Protein Structure Prediction using Coarse Grain Force Fields

Dissertation
zur Erlangung des Doktorgrades
des Department Chemie der
Universität Hamburg

vorgelegt von

Nasir Mahmood

geboren am 18.01.1978 in Multan
Pakistan

Hamburg, den 21. Dezember 2009

1. Supervisor: Prof. Dr. Andrew Torda
2. Supervisor: Prof. Dr. Ulrich Hahn

Date of Disputation: February 12, 2010

Abstract

Protein structure prediction is one of the classic problems from computational chemistry. Experimental methods are the most accurate in protein structure determination but they are expensive and slow. That makes computational methods (i.e. comparative modeling and ab initio or de novo modeling) significant. In ab initio methods, one tries to build three-dimensional protein models from scratch rather than modeling them onto known structures. There are two aspects to this problem: 1) the score or quasi-energy function and 2) the search method. Our interest has been the development of quasi-energy functions. These could be seen as low-resolution special purpose coarse-grain or mesoscopic force fields, but they are rather different to most approaches. There is no strict physical model and no assumption of Boltzmann statistics. Instead, there is a mixture of Bayesian probabilities based on normal and discrete distributions. This has an interesting consequence. If one works with a method such as Monte Carlo, one can base the acceptance criterion directly on the ratio of calculated probabilities.

Under a Bayesian framework, the probabilistic descriptions of the most probable set of classes were found by the classification of 1.5×10^6 *protein fragments, each* $k \leq 7$ *residues long. These fragments were extracted from the known protein structures (with sequence identity less than 50% to each other) in the Protein Data Bank (PDB). Sequence, structure (*ϕ, ψ *dihedral angles of the backbone) and solvation features of the fragments were modeled by multi-way discrete, bivariate normal, and simple normal distributions, respectively. An expectation minimization (EM) algorithm was used to find the most probable set of parameters for a given set of classes and the most probable set of classes in the fragment data irrespective of parameters. With the obtained classification, one can calculated the probability of a protein conformation as a product of the sums of probabilities of its constituent fragments across all classes. The ratio of these probabilities then allows us to replace the ratio which is derived from the Boltzmann statistics in traditional Metropolis Monte Carlo methods. The search method, simulated annealing Monte Carlo, makes three kinds of moves (i.e. biased, unbiased, and 'controlled') to explore the conformational space. It has an artificial scheme to control the smoothness of the distributions.*

In initial results, the score function with sequence and structure terms only could produce protein-like models of the target sequences. Interestingly, these rather less compact models had good predictions of secondary structures. Incorporation of solvation term into the score function led to the generation of comparatively compact and sometimes native-like models, particularly for small targets. Models for relatively large and hard targets could also be generated with close secondary structure predictions. Secondary structures, particularly beta sheets, in these models often failed to properly pack themselves in the overall globular conformations. An ad hoc hydrogen bonding term based on an electrostatic model was introduced to entertain the long-range interactions. It could not make much difference probably due to its inconsistency with the score function.

Dedication

To my mother

Acknowledgments

I deem utmost pleasure to express my gratitude to my supervisor Prof. Dr. Andrew Torda for providing me an opportunity to work on an important scientific problem. He has been very kind, friendly and supportive throughout my doctoral research and dissertation writing. This work would not have been possible without his continuous guidance and scientific advice.

I am grateful to my colleagues, Stefan Bienert for making Harry Potter series available, Thomas Margraf and Joern Lenz for all their best wishes during writing, Paul Reuter for his greetings half an hour before lunch, and Tina Stehr (who was used to sit in front of me until she had not become mother) for helping me to get familiar with Wurst code, and Gundolf Schenk who has always been available for all kinds of favors from the proof-reading chapters of this dissertation to the discussions on a wide range of topics. It would be unfair to not mention Marco Matthies, Tim Wiegels, Martin Mosisch, and Jens Kleesiek with whom I have had dinners and discussions at various occasions.

I would also like to thank Annette Schade, secretary of Prof. Andrew Torda, for providing assistance in numerous ways particularly by quick preparation of my letters for German bureaucracy.

Finally, I am indebted to my wife, Mussarat Abbas, for her endless patience, understanding, and support in bad and good times. I do not have words at my command to express my deepest gratitude to my parents, brother and sisters for their spiritual and moral encouragement when it was most needed.

Nasir Mahmood

Contents

Chapter 1

Introduction

Protein Structure

Proteins are one of the most abundant molecules in all organisms and perform a diverse set of functions. They play important role for regulation of metabolic activity, catalysis of biochemical reactions and maintenance of cell membranes and walls for structural integrity of an organism (Hansmann 2003, Baker 2004). All proteins are hetero-polymer chains built from 20 types of amino acid by linking them through peptide bonds. Amino acid residues are characterized by their central alpha carbon (C_α) atoms to which side chains (R-groups) are attached. The type of amino acid at each position in a chain is determined by the genetic material of a cell. A single amino acid change can alter shape and function of protein (Anfinsen et al. 1954).

A strategy to predict three-dimensional structure of a protein from its sequence intends to establish a logical connection between chemical and geometric features of the main chain and those of its associated side chains (known from the existing protein structures) to its possible native structure. Geometric features of the peptide bond and the alpha carbon C_α (in figure 1.1) show a small variation in bond lengths and bond angles. It provides us an opportunity to cease the values of bond lengths and bond angles at the ideal geometry of the peptide bond by reducing degrees of freedom to a great extent. Thus, the backbone chain of a protein conformation can be defined completely by dihedral or torsion angles (ϕ, ψ, ω) of its peptide units (shown in the second chapter's figure 2.1). A dihedral angle is built from four successive atoms and three bonds of the backbone chain (as illustrated in the second chapter's figure 2.7).

Due to partial-double-bond character of the peptide bond, in 'trans' conformations, the dihedral angle ω mostly stays around 180° with a little variation of 10° whereas in 'cis' conformations, it remains 0°. Occurrence of 'trans' conformations in the world of proteins is far more abundant than 'cis' transformations (DePristo et al. 2003). As a consequence, the dihedral angle ω can also be kept constant at some ideal value along with bond angles and bond lengths.
Therefore, dihedral angle pairs ϕ, ψ represent the only degrees of freedom to (re)define the structural features of a protein structure. Such scheme of conformation definition with just two dihedral angles ϕ, ψ may introduce slight errors in those structural fea-

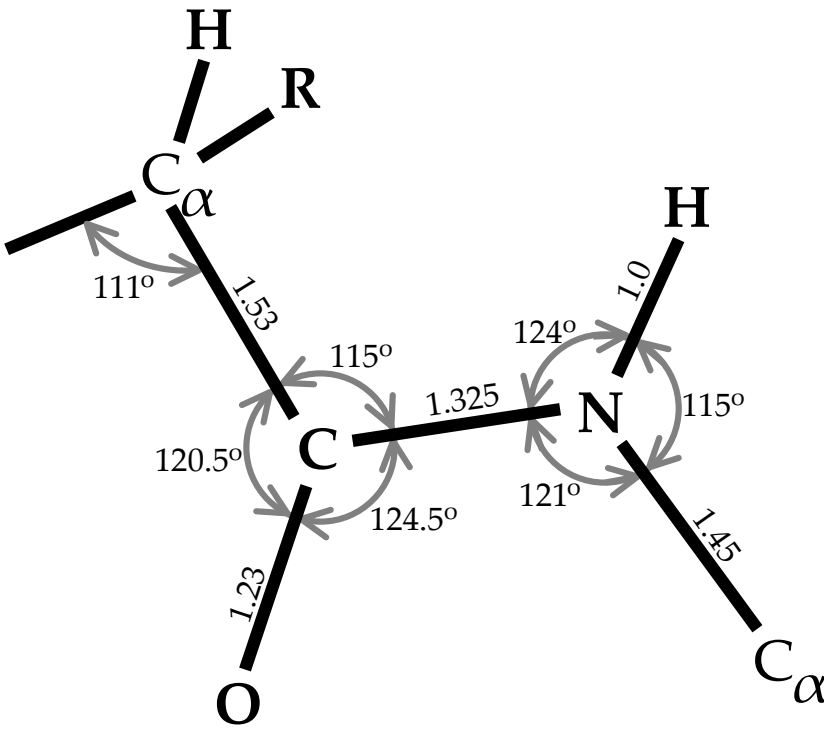


Figure 1.1: *Geometry of peptide bond.*

tures which require global definitions.

Most frequently observed secondary structures (i.e. helices and beta sheets) in protein structures are actually enforced or consolidated by physical hindrance caused by the steric properties of a protein backbone. The physical size of atoms or the groups of atoms in a protein backbone allow formation of a limited number of shapes without any clashes. In this regard, influence of weaker non-covalent interactions, called hydrogen bonds, is quite significant in stabilizing these shapes (i.e. secondary structures, helices and β sheets) and holding the entire structure together. Strength and effect of hydrogen bonds depend upon the environment. Backbone geometries of helices and β sheets facilitate in the establishment of systematic and extensible intramolecular hydrogen bonding. If these intramolecular hydrogen bonding patterns are not formed, the folding equilibrium would lead to unfolding by developing intermolecular hydrogen bonds with the surrounding water (Baldwin and Rose 1999, Petsko and Ringe 2004). Hydrogen bonds involve electrostatic attractions either between actual charges (Glu-Lys) or between peptide dipoles (N-H and C=O) to share a proton. Helices involve a repeated pattern of local hydrogen bonds between i and $i+3$ (in 3_{10} helix) or $i+4$ (in α helix) residues. In β sheets, these repeated patterns are between distant residues of the backbone (Fasman 1989).

