We used the PFF02 energy function to study two different aspects of protein-protein interactions: First, we studied the aggregation of the oligomers of 16-22 segment of $A\beta$ protein. The native $A\beta$ protein has partial helix structure. However, the oligomers of $A\beta$ form insoluble fibrils which are neurotoxic. We studied the aggregation of dimers of $A\beta$ and found that it forms anti-parallel sheets. We also performed simulation on the mutants of $A\beta$ and found that they were also prone to aggregation. The trimers of $A\beta$ were also found to form anti-parallel beta sheets. Our study confirmed the absence of proposed alpha-helix intermediate that were observed in some earlier studies.

Our next goal was to test and implement a computational scheme for studying protein-protein docking. We chose two simple systems to investigate protein dimers. Our protocol was able to identify the native interactions which stabilize the dimer conformations. We find that lowest energy structures agree with the experimental structures within 1.0 Å bRMSD.

7.1 Outlook

We have discussed the application of the all-atom free-energy approach for protein folding, protein structure prediction and protein-protein interactions. Though we were quite successful in each of these problems, there is room for further improvements. The protein folding mechanism illustrated in this work is just one step towards the full description of the problem. We still need to characterize the kinetics of the folding problem. Since our method exhaustively samples the low-energy regions, we can perform an analysis of all relevant low-energy conformations by using a master equation approach assuming the diffusive processes between similar conformations. In addition, such a study will qualitatively validate our method for the folding dynamics.

The prediction protocol used for CASP7 can surely be improved further. Our predictions for the free modeling section were satisfactory, but we were unable to generate good predictions (compared to the knowledge based methods) for targets with highly homologous structures. Comparative modeling methods dominate this section and we need to develop a scheme which combines comparative modeling and free-energy refinement. Such a method will be in principle as good as the comparative modeling for high homology structures. We need computational resources for extending the prediction protocol to larger proteins. One of the approaches to gather much computational power is the distributed computing approach. We are working towards a distributed computing project for protein structure prediction using the the Berkeley open infrastructure for network computing (BOINC) platform. The personal computer (PC) users around world can install the software on their computer and contribute computational resources to chosen applications. Distributed computing projects such as search for extraterrestrial intelligence (SETI) and Folding@home have been successful. There is two-fold advantage with such schemes. First, one can tap into the enormous but under-utilized calculating power of world-wide PC. Secondly such projects increase the public awareness of science and to a certain extent, democratize the allocation of research resources.

Our aggregation study was limited to trimers. Larger oligomeric systems are needed for complete characterization of aggregates. We have taken a first step towards studying such systems. We are also studying the the mutants for the $A\beta$ peptide. Currently there are few theoretical attempts which target the inhibition of the aggregation. We are interested in designing inhibitors which prevent aggregation.

The biological functions of protein are generated by interaction with environment. Protein-protein interactions is a crucial process for biological activity. We had demonstrated a test case for protein-protein docking for simple systems. Now we are interested in extending our methodology for complex systems. We are working on a improved protocol for enhanced sampling of conformations. Currently we are testing this on a protein-protein docking benchmark. The far future aim will be development of efficient technique which can be applied to the blind prediction challenge critical assessment of predicted interactions (CAPRI).

By combination of an accurate forcefield PFF02 and efficient optimization methods, we were able to study a broader aspect of protein structure. This work demonstrated the folding of a mixed fold protein, a methodology for a de novo protein structure prediction and protein-protein interactions. This work is one small step towards the understanding of protein folding, structure and interactions.

8

Appendix A

8.1 Test cases for protein-protein docking

We have studied the aggregation behavior for two small systems in the Chapter 6. The method used there is in general applicable to any system of protein oligomers. As a test case for protein oligomers we study two systems of oligomers. Presently we concentrate only on the rigid protein-protein docking. The MC move probabilities are modified for this purpose (Tab. 8.1).

Type	Non-rigid body moves			Rigid body moves		
Backbone	LIB	0.0%	**0.0%**			
	REL	0.0%				
	PIV	0.0%				
Sidechain	REL		**33.3%**			
Interchain				TRN	33.3%	**66.6%**
				ROT	33.3 %	

Table 8.1: The probabilities of MC moves for rigid docking are shown. LIB: library move,REL: relative move, PIV: pivot move, ROT: rigid body rotational move and TRN: rigid body translational move.