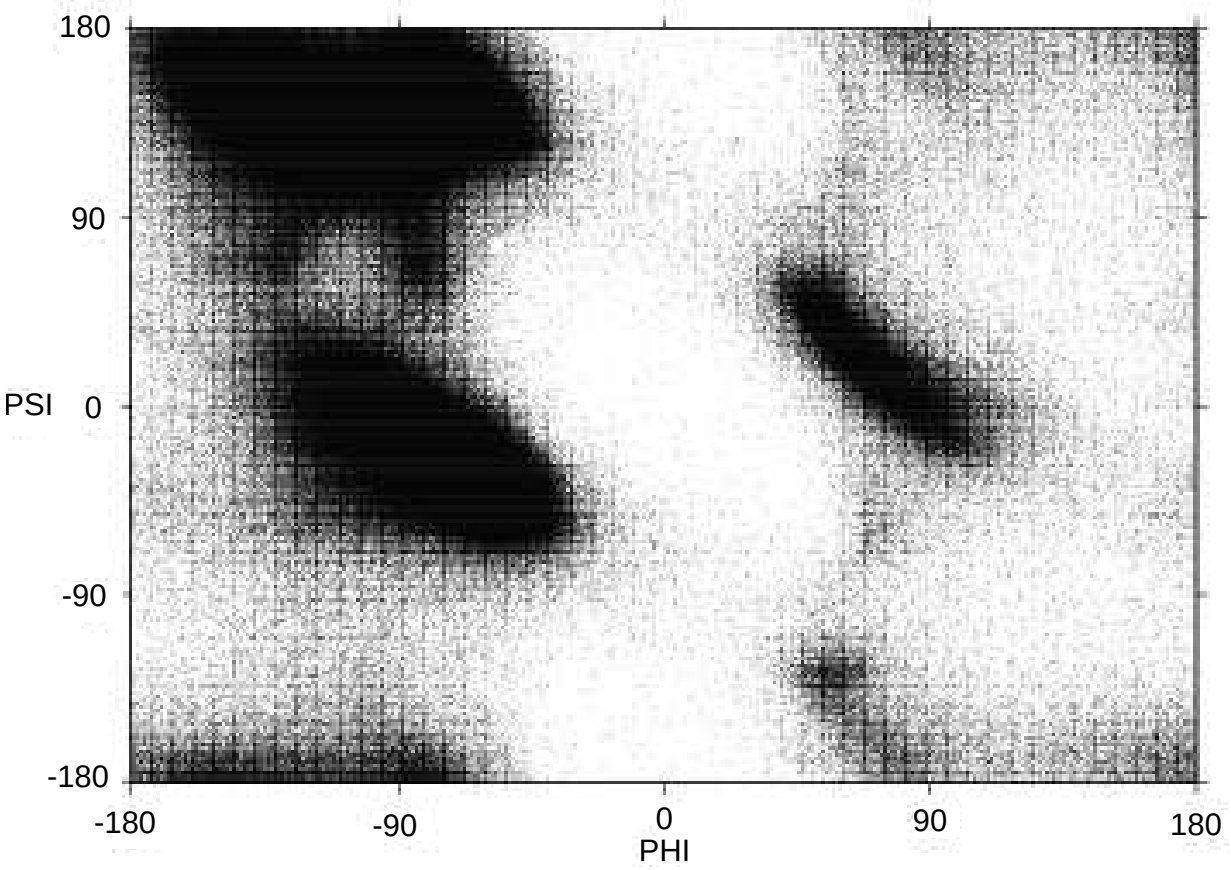


Figure 1.2: *Ramachandran plot - a survey of high resolution protein structures taken from the Protein Data Bank (PDB).*

Together, steric constraints and hydrogen bonding dictate dihedral angles (ϕ, ψ) to land in an even smaller space by reducing conformational space further. The allowed dihedral space can be viewed by plotting a ϕ versus ψ plot, called Ramachandran plot (Ramachandran et al. 1963), from dihedral angles (ϕ, ψ) of the known protein structures in the Protein Data Bank (PDB) (Berman et al. 2000). If we look at the Ramachandran plot (given in figure 1.2), there are two predominantly broader regions: the lower left represents right-handed α helices, and the upper left represents extended β or pleated sheets (Creamer et al. 1997). A smaller region on the upper right is of left-handed α helices. α helices are mostly right-handed with ϕ and ψ values around -60° and -40° respectively. Right-handed α helices are preferred over left-handed α helices because of two reasons: 1) cumulative effect of a moderate energy for each amino acid residue of a helix, and 2) no collision of C_β atoms with the following turn (Baldwin and Rose 1999). Left-hand conformations of helix are commonly observed in the isolated residues, for instance glycine, with ϕ and ψ values near 60° and 40° respectively. In the upper left region, ϕ and ψ values of extended β sheets remain around -120° and 140° respectively.

Protein Structure Determination

Three-dimensional protein structure determination involves a sequence of preparative and analytical steps from coding of DNA to optimization of 3D structures (as shown in figure 1.3). No matter which method is used, many of the preparative and analytical steps are common and have to be performed in sequence.

The main analytical techniques used for protein structure determination at atomic level are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Both techniques are briefly discussed in the following:

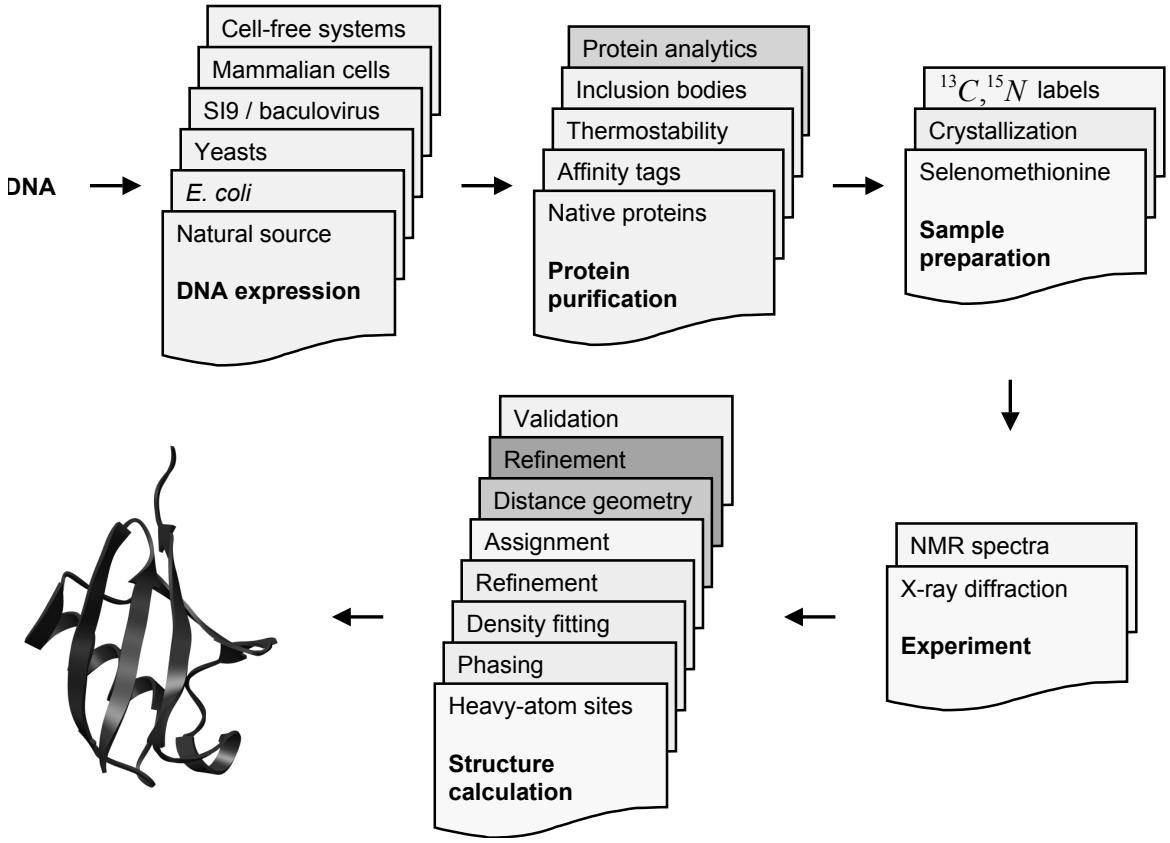


Figure 1.3: *Protein structure determination. Image taken from (Heinemann et al. 2001)*

X-ray Crystallography

The history of X-ray crystallography, the most favored and accurate technique, dates back to 1934 when Bernal and Crowfoot took the first X-ray photograph of a crystalline globular protein, pepsin (Bernal and Crowfoot 1934). After that revelation, it took about two and a half decades to make the progress for structure determination of a complete

protein (Kendrew 1959). Being a mature technique, protein X-ray crystallography has already been integrated into high-throughput technology.

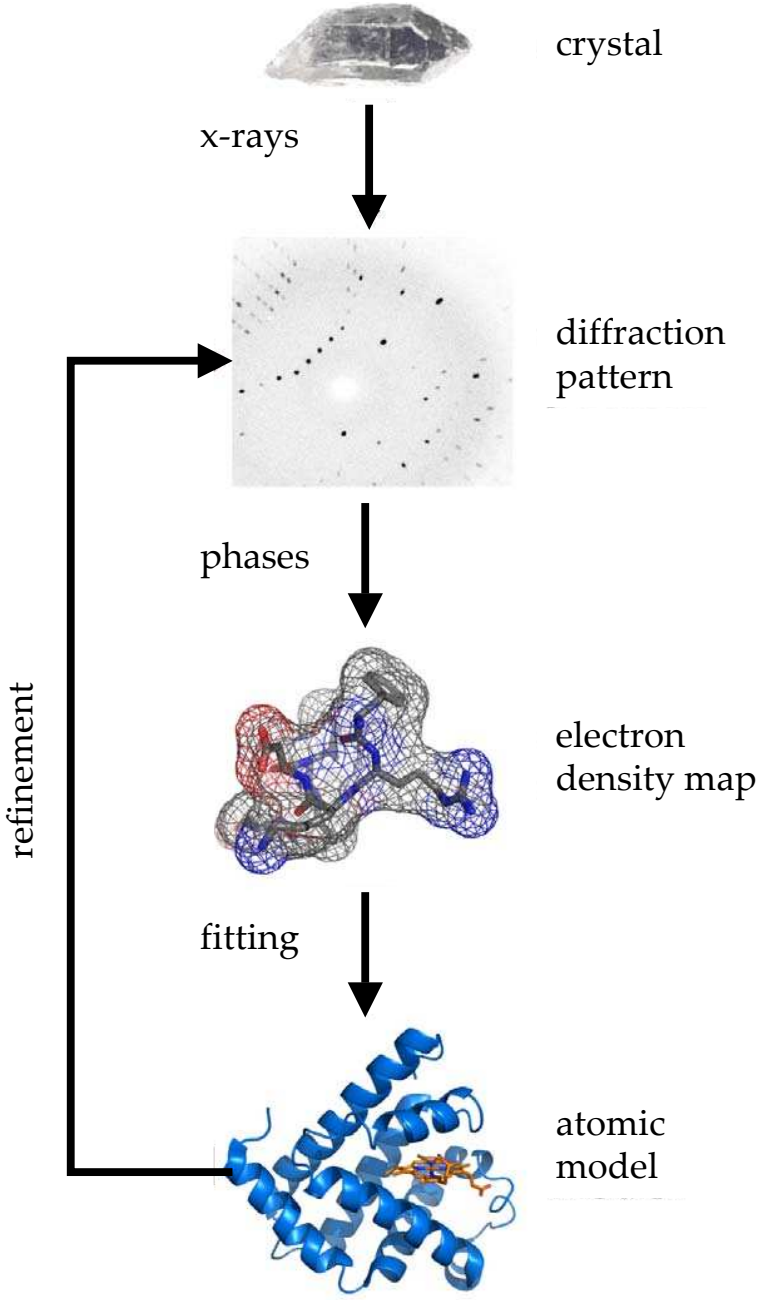


Figure 1.4: *Structure determination of a molecule by X-ray crystallography. Image taken from www.wikipedia.org*

The steps of isolation, purification, and crystallization of a purified sample at high concentration (as shown in figure 1.3) are performed to yield a protein crystal of sufficient quality. Protein crystal growth still remains the most time limiting and the least

well understood step in protein X-ray crystallography. Once protein crystal has been grown, then it is exposed to an X-ray beam to get three-dimensional molecular structure out of it. The exposure of crystal to an X-ray beam results into diffraction patterns. Diffraction patterns are then processed to yield information about the packing symmetry and the size of repeating units in that crystal. All the interesting information actually comes from the pattern of diffraction spots. The "structure factors" determined from the spots intensities of diffraction pattern are then used to construct electron density maps. Three-dimensional molecular structure can be built out of electron density maps by using protein sequence (Smyth and Martin 2000). Figure 1.4 shows the key steps of X-ray crystallography.

Nuclear Magnetic Resonance (NMR)

NMR, another experimental method for protein structure determination, was originally proposed by Harvard (Purcell et al. 1946) and Stanford (Bloch et al. 1946) in late 1945 and later in 1980s was extended by Wüthrich and Ernst (Wüthrich 1986, Oschkinat et al. 1988) for proteins and nucleic acids. NMR has an advantage over X-ray crystallography in that it is recorded in solution closer to physiological conditions (Wüthrich 2003).

A suitable protein sample is used to perform a set of multidimensional NMR experiments. These experiments generate NMR spectra which are then used to measure the resonance frequencies of NMR-active spins in a protein. Conformation-dependent resonance frequencies play a significant role in the derivation of constraints from NMR experiments (such as Nuclear Overhauser Effect (NOE), scalar coupling, and diploar coupling data). The derivation of structural constraints and the calculation of structures are done in an iterative manner and it stops when the majority of experimentally derived constraints verifies conformations representing the NMR structure. Variation in the conformation structures reflects the precision of NMR structure determination (Montelione et al. 2000).

The advantage of NMR is that it avoids the time limiting step of protein crystallization. On the other hand, the main disadvantages of NMR are the limitation of protein size (i.e. 150-200 amino acids) and requirement of relatively soluble proteins (Lesk 2004).

Protein structure determination with experimental methods is considered very reliable as it provides vital information about the general characteristics of protein structures. Thereby, the computational biologists can build their prediction methods by relying on or learning from the detailed information obtained from the experimentally solved structures (Dodson 2007, Smyth and Martin 2000).

Viewpoints to Protein Folding

About 50 years ago, it was proposed that the amino acid sequence of a protein contains all the necessary information needed to make it folded into a three-dimensional structure (Anfinsen et al. 1961, Sela et al. 1957). Nevertheless, the unearthing of fundamental principles which govern the folding process of the polypeptide chain of a protein into a compact three-dimensional structure is still a grand challenge in modern biology (Dobson 2003).

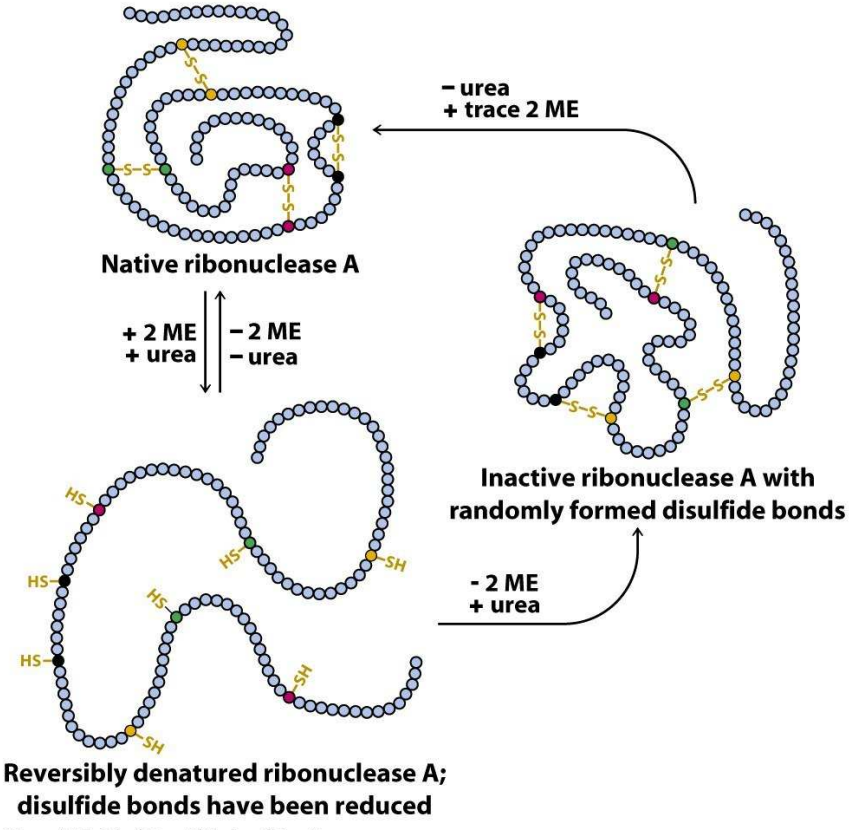


Figure 1.5: *The Anfinsen experiment in protein folding. Image taken from (Horton et al. 2006)*

Amazingly, a polypeptide chain of N amino acids takes a very short time of $\sim \exp(N^{2/3})$ nanoseconds to fold into its native structure by avoiding 2^N possible conformations. The value of N may range between 50 and 5000 amino acids (Beiner 2007, Finkelstein et al. 1996). In practice, however, a folding process based on some statistical principle (i.e. random, unbiased sampling of configurations) would take billions of years to find native state of a protein of just 100 amino acides (Finkelstein 1997, Hunter 2006, Zwanzig et al. 1992, Radford 2000). This apparent contradiction between finite time of folding and infinite possible conformations is called Levinthal's paradox

(Levinthal 1969).

Levinthal 's paradox suggests that there is some mechanism which simplifies the process of protein folding. Numerous points of view or models about that suspected mechanism of protein folding have been presented over the last few decades. In the following, few important ones are discussed briefly.

Anfinsen 's Hypothesis

According to Anfinsen's hypothesis, a protein in normal physiological milieu is a system with lowest Gibbs free energy and its native structure is determined by the totality of its inter-atomic interactions (i.e. amino acid sequence) under the given environment. A protein unfolds and loses its activity under the changed environmental conditions but it re-folds and becomes biologically active on return of proper environment (Anfinsen 1973).

Anfinsen made his point with ribonuclease A (RNaseA), an extracellular enzyme of 124 residues with four disulfide bonds, as his model. Disulfide bonds were reduced, but after the reductant was removed, protein refolded and regained its activity. This was interpreted as showing that the primary structure completely determines the tertiary structure (Horton et al. 2006). Figure 1.5 shows the discussed experiment.

Sequential Model

There are numerous hypotheses about hierarchical or sequential model of protein folding but the two most popular ones are: the framework model (Anfinsen 1972, Ptitsyn

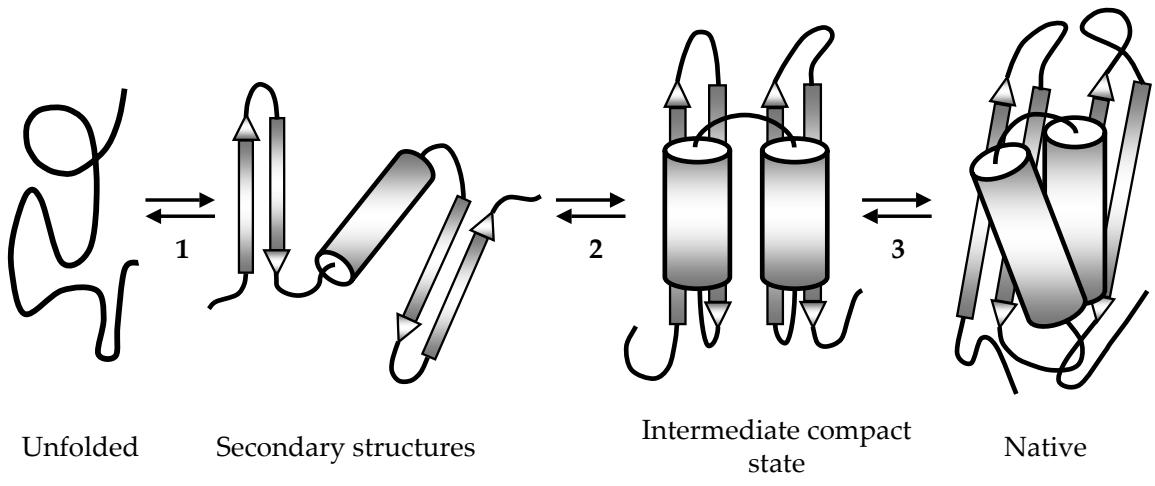


Figure 1.6: *Sequential mechanism of protein folding. Image adapted from (Ptitsyn 1987)*

et al. 1972) and the modular assembly model (Ptitsyn and Rashin 1973). According

to the framework model, protein folding starts with the formation of secondary structures in an unfolded chain. Whereas the modular assembly model assumes that the folding process starts with independent folding of separate parts of protein molecule. However, in both models, secondary structures have a central role in determining the folding pathway. Together these two models assume that the quasi-independent domains or sub-domains of a protein may fold at different times but each one's folding starts with the formation of secondary structures (Ptitsyn 1987).

According to the framework model, protein folding happens in three hypothetical steps : 1) the formation of secondary structures by local interactions at the fluctuating regions of the unfolded polypeptide chain, 2) the collapse of secondary structures into a compact structure due to long-range interactions of side groups with the surrounding medium, and 3) the re-arrangement of already intermediately compact structures into a unique native structure. Specific long-range interactions between spatially close amino acid residues play a key in this re-arrangement (Ptitsyn and Rashin 1975). See figure 1.6.