The rigid-body moves constitute 66.6% of the total moves. The side-chain moves contribute the rest. The absence of back-bone moves is not crucial for test case proteins, since the monomers doesn't undergo a conformational change when they interact.

We study the docking of two olgiomeric proteins.

- **Rop-dimer**: The Rop-dimer (Repressor of primer) is a homo-dimer which is able to bind the RNA. The molecule consists of two helix-turn-helix chains are in parallel orientation (chain A 63 residues and chain B 63 residues). The structure is found to be stabilized by the hydrophobic

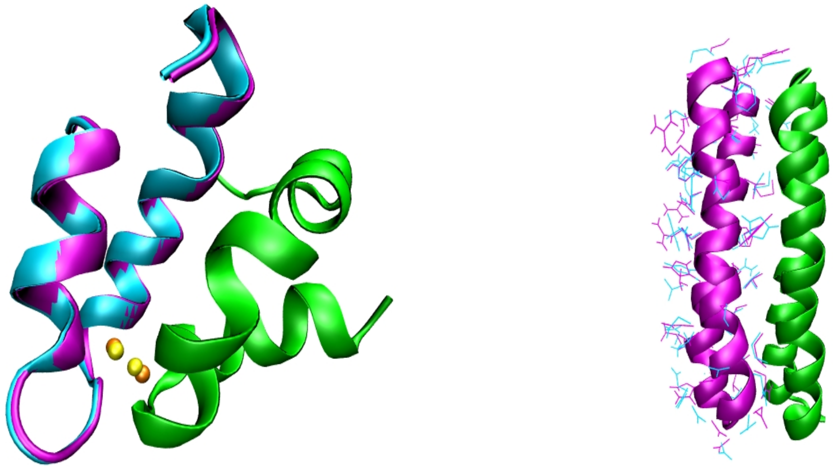


Figure 8.1: The lowest energy structures for rigid docking of rop dimer(right) and ectatomin (left). For ectatomin(left) the sulphur atoms of cysteines(experimental : yellow, predicted : orange) which form interchain disulphide bonds are also shown.

interface between the monomers. We have chosen one variant of ROP dimer (PDBID:1RPR) for this study.

The starting structures for the simulation is an unbound monomers. We also randomized the side-chain of monomers by performing high temperature MC. The 100 cycles of gBHT simulations are performed on 20 copies of these unbound species. As mentioned above, the back-bone moves were frozen. of our gBHT simulations docked the monomers within 2.5 Å bRMSD with respect to native structure. The lowest energy structure differed by just 0.48 Å bRMSD and had an energy of -266.26 kcal/mol. This structure is shown in Fig. 8.1.

- **Ectatomin**: Ectatomin is an ant venom responsible for the toxic effect in the both mammals and insects. It is found to form channel insert to the plasma membrane by forming channels. The molecule consists of homologous chains (chain A 36 residues and chain B 34 residues). Each monomer consists of two αhelices with a connecting hinge region. The two chains are bound by an intrachain disulphide bridge in the hinge regions. Our forcefield PFF02 lacks a explicit descriptor disulphide bridges, so we were interested in finding effect disulphide bridge for bound monomers.

  The 100 cycles of gBHT simulations were performed on 10 identical unbound monomers. Five out of 10 simulations found the docked monomers within 1.0 Å. The lowest energy structure has an energy of -127.060 kcal/mol and differs from native structure by .5 Å(Fig. 8.1). The sulphur atoms of cystenies in chain A and B are separated by 3.0 Å in this structure compared to 2.0 Å separation in the native (Fig.8.1)

9

Appendix B

9.1 Software programs used

- **POEM**: All the simulations are performed using the POEM software. POEM allows several optimization methods and is used in conjunction with PFF02.

- **Rosetta++**: Rosetta++ software suite in conjunction with Rosetta ab-initio protocol is used for generating decoys in the CASP7 exercise. This software is available for free to academic labs at *http://www.rosettacommons.org/*

- **Psipred**: Psipred is used for assigning secondary structure predictions for the fragment generation phase of Rosetta ab-initio protocol. Psipred is available at *http://bioinf.cs.ucl.ac.uk/psipred/*

- **MMTSB**: Software tools from MMTSB suite are used for analyzing pdb files. MMTSB is available at *http://mmtsb.scripps.edu*.

- **TMscore**: The GDT_TS and $C_\alpha$rmsds were calculated using the TMscore. TMscore can be obtained from *http://zhang.bioinformatics.ku.edu/TM-score/*

- **Pymol**: Pymol is used for analyzing the protein structures. Some of protein images are rendered with Pymol. This software is available at *http://pymol.sourceforge.net*.

- **VMD**: VMD is used for analyzing the protein structures. Some of protein images are rendered with VMD. This software is available at *http://www.ks.uiuc.edu/research/vmd*.

- **Scwrl**: Scwrl is used to add the side-chain rotamer. This software is available at *http://dunbrack.fccc.edu/SCWRL3.php*

- **Kile**: This thesis is typesetted using Kile, a front end to LaTeX

9.2 Definitions

RMSD

RMSD measure is used to compare two protein conformations. It is defined as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta^2}$$

where N is the number of atoms and δ is the distance between pair of equivalent atoms in these conformations. The structures which are to be compared are translated and rotated to obtain the best possible fit before computing the RMSD. We have used two variants of RMSD : bRMSD which uses the backbone atoms N, C_α, C and O or C_αRMSD considering only C_α atoms.

DSSP

DSSP (Kabsch and Sander, 1983) is used to characterize the secondary structure of proteins. The secondary structure is assigned based on hydrogen bonding patterns. DSSP defines an existence if the electrostatic energy defined by equation is $\leq$ 0.5 Kcal/mol.

$$HB_{12} = 332 q_1 q_2 \left(\frac{1}{r_{C_2H_1}} - \frac{1}{r_{C_2N_1}} - \frac{1}{r_{O_2H_1}} + \frac{1}{r_{O_2N_1}} \right)$$

where $r_{A_iB_j}$ are interatomic distances between atoms A_i and B_j. The value of q_1 and q_2 is set to 0.3e and 0.2e respectively which correspond to absolute partial charge on backbone carbonyl $C = O$ and amide $N - H$.

The following secondary structure elements are assigned.

- G = 3-turn helix (3_{10} helix). Min length 3 residues.
- H = 4-turn helix (right handed α helix). Min length 4 residues.
- I = 5-turn helix (π helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand which is component of beta sheet conformation. 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)

GDT_TS

The GDT_TS measure to compare two different structures.

$$GDT\_TS = (GDT\_P1 + GDT\_P2 + GDT\_P4 + GDT\_P8)/4$$

where GDT_Pn denotes percent of residues under distance cutoff $\leq$ n Å. Only Cα atoms are considered for the alignment and calculation. GDT_TS can detect the weak similarities in contrast to the RMSD measure.

Z-score

The Z-score measure of an observation is the distance from the mean measured in terms of standard deviation. In case of a protein ensemble, it is distance of the native structure from the average of the energy of the ensemble. Lower the Z-score, the better is the discrimination between the native and non-native structures.

$$Z\text{-}score = \frac{E_{native} - < E >}{\sigma_E} \tag{9.1}$$

where E_{native} is the energy of native structure, $< E >$ average energy and σ_E, standard deviation of the ensemble.

9.3 Abbreviations

Aβ	Amyloidβ peptide
APP	Amyloid precursor protein
bRMSD	Backbone root mean square deviation
BHT	Basin hopping technique
CASP	Critical assessment of structure prediction
CM	Comparative modeling
DSSP	Dictionary of secondary structure of proteins
ES	Evolutionary strategy
FM	Free modeling
gBHT	Greedy basin hopping technique
HA_TBM	High accuracy template based modeling
GDT_TS	Global distance test total score
MC	Monte Carlo
MD	Molecular dynamics
PDB	Protein data bank
POEM	Protein optimization with energy methods
PES	Potential energy surface
PSP	Protein structure prediction
PT	Parallel tempering
QA	Quality assessment
REMD	Replica exchange molecular dynamics
ROP	Repressor of primer
TBM	Template based modeling
WT	Wild type

Bibliography

A. Abagyan and M. Totrov. Ab Initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. *J. Comput. Phys.*, 402-412:151, 1999.