The important feature of the framework model is the stabilization of folding states by three different kinds of interactions. In the first step, secondary structures are stabilized by hydrogen bonds, in the second step, intermediately compact structure by hydrophobic interaction, and in the last and third step, native globular structure by van der Waals interactions (Ptitsyn 1987).

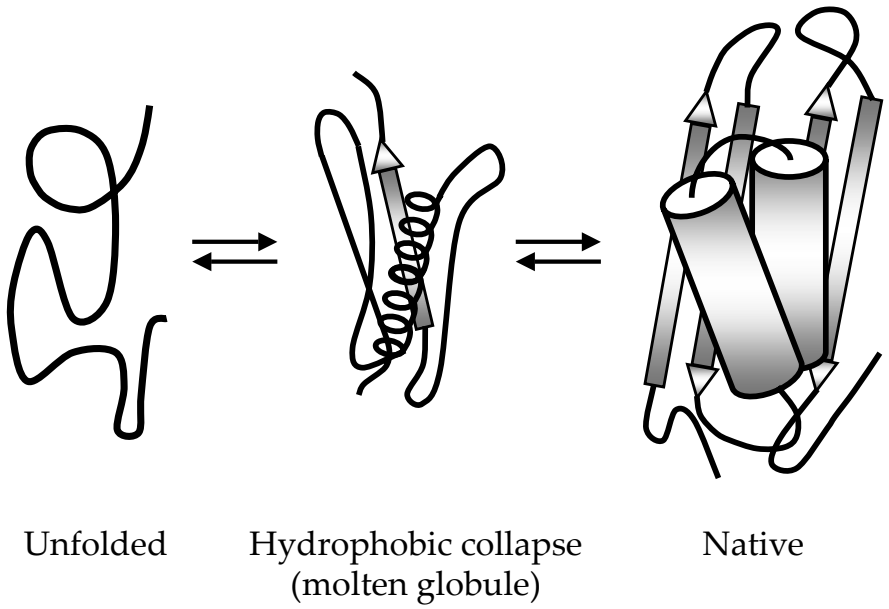


Figure 1.7: *Hydrophobic collapse. Image redefined from (Radford 2000, Ptitsyn 1987)*

Hydrophobic Collapse

According to hydrophobic collapse model, globular proteins have a 'hydrophobic core' (consisting of non-polar residues) at the interior of their native structures and most of the polar or charged residues are situated on the solvent exposed surface. 'Hydrophobic core' supposedly leads to structural collapse of the unfolded polypeptide chain of a protein. The squeezing of hydrophobic side chains from the surrounding solvent is thought to be a source of energetic stabilization of intermediately compact structures. The collapsed structure, also called molten globule, represents a partially folded state and its energy is lower than the unfolded state but higher than that of the native state.

Unlike sequential model, the event of hydrophobic collapse (step 1 of figure 1.7) happens first and then the formation of secondary structures and medium to long-range interactions occur in the folding pathway (Radford 2000, Schellman 1955).

Shakhnovich's Critical Residue Model

Shakhnovich's model proposes that the transition state of protein folding depends upon the formation of a specific subset of native structure called protein folding nucleus. Protein folding nucleus corresponds to the transition state between a structureless compact intermediate without unique structure and molten globule with elements of a native-like fold but not to the one between the native state and molten globule. Nucleus growth is a necessary condition for the subsequent fast folding to the native state. Search for protein folding nucleus takes a considerable time to overcome a major free energy barrier. See figure 1.8

To test this model, a number of designed sequences were generated and their random coils were folded with lattice Monte Carlo simulations. Polypeptide chains were observed to reach their native conformations through the formation and the growth of protein folding nuclei (Abkevich et al. 1994).

Protein folding nucleus is a localized sub-structure made of 8 to 40 native contacts scattered along a protein sequence. That shows, the nucleus contacts are both long-range as well as short-range (Shakhnovich 1997). Shakhnovich's model recognizes role of two kinds of residues: 1) critical residues which make most of the contacts in the transition state, and 2) those which only form contacts on reaching of the native state. The mutations of critical residues may affect the stability of the transition state and that of the folding kinetics. That is why, critical residues are supposed to be evolutionarily conserved (Shakhnovich et al. 1996). In the subsequent studies, it has also been revealed that the transition state contains specific interactions which are not found in the native state. These non-native contacts slow down folding process without affecting the stability of the native state. This decoupling of the folding kinetics from the thermodynamics

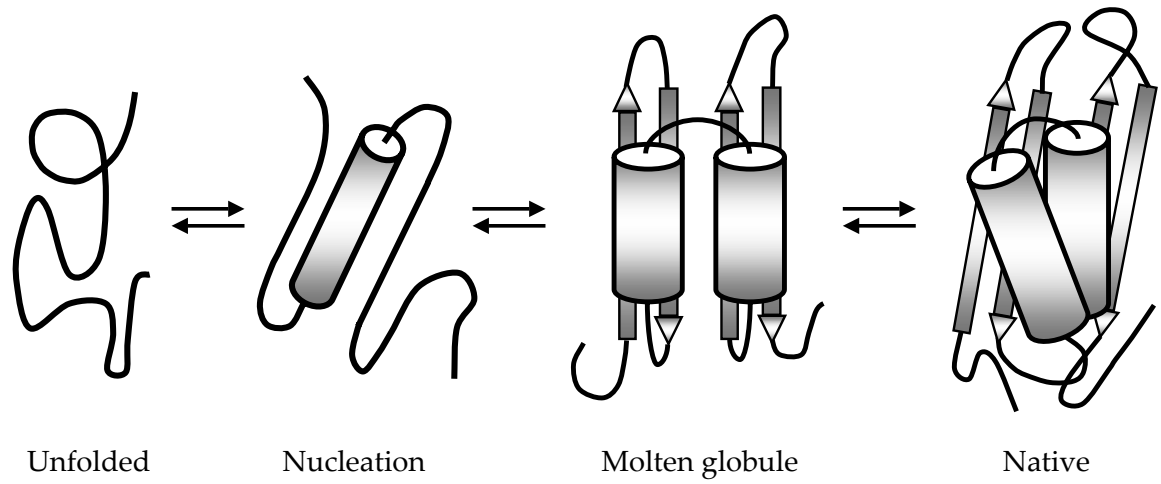


Figure 1.8: *Shakhnovich's model of critical residues. Image adapted from (Radford 2000)*

was observed in many proteins (Li et al. 2000).

Protein Structure Prediction

In the recent past, methodological advances in the field of DNA sequencing have led to a revolutionary growth of sequence databases (Moult 2008). The PDB has only 1% experimentally determined three-dimensional structures of the millions of known protein sequences and this percentage is falling fast with the sequencing of new genomes (Eramian et al. 2008). Figure 1.9 shows the current statistics of the PDB.

The growth of protein structure database is mainly restricted by two factors: 1) experimental methods for protein structure determination are slow because many of the steps (shown in figure 1.3) involved in experimental techniques have to be repeated tediously in a trial-and-error search of the optimum conditions (Heinemann et al. 2001), and 2) protein structure determination by experimental methods is quite expensive. On average, the amount spent on experimental structure determination of a protein ranges from $250,000 to $300,000 (Service 2005, Lattman 2004). Therefore, it would always be useful to build models for protein structures even if they were not perfect.

A number of initiatives, like structural genomics, have also been launched to fill the widening gap where experimentally determined structures are two orders of magnitude less than the known protein sequences (Chandonia and Brenner 2006, Burley 2000, Burley et al. 1999). In all such initiatives, computer-based structure prediction methods in combination with experimental structure determination methods have a significant role to provide a fast information about the structural and functional properties of proteins.

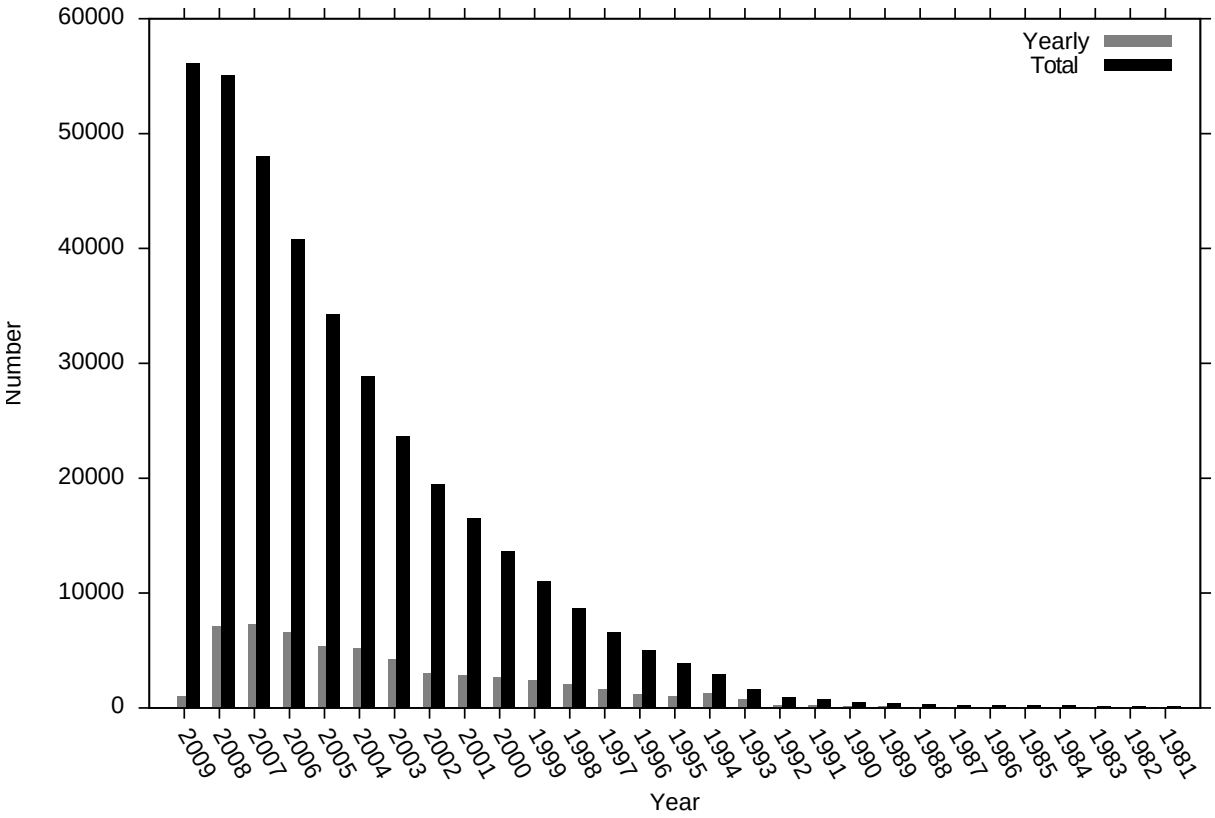


Figure 1.9: *Yearly growth of total structures in Protein Data Bank (PDB). Image adapted from (PDB 2009)*

Molecular and cell biologists are always waiting for the information generated by these initiatives (Zhang 2008).

The experimentally determined structures are always ideal and essential for some typical application, e.g. structure-based drug design. However, protein structure database cannot be filled rapidly with the missing structures due to the unavoidable experimental difficulties and the cost of structure determination (mentioned above). Three-dimensional models of proteins generated by computer-based prediction methods also have a broad range of useful applications.

Protein structure prediction methods produce models of different quality (Moult 2008). The low resolution models are usually used for the recognition of approximate domain boundaries (Tress et al. 2007) and the assignment of approximate functions (Nassal et al. 2007). The medium resolution models are important to approximate the likely sites of protein-protein interaction (Krasley et al. 2006), role of disease-associated substitutions (Ye et al. 2006), and the likely role of alternative splicing in protein function (Wang et al. 2005). The applications of high resolution models include molecular

replacement in X-ray crystallography (Qian et al. 2007), interpretation of disease mutations (Yue and Moult 2006) and identification of orthologous functional relationship (Murray and Honig 2002).

The existing protein structure prediction methods basically deal with two situations: 1) when the PDB has structures related to a target sequence, how to identify those structures (especially when the similarity between the target sequence and the PDB structures is very weak or distant) to build model on them, and 2) when no related structure is found in the PDB for a target sequence, how to build model for that target from scratch (Zhang 2008). The methods for these two kinds of structure predictions are categorized as: comparative protein modeling, and *ab initio* structure prediction.

Comparative Protein Modeling

Comparative modeling, also known as homology modeling, is a widely used and well established class of protein structure prediction methods (Kolinski and Gront 2007, Eramian et al. 2008). Comparative modeling methods take advantage of the structural information available through already experimentally solved protein structures. The known protein structures are used as templates to predict the structures of target sequences (Sanchez and Sali 1997). Both an efficient modeling algorithm and the presence of a diverse set of experimentally solved structures in the PDB are equally important factors for the generation of correct models (Zhang 2008).

A classical comparative modeling approach consists of four steps: 1) identifying the templates that are related to target sequence, 2) aligning target sequence with the templates, 3) building a model from the alignment of target sequence with the templates, and 4) assessing final model using different criteria (Fiser et al. 2002, Zhang 2008, Marti-Renom et al. 2000, Blundell et al. 1987, Greer 1981, Johnson et al. 1994, Sali and Blundell 1993, Sali 1995, Fiser and Sali 2003, Lushington 2008). Figure 1.10 shows a detailed flow chart of the steps often used by various comparative modeling methods. The accuracy of a method solely depends upon the correct identification of templates relevant to a target sequence because the wrong templates will generate a model full of errors. The correct identification of templates is impossible unless the alignment between the target sequence and the template is accurate (Zhang 2008). The existing comparative modeling methods can be put into two categories depending upon the alignment methods or score functions they use to find related templates (Ginalski et al. 2005).

Sequence Similarity Centric Methods

Sequence similarity between target sequence and template structure(s) is precondition for the success of any comparative modeling method. The pairwise comparison methods discussed here can basically detect sequence similarities higher than some length-

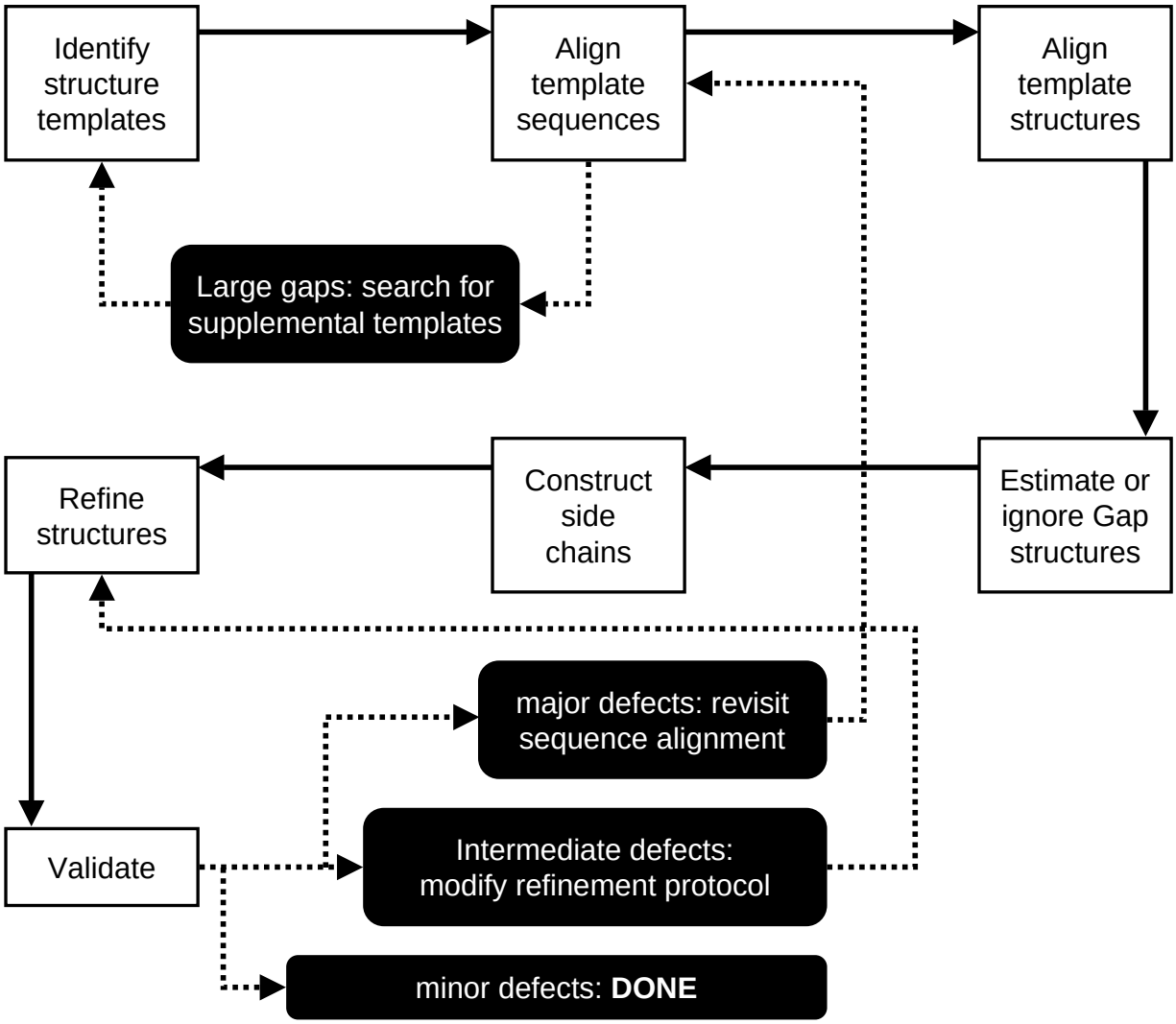


Figure 1.10: *Comparative modeling flow chart showing standard process (solid) and feedback/refinement mechanism (dashed). Diagram was taken from (Lushington 2008)*

dependent sequence similarity threshold (Rost 1999, Zhang 2008, Eramian et al. 2008, Sánchez et al. 2000, Deane and Blundell 2003).

Sequence-Sequence Comparison

Sequence-sequence comparison methods are simple and still popular to find homologous structures that are closely related to a target sequence. FASTA (Pearson and Lipman 1988) and BLAST (Altschul et al. 1990, Altschul et al. 1997) are the most widely used sequence-sequence comparison methods. In addition to a substitution matrix (Henikoff and Henikoff 1992), they require parameters for defining the penalty for gap initiation and extension in the alignments. The discrimination between scores for ho-

mologs and expected random scores is achieved by proper evaluation of the substitution matrices and alignment parameters.

Sequence-sequence comparison methods, like BLAST, are extremely fast because of an initial screening of the sequences which have a chance of providing a better alignment score. The main disadvantages of such methods are the equal treatment of variable and conserved positions and inability to detect distant or remote homologs.

Sequence-Profile Comparison

Profile-sequence or sequence-profile comparison methods involve position-specific substitution matrices (Bork and Gibson 1996) to preferably take the conserved motifs into account. A profile ($N\times$ 20 substitution matrix) is built from the variability in multiple sequence alignment of the target sequence with closely homologous structures. This profile carrying information about the family of homologs (than a single sequence) is then used to score the alignment of any of 20 amino acids to each of N residues of a protein. The additional computational cost of profile calculation does not make profile-sequence comparison slower than sequence-sequence comparison because the alignment score of two positions is calculated through a lookup in both profile and matrix.

PSI-BLAST (Altschul et al. 1997) is the most popular profile-sequence comparison method due to its ability to generate multiple sequence alignments and profiles iteratively. PSI-BLAST also makes use of an initial screening technique of potential hits to enhance its speed. PDB-BLAST (Jaroszewski et al. 1998) is a PSI-BLAST based method which involves two steps: 1) it builds a sequence profile from NCBI non-redundant database and other protein sources, and 2) the sequence profile is then used to search the PDB. RPS-BLAST (Marchler-Bauer et al. 2003) is another BLAST-based method for searching a query sequence against a database of profiles of conserved motifs.

Another related type of comparison methods uses hidden Markov models (HMMs) to describe the sequence variability in the homolog family by specifying the probability of occurrence of each of 20 amino acids at each position of a target or query sequence (Eddy 1998, Karplus et al. 1999, Soding 2005). HHM-based searches are slower due to the lack of initial screening of a database but more sensitive than those based on PSI-BLAST.