S. A. Adcock and J. A. McCammon. Molecular Dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106:1589–1615, 2007.

M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, 1987.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–10, 1990.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389–3402, 1997.

C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 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.

D. Baker and A. Sali. Protein Structure Prediction and Structural Genomics. *Science*, 294:93–96, 2001.

J. J. Balbach, Y. Ishii, O. N. Antzutkin, R. D. Leapman, N. W. Rizzo, F. Dyda, J. Reed, and R. Tycko. Amyloid fibril formation by A beta 16-22, a seven-residue fragment of the alzheimer's beta-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry*, 39:13748–13759, 2000.

J. M. Berg, J. L. Tymoczky, and L. Stryer. *Biochemistry, fifth edition*. W. H. Freeman and Company, 2002.

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 Struct. Funct. Bioinf.*, 45:119–126, 2001.

Richard Bonneau and David Baker. AB INITIO PROTEIN STRUCTURE PREDICTION: Progress and Prospects. *Annu. Rev. Biophys.*, 30:173–89, 2001.

Philip Bradley, Kira M. S. Misura, and David Baker. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science*, 309(5742):1868–1871, 2005.

J. Brange, L. Andersen, E. D. Laursen, G. Meyn, and E. Rasmussen. Toward understanding insulin fibrillation. *J. Pharm. Sci.*, 86:517–525, 1997.

C. L. Brooks, J. N. Onuchic, and D. J. Wales. Taking a Walk on a Landscape. *Science*, 293:612–613, 2001.

S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, W. Studier, and S. Swaminathan. Structural genomics: beyond the human genome project. *Nat. Genet.*, 23:151–157, 1999.

CASP5. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins Struct. Funct. Bioinf.*, 53:334–339, 2003.

CASP6. Critical assessment of methods of protein structure prediction (CASP)-Round 6. *Proteins Struct. Funct. Bioinf.*, 58:3–7, 2005.

M. Cecchini, R. Curcio, M. Pappalardo, R. Melki, and A. Caflisch. A Molecular Dynamics Approach to the Structural Characterization of Amyloid Aggregation. *J. Mol. Biol.*, 357:1306–1321, 2006.

M. S. Cheung, A. E. Garcia, and J. N. Onuchic. Protein folding mediated by solvation: Water explusion and formation of the hydrophobic core occur after the structure collapse. *PNAS*, 99:685–690, 2002.

D. Chivian, D. E. Kim, L. Malmstrom, J. Schonbrun, C. A. Rohl, and D. Baker. Prediction of CASP6 structures using automated robetta protocols. *Proteins Struct. Funct. Bioinf.*, 61:157–166, 2005.

R. Czerminski and R. Elber. Computational studies of ligand diffusion in globins: I. Leghemoglobin. *Proteins Struct. Funct. Bioinf.*, 10:70, 1991.

K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7155–8133, 1990.

K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.

R. I. Dima and D. Thirumalai. Exploring protein aggregation and self-propogation using lattice models: Phase diagram and kinetics. *Protein Sci.*, 11:1036–1049, 2002.

C. M. Dobson. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B*, 356:133–145, 2001.

C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

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.

R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6:1661–1681, 1997.

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

G. Favrin, A. Irbaeck, and S. Mohanty. Oligomerization of amyloid $A\beta_{16-22}$ Peptides Using Hydrogen Bonds and Hydrophobicity Forces. *Biophys. J.*, 87:3657–3664, 2004.

Alan R. Fersht and Valerie Daggett. Protein Folding and Unfolding at Atomic Resolution. *Cell*, 108: 573–582, 2002.

D. Frenkel and B. Smit. *Understanding Molecular Simulations : From Algorithms to Applications*. Academic Press, 2001.

K. Ginalski, A. Elofsson, D. Fischer, and L. Rychlewski. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, 19(8):1015–1018, 2003.

Nobuhiro Go. Theoretical Studies of Protein Folding. *Annu. Rev. Biophys*, 12(1):183–210, 1983.

Srinivasa M. Gopal and Wolfgang 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(46):7726–7728, 2006.

J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J. Mol. Biol.*, 331:281–299, 2003.

Thomas A. Halgren and Wolfgang Damm. Polarizable force fields. *Curr. Opin. Struct. Biol*, 11: 236–242, 2001.