Profile-Profile Comparison

Instead of making a comparison between a query sequence and the template profile or the query profile and a template protein, a direct comparison between two profiles can be made by converting their positional vectors into a score matrix. Profile-profile comparison methods have been claimed to be more sensitive in detecting similarity between two families than profile-sequence and sequence-profile comparison methods

(Rychlewski et al. 2000). ORFeus, a profile comparison tool, has introduced a three-valued secondary structure information to the position-specific substitution scores (Ginalski et al. 2003). The calculation of secondary structure is entirely based on the sequence profiles. Addition of secondary structure information reportedly improves the sensitivity towards similarity between protein families.

Threading or Fold Recognition

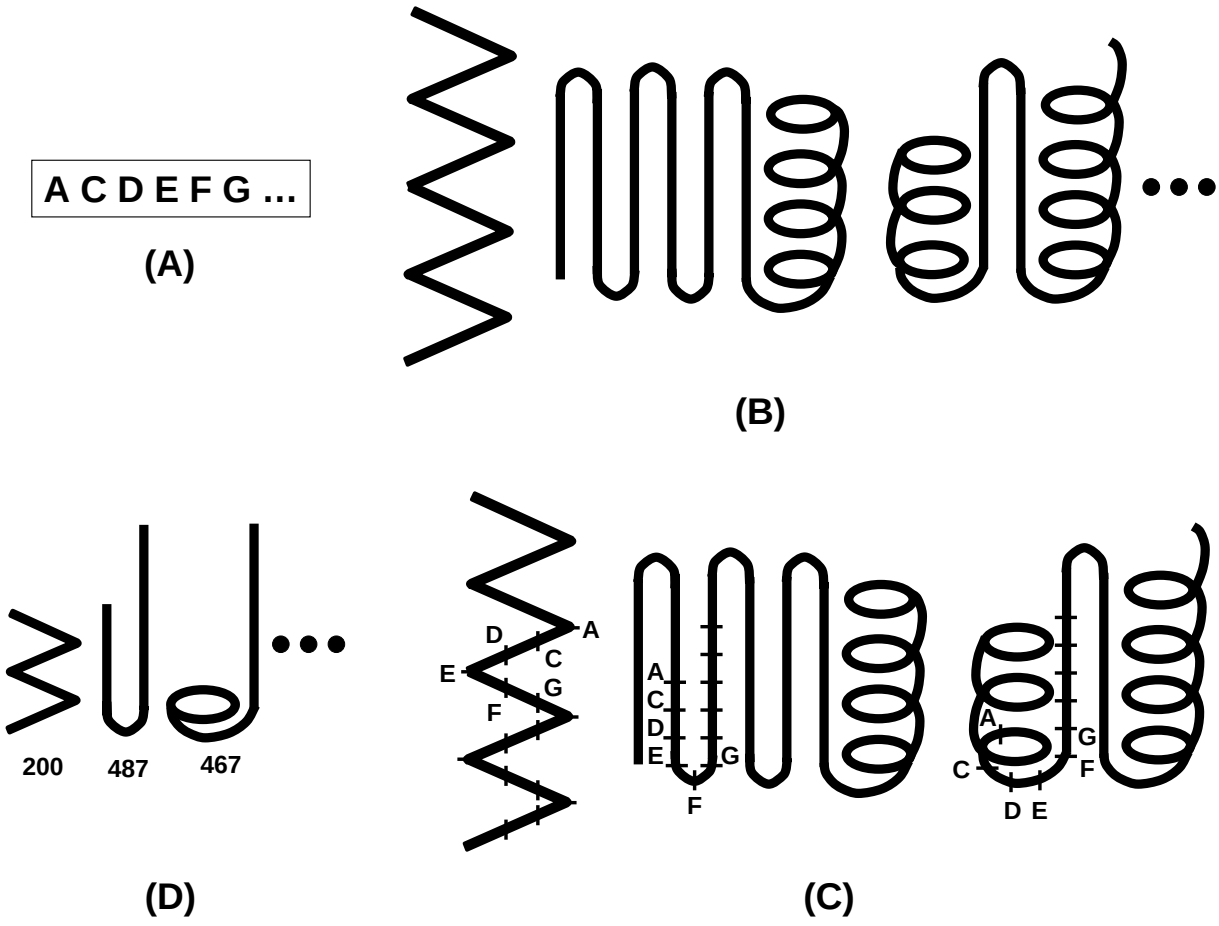


Figure 1.11: *Protein threading: A) a target sequence whose structure is not known yet and does not have detectable sequence similarity with the known structures, B) a template library consisting of a collection of the known structures, C) threading and alignment of the target sequence to the structures in the template library, D) a set of candidate structures for the target sequence. Image adapted from (Torda 2005).*

Threading or fold recognition methods basically employ sequence to structure alignment scoring techniques and are useful to identify the templates of those targets whose

sequences have no significant similarity to those of templates. The basic premise of threading methods is that there is are limited number of protein folds in nature (Hubbard 1997, Zaki and Bystroff 2007, Lemer et al. 1995, Chothia 1992, Orengo et al. 1994) and fold of a target sequence may be similar to that of a known structure either because of too remote undetectable evolutionary relationship or the two folds have converged to be similar (Moult 2005)

Successful recognition of related structure(s) through sequence to structure alignment ultimately results into a useful model. A typical threading method has three components (as shown in figure 1.11): 1) a comprehensive and representative library of templates, 2) an efficient and accurate sequence to structure alignment method, and 3) a score function for the ranking of final templates (Torda 2005).

In existing methods, the sequence to structure alignment is based on dynamic programming (Thiele et al. 1999, Taylor and Orengo 1989), Monte Carlo search methods (Bryant and Lawrence 1993, Bryant and Altschul 1995, Abagyan et al. 1994, Madej et al. 1995), genetic algorithms (Yadgari et al. 1998) or a branch and bound search (Lathrop and Smith 1996). A score function for sequence to structure alignment may be different from the one used in the template's ranking. Contemporary score functions are based on structural environment around (Bowie et al. 1991), statistical potentials based on pairwise interaction between residues (Sippl 1990, Sippl 1995, Jones et al. 1992), and secondary structures and solvation information (Shi et al. 2001). In hybrid methods, threading score functions have also been used in combination with multiple sequence alignments or sequence profiles (Panchenko et al. 2000, Zhou and Zhou 2004, Torda et al. 2004).

As sequence to structure alignment is an NP-complete multi-minima problem (Lathrop 1994), it usually takes hours of large computational resources to generate alignments of the target sequence with its templates in a reasonably large library (Yadgari et al. 1998, Ginalski et al. 2005). Apart from the computational requirements, the energy functions are also poor in dealing with the energetics of missing or dissimilar details of the alignments generated from weakly homologous templates. These energetic errors are more severe in the 'twilight' zone where sequence similarity is below 30%. Sometimes, minor errors in the alignment can spoil energy by seriously damaging the accuracy of models (Finkelstein 1997).

Recent benchmark studies have shown that only $\sim 50 - 65\%$ of proteins can be assigned correct templates by threading methods in the absence of a significant sequence similarity (Ginalski et al. 2005, Zhang 2008). Threading methods are also accused of not proposing novel folds because their predictions are entirely based on already known structures. That is why, *ab initio* structure prediction methods are important for prediction of both novel as well as weakly homologous folds (Kihara et al. 2002, Hardin et al. 2002, Skolnick et al. 2000).

Ab Initio Structure Prediction

Target sequences for which no relationship to the known structures of proteins can be detected are potentially addressed by *ab initio* or template-free methods. In *ab initio* structure prediction, the structure of a protein in question is not adopted from the known structure(s) of an evolutionarily related protein(s) but it is rather built from scratch. These methods are computationally more expensive than template-based methods.

As physical forces on atoms of a protein sequence push it to fold into a native-like conformation, the most natural and accurate starting point would be an all-atoms molecular dynamics simulation (Levitt and Sharon 1988) of protein folding or prediction problem by simulating both protein and the surrounding water (Skolnick and Kolinski 2001). Unfortunately, it is not a way forward due to two reasons: 1) molecular dynamics simulation with an all-atoms force fields and explicit representation of the surrounding water molecules is computationally very expensive, and 2) only inadequate potential functions are available for the current molecular dynamics simulations (Bonneau and Baker 2001).

A number of *ab initio* structure prediction methods have been proposed over the last decades. All methods search for a native-like conformation with lowest free energy and have three aspects in common: 1) a suitable representation of protein molecule, 2) an energy or score function, and 3) an efficient search method. In following sections, these three aspects have been discussed in detail.

Structure Representation Schemes

A protein with N atoms has $3N$ degrees of freedom and can be represented by $3N$ variables. Knowing the bond lengths and the bond angles are almost constant, one may think to have a dihedral or torsional representation in order to bring the space complexity threefold down (Xu et al. 2006).

As discussed earlier in section 1.1, the dihedral space can be further reduced by ignoring the dihedral angle ω (shown in the second chapter's figure 2.1) and assuming that the peptide bond is planar and remains almost constant. Thus, the main chain can be represented with only dihedral angle pairs (ϕ and ψ). The amino acid side chains may be represented by pseudo-atoms. Pseudo-atoms are either a weighted average of the real atoms or a combination of the most commonly observed side-chain angles called rotamers (Samudrala et al. 1999).

In the Cartesian space, a regular grid of tetrahedral or cubic cells (lattice) can be used to reduce the search space. Lattice methods are used for the simplification of conformations by digitizing them onto lattice. These methods use a very simple one or

two atoms representation of conformations to perform a fast search through the space. For smaller proteins, an exhaustive search on all possibly enumerated states can be performed but, in case of large proteins, special grid-based moves and optimization techniques are needed to search heuristically through all possible states.

Fragment assembly (Bowie and Eisenberg 1994), another conformational space reduction method and one of the most successful techniques in *ab initio* structure prediction techniques, is an extension of template-based methods but uses rather small fragments from multiple sources to build three-dimensional models. Fragment assembly does not transfer parent fold but the homology and the knowledge based information. Fragments carry information about the geometry (super-secondary, secondary, no clashes etc.) depending upon their size (Zhang and Skolnick 2004, Jones and McGuffin 2003, Simons et al. 1997). The assembly of fragments happens in two steps: 1) picking homologous fragments from a fragment library or dihedral angles randomly from continuous dihedral distributions, and 2) the minimization of conformations formed by the chosen fragments. With fragment assembly based methods, Monte Carlo simulated annealing (MCSA) has been used as a heuristic technique to find the energy minima.

The methods which involve fragment assembly to make their conformational search tractable include: Rosetta (Rohl et al. 2004), SimFold (Fujitsuka et al. 2006), PROFESY (Lee et al. 2004), FRAGFOLD (Jones 2001), UNDERTAKER (Karplus et al. 2003) and ABLE (Ishida et al. 2003).