K. Hamacher and W. Wenzel. Stochastic Tunneling Approach for Global Minimization of Complex Potential Energy Landscapes.

F. U. Hartl and M. H. Hartl. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded protein. *Science*, 295:1852–1858, 2002.

T. Head-Gordan and S. Brown. Minimalist models for protein folding and design. *Curr. Opin. Struct. Biol.*, 13:160–167, 2003.

T. Herges and W. Wenzel. An All-Atom Force Field for Tertiary Structure Prediction of Helical proteins. *Biophys. J.*, 87(5):3100–3109, 2004.

D. Hong, A. Ahamad, and A. L. Fink. Fibrillation of Human Insulin A and B Chains. *Biochemistry*, 45:9342–9353, 2006.

H. Jang, C. K. Hall, and Y. Zhou. Assembly and Kinetic Folding Pathways of a Tetrameric β-Sheet Complex: Molecular Dynamics Simulations on Simplified Off-Lattice Protein Models. *Biophys. J.*, 86:31–49, 2004.

S. Jang and S. Shin. Amyloid β -Peptide Oligomerization in Silico: Dimer and Trimer. *J. Phys. Chem. B*, 110:1955–1958, 2006.

G Jayachandran, V. Vishal, and V. Pande. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys*, 124: 164902, 2007.

F. Jeanmougin, J. D. Thompson, M. Gouy, D. G. Gibson, and T. J. Higgins. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.*, 23:403–405, 1998.

D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

W. L. Jorgensen, D. S. Maxwell, and J. J. Tiradorives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79 (2):926–935, 1983.

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

K. Karplus, R. Karchin, C. Barrett, S. Tu, M. Cline, M. Diekhans, L. Grate, J. Casper, and R. Hughey. What is the value added by human intervention in protein structure prediction? *Proteins Struct. Funct. Bioinf.*, S5:86–91, 2001.

D. K. Klimov and D. Thirumalai. Dissecting the assemble of $a\beta_{16-22}$ Amyloid Peptides into Antiparallel β Sheets. *Structure*, 11:295–307, 2003.

Andriy Kryshtafovych, Ceslovas Venclovas, Krzysztof Fidelis, and John Moult. Progress over the first decade of CASP experiments. *Proteins Struct. Funct. Bioinf.*, 61(S7):225–236, 2005.

B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302:1364–1368, 2003.

J. H. Laity, H. J. Dyson, and P. E. Wright. DNA-induced α-helix capping in conserved linker sequences is a determinant of binding affinity in $Cys_2 - His_2$ zinc fingers. *J. Mol. Biol.*, 295:719–727, 2000.

J. H. Laity, B. M. Lee, and P. E. Wright. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol*, 11:39–46, 2001.

B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55(3):379–400, 1971.

MS Lee, GP Gippert, KV Soman, DA Case, and PE Wright. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science*, 245(4918):635–637, 1989.

Cyrus Levinthal. Are there pathways for protein folding ? *J. Chim. Phys.*, 65:44–45, 1968.

A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997a.

A.. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations .II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J. Comput. Chem.*, 18(7):874–887, 1997b.

A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga. United-residue force field for off-lattice protein-structure simulations: Iii. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J. Comput. Chem*, 19(3):259–276, 1998.

T. Lührs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert, and R. Riek. 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *PNAS*, 102:17342–17347, 2005.

A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

Alexander D. Mackerell. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.*, 25:1584–1604, 2004.

O. S. Makin and L. C. Serpell. Structures for amyloid fibrils. *FEBS J.*, 272:5950–5961, 2005.

Marc A. Marti-Renom, Ashley C. Stuart, Andras Fiser, Roberto Sanchez, Francisco Melo, and Andrej Sali. Comparative protein structure modeling of genes and genomics. *Annu. Rev. Biophys.*, 29(1): 291–325, 2000.

J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

I. K. McDonald and J. M. Thornton. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.*, 238:777–793, 1994.

J. Meiler, M. Mueller, A. Zeidler, and F. Schmaeschke. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.*, 7: 360–369, 2001.