Score Functions

about depend upon the type of problem to be tackled. For example, a homology modeling score function would calculate a score based on the interactions between pairs of side chains, and side chains with the backbone whereas an *ab initio* folding or threading score function would be taking care of the topology of protein conformation (Huang et al. 2000).

There are two categories of score functions which may be employed by an *ab initio* structure prediction method to represent the forces, such as solvation, electrostatic interactions, van der Waals interactions, etc., which determine the energy of protein conformation.

Physics-Based Energy Functions

The biologically active native structure of a macromolecule is commonly found at the lowest free energy minima of its energy landscape (Anfinsen 1973, Edelsbrunner and Koehl 2005, Chivian et al. 2003, Sohl et al. 1998). In theory, *ab initio* energies of a molecular system can be determined by the laws of quantum mechanics i.e. Schrödinger's equation (Hartree and Blatt 1958, Hohenberg and Kohn 1964). In practice, however,

only small and simple systems can be solved by quantum mechanical methods. Therefore, protein molecules cannot be given quantum mechanical treatment due to their larger size, flexibility, and protein-solvent (Kauzmann 1959, Dill 1990) and higher-order solvent-solvent interactions in non-uniform polar aqueous environment (Frank and Evans 1945).

Ab initio calculations can be useful for the derivation of empirical physics-based energies by applying some approximations and simplifications. For instance, quantum mechanical calculations of simple systems can be used to approximate hydrogen bond geometries (Morozov et al. 2004). Similarly, electrostatic calculations using classical point charges and Lennard-Jones potentials can provide the approximation of protein-solvent polarizability and van der Waals interactions respectively. Molecular dynamics simulations have made use of such functions to determine the force fields e.g. CHARMM (Brooks et al. 1983), AMBER (Weiner and Kollman 1981), and ENCAD (Levitt et al. 1995). The parameterization of these energies is done by fitting to the experimental data.

Generally, physics-based energies work poorly in protein folding simulations due to weaknesses in solvent and electrostatic interaction modeling. However, these energies work adequately for small perturbations around a known native structure and have been used by coupling with experimental constraints obtained from NMR data to get accurate structures. PROFESY (Lee et al. 2004), a fragment assembly based structure prediction method, has reportedly used physics-based energies for prediction of novel structures.

Knowledge- or Statistics-based Functions

Knowledge-based functions are empirically derived from the properties observed in a database of already known protein structures (Tanaka and Scheraga 1976, Ngan et al. 2006, Wodak and Rooman 1993, DeBolt and Skolnick 1996, Gilis and Rooman 1996, Samudrala and Moult 1998, Sun 1993) with assumptions: 1) a free energy function can describe the behavior of a molecular system, 2) energy approximation is possible by capturing some aspects of the molecule, and 3) more frequently observable structures correspond to low-energy states. The last assumption is due to the Boltzmann principle which states that the probability density and the energies are closely related quantities (Sippl 1995). Knowledge-based functions are considered useful with the reduced models particularly for the treatment of less understood hydrophobic effects of protein thermodynamics (Kocher et al. 1994).

Rosetta (Rohl et al. 2004, Das and Baker 2008), an *ab initio* structure prediction method, uses knowledge-based score functions. In Rosetta, a target sequence is treated as a set of segments and for each segment, 25 structural fragments of 3-9 residues are extracted

from a database of fragments generated from the known protein structures. Fragment extraction is made on the basis of a multiple sequence alignment and secondary structure prediction. Protein conformational space is then searched by randomly inserting the extracted fragments and scoring the evolving conformation of target sequence by a knowledge-based function. The score function consists of hydrophobic burial, electrostatic, disulfide bond bias, and sequence-independent secondary structure terms.

Search Methods

In addition to a score function, one also needs a search method which can efficiently explore the conformational space to find the probable structures of target sequence. As the number of possible structures increases exponentially with increase in degree of freedom, the search for native-like structures is computationally an expensive NP-hard problem (Li et al. 1998, Derreumaux 2000). In the following, we have discussed few commonly used search methods.

Molecular Dynamics

In molecular dynamics simulations, classical equations of motion are solved to determine the positions, velocities and accelerations of all the atoms. A system in molecular dynamics simulations often starts with a conformation described either with Cartesian or internal coordinates (Edelsbrunner and Koehl 2005). As the system progresses towards the minimum-energy state, it reacts to the forces that atoms exert on each other. Newton's or Langrange's equations depending upon the empirical potential are solved to calculate the positions and the momenta of atoms and of any surrounding solvent throughout the specified period of time. Molecular dynamics simulations are time-consuming and require an extensive computer power to solve equations of motion (Scheraga et al. 2007, Contact and Hunter 2006).

Besides the issue of accuracies in the description of forces which atoms apply on each other, the major drawback with molecular dynamics is that when one atom moves under the influence of other atoms in the system, all the other atoms are also in motion. That means, the force fields which control atoms are constantly changing and the only solution is to recalculate the positions of atoms in a very short time slice (on the order of 10^{-14}s). The need to recalculate forces is considered the main hurdle. In principle, N^2 calculations are required for a system consisting of N atoms (from protein and water both). This constraint makes the simulation of a process, which takes just 1s in nature, out of reach fro the contemporary computers (Karplus and Petsko 1990, Unger 2004, Nölting 1999).

Efficient sampling of the accessible states and the kinetic data of transitions between states of a system are the main strengths of molecular dynamics (Cheatham III and

Kollman 2000, Karplus and McCammon 2002). At the moment, the only application of molecular dynamics is either modeling of smaller molecules (Fiser et al. 2000) or refinement and discrimination of models generated by the ordinary *ab initio* methods (Bonneau and Baker 2001).

Genetic Algorithms

Genetic algorithms use gene-based optimization mechanisms (i.e. mutations, cross-overs and replication of strings) and have been used as search method for *ab initio* protein structure prediction (Pedersen and Moult 1995, Karplus et al. 2003, Sun 1993). Mutations are analogous to the operations within a single search trajectory of a traditional Monte Carlo procedure. Cross-overs can be thought of as means of information exchange between trajectories. Genetic algorithms are co-operative search methods and can be used to develop such protocols where a number of searches can be run in parallel (Pedersen and Moult 1996)

Many studies have demonstrated superiority of genetic algorithms over Monte Carlo methods for protein structure prediction, but no method based on naive pure implementation of genetic algorithms has been able to demonstrate a significant ability to perform well in real predictions. In fact, genetic algorithms rather improve themselves on the sampling and convergence of Monte Carlo approaches (Unger 2004).

Monte Carlo Sampling Methods

Monte Carlo methods are powerful numerical techniques with a broad range of application in various fields of science. Although the first real use of Monte Carlo methods as search tool dates back to the second world war, the systematic development happened in the late 1940s when Fermi, Ulam, von Neumann, Metropolis and others made use of random number to solve problems in physics (Landau and Binder 2005, Doll and Freeman 1994). The most frequent use of Monte Carlo methods is the treatment of analytically intractable multi-dimensional problems, for example, evaluation of difficult integrals and sampling of complicated probability density functions (Kalos 2007, Binder and Baumgartner 1997).

Monte Carlo methods allow fast and reliable transformations of a natural process or its model by sampling its states through stochastic moves and calculating the average physical or geometric properties (Edelsbrunner and Koehl 2005). For instance, a molecular dynamics simulation of the same process will take a very long time than Monte Carlo simulation (Siepmann and Frenkel 1992). To perform faster conformational search for native-like states or conformations, Monte Carlo methods have been used extensively in protein folding (Avbelj and Moult 1995, Rey and Skolnick 1991, Hao

and Scheraga 1994, Yang et al. 2007) and structure prediction (Das and Baker 2008, Simons et al. 2001, Rohl et al. 2004, Fujitsuka et al. 2006, Gibbs et al. 2001, Zhou and Skolnick 2007, Kolinski 2004) studies.

In Monte Carlo method, we are supposed to estimate certain "average characteristics" of a complex system by sampling in a desired probability distribution $P(\mathbf{x})$ of that system (Cappe et al. 2007). The random variable $\mathbf{x}$ is called configuration of the system (e.g. $\mathbf{x}$ may represent three-dimensional structure of a protein conformation) and the calculations of $\mathbf{x}$ are based on the statistical inferences or basic laws of physics and chemistry (Liu 2001). Mathematically, the average characteristics or state of the system can be expressed as expectation $\bar{g}$ for some useful function g of configuration space $\mathbf{x}$ such that

$$\bar{g} = \int g(\mathbf{x}) P(\mathbf{x}) dx \tag{1.1}$$

It is often difficult to compute the expectation value $\bar{g}$ analytically, therefore the approximation problem is solved indirectly by generating (pseudo-) random samples from the target probability distribution P. In statistical mechanics, the target probability distribution P is defined by the Boltzmann distribution (or Gibbs distribution)

$$P(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})/k_{\mathrm{B}}T} \tag{1.2}$$

where $\mathbf{x}$ is the configuration of a physical system, $E(\mathbf{x})$ is the energy function, k_{B} is the Boltzmann constant, T is the system's temperature and Z is the partition function. The partition function Z can be written as

$$Z = \sum_{\mathbf{x}} e^{-E(\mathbf{x})/k_{\mathrm{B}}T} \tag{1.3}$$

If we are able to generate random samples $\{\mathbf{x}_r\}_{r=1}^{N}$ with probability distribution $P(\mathbf{x})$, then the mean of $g(\mathbf{x})$ over samples, also called Monte Carlo estimate, can be computed such that

$$\hat{g} \equiv \frac{1}{N} \sum_{r} g(\mathbf{x}_r) \tag{1.4}$$

Practically, the computation of Monte Carlo estimate $\hat{g}$ is not an easy task because of two reasons: 1) it is difficult to calculate the partition function Z in order to have a perfect probability distribution $P(\mathbf{x})$ where the samples come from, and 2) even if the partition function Z is knowable, the sampling with probability distribution $P(\mathbf{x})$ still remains a challenge particularly in case of high-dimensional spaces.

Importance Sampling

When it is not possible to sample from the Boltzmann probability distribution directly due to the inability to calculate the partition function Z (Trigg 2005), we can scale down the sampling domain by importance sampling (Metropolis et al. 1953). In order to compute final result, importance sampling focuses on important regions than covering the entire domain randomly (Doll and Freeman 1994). An importance region is comprised of a subset $\{\mathbf{x}_r\}_{r=1}^{M}$ of finite number of configuration states which occur according to their Boltzmann probability. The subset of states $\{\mathbf{x}_r\}_{r=1}^{M}$ is generated according to stochastic principles through random walks in the configurational space $\mathbf{x}$.