J. H. Meinke and U. H. E. Hansmann. Aggregation of $\beta - amyloid$ fragments. *J. Chem. Phys.*, 126: 014706, 2007.

M. Mezei. Chameleon sequences in the PDB. *Protein Eng.*, 11:411–414, 1998.

John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol*, 15(3):285–289, 2005.

A. Nagadoi, K. Nakazawa, H. Uda, K. Okuno, T. Maekawa, S. Ishii, and Y. Nishimura. Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. *J. Mol. Biol.*, 287:593–607, 1999.

M. Nanias, C. Czaplewski, and H.A. Scheraga. Replica Exchange and Multicanonical Algorithms with the Coarse-Grained United-Residue (unres) Force Field. *J. Chem. Theory Comput.*, 2(3): 513–528, 2006.

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. Comput. Chem.*, 12(5):594–605, 1991.

E. Neria and M. Karplus. Molecular dynamics of an enzyme reaction: proton transfer in TIM. *Chem. Phys. Lett.*, 267:23, 1997.

S. Oldziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kazmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, and H. A. Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *PNAS*, 102(21):7547–7552, 2005.

J. N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. THEORY OF PROTEIN FOLDING.

M. Orozco and F. J. Luque. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.*, 100:4187, 2000.

NP Pavletich and CO Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007):809–817, 1991.

D.A. Pearlman, D.A. Case, J.W. Caldwell, W.R. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.*, 91:1–41, 1995.

W. R. Pearson. Emperical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, 276: 71–84, 1998.

G.N. Ramachandran and V. Sasiskharan. CONFORMATION OF POLYPEPTIDES AND PROTEINS.

Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande. Simulations of the role of water in the protein-folding mechanism. *PNAS*, 101:6456–6461, 2004.

H. Roder and W. Colon. Kinetic role of early intermediates in protein folding. *Curr. Opin. Struct. Biol.*, 7:15–28, 1997.

C. A. Rohl and D. Baker. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc.*, 124:2723–2729, 2002.

C. A. Rohl, C. E. Strauss, D. Chivian, and D. Baker. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins Struct. Funct. Bioinf.*, 55:656–677, 2004a.

C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein Structure Prediction Using Rosetta. *Methods Enzymol.*, 383:66–93, 2004b.

B. Rost. Twilight zone of protein sequence alignments. *Protein Eng.*, 12:85–94, 1999.

B. Roux and K. Schulten. Computational Studies of Membrane Channels. *Structure*, 12:1343–1351, 2004.

Leszek Rychlewski and Daniel Fischer. LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.*, 14(1):240–245, 2005.

C. Sagui and T. A. Daren. MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULES: Long-Range Electrostatic Effects. *Annu. Rev. Biophys.*, 28:155–179, 1999.

A. Sali. 100,000 protein structures for the biologist. *Nat. Struct. Biol.*, 5:1029–1032, 1998.

A. Schug and W. Wenzel. Predictive in Silico All-Atom Folding of a Four-Helix Protein with a Free-Energy Model.

A. Schug, T. Herges, and W. Wenzel. Reproducible Protein Folding with the Stochastic Tunneling Method. *Phys. Rev. Lett.*, 91:158102, 2003.

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

A. Schug, B. Fischer, A. Verma, H. Merlitz, W. Wenzel, and G. Schoen. Biomolecular Structure Prediction Stochastic Optimization Methods. *Adv. Eng. Mater.*, 7(11):1005–1009, 2005a.

A. Schug, T. Herges, A. Verma, and W. Wenzel. Investigation of the parallel tempering method for protein folding. *J. Phys. Condens. Matter*, 17:1641–1650, 2005b.

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. A*, 103:3596–3607, 1999.

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.

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

K. T. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins Struct. Funct. Bioinf.*, 34:82–95, 1999.

M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials from crstal cata. 6. Determination of emperical potentials for O-H$\cdots$O=C hydrogen bonds from packing configurations. *J. Phys. Chem.*, 88:6231–6233, 1984.

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.*, 124:14548–14549, 2002.

Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

M. Sunde and C. C. F. Blake. The Structure of Amyloid Fibrils by Electron Microscopy and X-ray Diffraction. *Adv. Protein Chem.*, 50:123–159, 1997.

P. J. Thomas, B. Qu, and P. L. Pederson. Defective protein folding as a basis of human disease. *Trends Biochem. Sci.*, 20:456–459, 1995.

L. O. Tjernberg, J. Naslund, F. Lindqvist, J. Johansson, A. R. Karlstrom, J. Thyberg, L. Terenius, and C. Nordstedt. Arrest of β-amyloid Fibril Formation by a Pentapeptide Ligand. *J. Biol. Chem.*, 271: 8545–8548, 1996.

A. E. Torda. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.*, 7:200–205, 1997.

Jerry Tsai, Richard Bonneau, Alexandre V. Morozov, Brian Kuhlman, Carol A. Rohl, and David Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins Struct. Funct. Bioinf.*, 53:76–87, 2003.

Fyodor D. Urnov, Jeffrey C. Miller, Ya-Li Lee, Christian M. Beausejour, Jeremy M. Rock, Sheldon Augustus, Andrew C. Jamieson, Matthew H. Porteus, Philip D. Gregory, and Michael C. Holmes. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, 435(7042):646–651, 2005. ISSN 0028-0836.

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

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

Abhinav Verma. *Development and Application of a Free Energy Force Field for All Atom Protein Folding*. PhD thesis, Universität Dortmund, 2007.

D. J. Wales and Jonathan P. K. Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing upto 110 Atoms. *J. Phys. Chem.*, 101:5111–5116, 1997.

G. Wei, N. Mousseau, and P. Derreumaux. Computational Simulations of the Early Steps of Protein Aggregation. *Prion*, 1:3–8, 2007.

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

Scot A. Wolfe, Lena Nekludova, and Carl O. Pabo. DNA RECOGNITION BY BY CYS2HIS2 ZINC FINGER PROTEINS. *Annu. Rev. Biophys.*, 29(1):183–212, 2000.

B. Zagrovic, E. J. Sorin, and V. Pande. β-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 313:151–169, 2001.

Publications

Journals

- Srinivasa 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 2006*, **45**, 7726-7728.
- A. Verma, Srinivasa M. Gopal, J. Oh, K. H. Lee and W. Wenzel All atom de- novo folding of a 40 amino acid three-helix bundle protein with a scalable evolutionary algorithm. *J. Comp. Chem. 2007*, **28**, 2552-2558.
- Srinivasa M. Gopal and W. Wenzel All atom folding studies of a DNA bindin g protein in a free energy forcefield. *J.Physics: Cond. Matter 2007*, **19**, 285210.

Proceedings

- A. Verma, Srinivasa M. Gopal, K. H. Lee, E. Starikov, Wolfgang Wenzel De-novo all atom folding of helical proteins. *NIC Series 2006*, **34**, 45-52.
- Srinivasa M. Gopal, Konstantin V. Klenin and W. Wenzel Aggregation of the Amyloid-β protein: Monte Carlo optimization study. *NIC Series 2007*, **36**, 177-180.
- A. Verma, Srinivasa M. Gopal, A. Schug, J. S. Oh, Konstantin V. Klenin, K. H. Lee, W. Wenzel Massively parallel all atom protein folding in a single day. *Advances in Parallel Computing 2008*, **15**, 527-534.

Acknowledgment

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, he guided me with my thesis on a highly interdisciplinary field of research.

I am indebted to my colleagues for providing a stimulating and fun environment to learn. I am especially grateful to Dr. Konstantin Klenin, Timo Strunk, Dr. Abhinav Verma, Dr. Alexander Schug, Dr. Bernhard Fischer, Dr. Holger Merlitz and Dr. Aina Quintilla who were always helpful and congenial. I would like to thank Institute of Nanotechnology for providing an excellent working environment during my stay. Additionally I would thank my colleagues and teachers at The National College Bengaluru, Indian Institute of Technology Madras, Institute of Mathematical Sciences for providing me motivation for scientific research.

I would also like to thank my friends Dr. Nagabhushana, Smt. Padmini, Dr. Rajagopala, Dr. Suryakant Gupta and Dr. Nirmal Thyagu who have been helping me get through the difficult times, and for all the emotional support and caring they provided.

I am grateful to the Christine Batsch and Erika Schütze for assisting me in various different tasks during my stay at INT. I am also grateful to 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.

I wish to thank my entire family for providing a loving environment for me.

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