The limit properties of M are important to make sure that the states are visited proportional to their Boltzmann factor and also to satisfy detailed-balance condition. Detailed balance ensures that the probability of going from some state k to state j is equal to that of going from state j to state k. In the configurational space of a system, an unbiased random walk is possible only if any state of the system can be reached from any other state within a finite number of moves and this property of the system is called ergodicity (Brasseur 1990).

In order to obtain an unbiased estimate of $\bar{g}$, a ratio $\mathcal{P}_i$, also called importance function, between the probabilities of final state i and initial state $i-1$ of the system is evaluated at each move step of the random walk. If the probabilities $p(\mathbf{x}_i)$ and $p(\mathbf{x}_{i-1})$ of states i and $i-1$ are given by

$$p(\mathbf{x}_i) = \frac{e^{-E(\mathbf{x}_i)/k_{\mathrm{B}}T}}{Z}, \tag{1.5}$$

and

$$p(\mathbf{x}_{i-1}) = \frac{e^{-E(\mathbf{x}_{i-1})/k_{\mathrm{B}}T}}{Z}, \tag{1.6}$$

respectively, then the the unknowable denominator Z can be canceled in the ratio $\mathcal{P}_i$ of individual probabilities:

$$\mathcal{P}_i = \frac{p\left(\mathbf{x}_i\right)}{p\left(\mathbf{x}_{i-1}\right)} = \frac{e^{-E(\mathbf{x}_i)/k_{\mathrm{B}}T}}{e^{-E(\mathbf{x}_{i-1})/k_{\mathrm{B}}T}} \tag{1.7}$$

In order to compute $\mathcal{P}_i$ in equation 1.7, we only need energy difference of the two states

$$\Delta E = E_i - E_{i-1} \tag{1.8}$$

So, the equation 1.7 becomes

$$\mathcal{P}_i = \exp\left(\frac{-\Delta E}{k_{\mathrm{B}}T}\right) \tag{1.9}$$

To decide whether the move is to be accepted or not, a random number ξ is uniformly generated over the interval (0, 1). If ξ is less than $\mathcal{P}_i$ the move is accepted, otherwise the system remains in the same state x_{i-1} and the old configuration is retained as a new state (Landau and Binder 2005, Allen and Tildesley 1989).

Relation to our Work

Although the Boltzmann distribution is the most commonly used in physics and chemistry, it is not the only possibility. In section 2.1, we describe a scheme for distributions based on descriptive statistics. Importance sampling can be used within this framework. Simply, given any arbitrary probability distribution, one can use it as the basis of sampling method. The advantage is that in a descriptive framework, there is no explicit energy term. The probability model used in this work is given in section 2.1.2.

A completely different non-Boltzmann probabilistic model $P(\mathbf{x})$ describes the states of the system. This probabilistic model allows computation of the probability $p(\mathbf{x}_k)$ of any state k of the system. In importance sampling, a ratio $p(\mathbf{x}_i)/p(\mathbf{x}_{i-1})$ of the non-Boltzmann probabilities can be used to perform search for protein conformations. These protein conformations would be consistent with the predefined distributions of the non-Boltzmann probabilistic model $P(\mathbf{x})$. Our sampling scheme has three kinds of move schemes: 1) totally unbiased, 2) biased, and 3) 'controlled'. The bias in second move scheme has been fixed appropriately to make the sampling scheme nearly ergodic.

Chapter 2

Monte Carlo with a Probabilistic Function

The representative features of protein conformations (Marqusee et al. 1989, Blanco et al. 1994, Callihan and Logan 1999) (discussed earlier in section 1.1) can be used for their statistical description. In order to extract descriptors for statistical analysis, a set of the already known protein structures in the Protein Data Bank (PDB) was taken and each protein was treated as a set of n-residues fragments. The number of descriptors associated to each fragment depends upon their features of our interest, for instance sequence, dihedral angles, etc.

Probabilistic clustering of protein fragments through the chosen descriptors was performed to understand what kind of fragments exist in the world of protein. Technically, what comes out of this clustering are the probability distributions and the weights of descriptors. By using the obtained probability distributions and weights, one can determine the probability of any given conformation by interpreting the probabilities of its constituent fragments. In Monte Carlo, the ratio of purely probabilistically determined probabilities of the current conformation $\mathbf{x}_{i-1}$ and the proposed conformations $\mathbf{x}_i$ allows us to (have $\mathcal{P}_j$ by replacing the right hand side of equation 1.9 and) decide which one of the two conformations to be preferred. The stated non-Boltzmann descriptive statistics for our sampling scheme were obtained through Bayesian classification.

In the following, we are going to discuss: first, our score function (in terms of probabilistic framework, attribute models, and the classification model), second, how the score function was applied to find the distributions of protein conformations by clustering fragments of the known protein structures in the PDB, third, how simulated Monte Carlo annealing was uniquely coupled with the score function, and lastly, some significant results generated by this scheme of protein structure prediction.

Probabilistic Score Function

Unsupervised classification (Herbrich 2002) based on Bayesian theory was used to discover the probabilistic descriptions of the most probable set of classes in protein structures. The parameterized probability distributions of the found classes mimic the processes that produce the observed data. Such a probabilistic classification allows a mixture of real and discrete attributes of cases (i.e. protein fragments). To avoid over-fitting

of data, there is a trade off between the predictive accuracy of classification and complexity of its classes.

A probabilistic classification is not supposed to have any categorical class definitions by partitioning of data but it rather consists of a set of classes and their probabilities. Each class is defined by a set of parameters and associated models. Parameters of a class specify regions in the attribute space where that class dominates at the other classes. The final set of classes in a classification provides a basis to classify the new cases and to calculate their probabilities. The best classification is the one which gets least surprised by an unseen case (Cheeseman et al. 1988). In *ab initio* structure prediction, such a classification can possibly have encouraging consequences in the predictions of novel folds.

Bayesian Framework

Bayesian theory (Bayes Rev. 1763) provides a framework to describe degrees of belief, its consistency and the way it may be affected with change in evidence. The degree of belief in a proposition is always represented by a single real number less than one (Cox 1946, Heckerman 1990). To understand it theoretically, let E be some unknown or possibly known evidence, H be a hypothesis that the system under consideration is in some particular state. Also consider that the possible sets E and H can be mutually exclusive and exhaustive. For given E and H, Bayes' theorem is

$$P\left(H|E\right) \;=\; \frac{L\left(E|H\right)P\left(H\right)}{P\left(E\right)}. \tag{2.1}$$

The prior $P(H)$ describes belief without seeing the evidence E whereas the posterior $P(H|E)$ is belief after observing the evidence E. $L(E|H)$, the likelihood of H on E, tells how likely it is to see each possible combination of the evidence E in each possible state of the system (Howson and Urbach 1991). The likelihood and prior can be used to have joint probability for E and H

$$J\left(EH\right) \equiv L\left(E|H\right)P\left(H\right). \tag{2.2}$$

According to Bayes' rule, the beliefs change with change in the evidence and it can be shown by normalizing the joint probability (Hanson et al. 1991) as

$$P\left(H|E\right) \;=\; \frac{J\left(EH\right)}{\sum\limits_{H} J\left(EH\right)} \;=\; \frac{L(E|H)P(H)}{\sum\limits_{H} L\left(E|H\right)P\left(H\right)}. \tag{2.3}$$

Let us consider the system we are dealing with is a continuous system, then a differential $dP\left(H\right)$ and integrals can be used to replace the priors $P\left(H\right)$ and sums over

H respectively. Similarly, the likelihood of continuous evidence E would also be given by a differential $dL\,(E|H)$. Consequently, the probability of real evidence ΔE will be a finite probability:

$$\Delta L\,(E|H) \;\approx\; dL\,(E|H)\,\frac{\Delta E}{dE}. \tag{2.4}$$

In short, given a set of states, the associated likelihood function, the prior expectations of states, and some relevant evidence are known, Bayes' rule can be applied to determine the posterior beliefs about states of the system. Then, these posterior beliefs can be used to answer further questions of interest (Hanson et al. 1991).

Mathematically, it is not an easy task to implement Bayesian theory in terms of integrals and sums. In order to make an analysis tractable, possible states of the system can be described by models on an assumption that the relevant aspects of the system can easily be represented by those models. A statistical model is supposed to combine a set of variables, their information and relationship function(s) (Bruyninckx 2002). All statistical models have specific mathematical formulae and can provide more precise answers about the likelihood of a particular set of evidences.

Here, we briefly discuss the relevant statistical models before explaining how Bayesian framework works in practice.

Attribute Models

We use a notation where a set of I cases represents evidence E of a model and each case has a set $\mathcal{K}$ of attributes of size K. The case attribute values are denoted by X_{ik}, where i and k are indexes over cases and associated attributes respectively.

Multi-way Bernoulli Distribution

A discrete attribute allows only a finite number of possible values $l \in [1, 2, \ldots, L]$ for any given instance X_i in space S. Since the model expects only one discrete attribute, the only parameters are the continuous parameters $V = \{q_1, \ldots, q_L\}$. The continuous parameters are the likelihood values $L\,(X_i|VS) = q_{(l=X_i)}$ for each possible value of l. $L-1$ free parameters are constrained such that $0 \leq q_l \leq 1$ and $\sum_l q_l = 1$.

"Sufficient statistics" for the model are generated by counting the number of cases with each possible attribute value $I_l = \sum_i \delta X_{il}$. There is a prior of form similar to that of likelihood, therefore it is also referred as (Dirichlet) conjugate prior,

$$dP\left(V|S\right) \equiv \frac{\Gamma\left(aL\right)}{\Gamma\left(a\right)^{L}} \prod_{l} q_{l}^{q-1} dq_{l}. \tag{2.5}$$

where a is a parameter to parameterize the formula. The parameter a can be assigned different values to specify different priors.

Normal Distribution

Real-valued attributes represent a small range of the real number line. As scalar attribute values can only be positive, like weight, it is preferred to represent them by their logarithms (Aitchison and Brown 1957). Real-valued attributes are modeled by the standard normal distributions. The sufficient statistics include the data mean $\bar{X} = \frac{1}{I}\sum_{i}^{I} X_i$ and the variance $s^2 = \frac{1}{I}\sum_{i}\left(X_i - \bar{X}\right)^2$. The continuous parameters V consist of a model mean μ and standard deviation σ. Given the parameters V and space S, the likelihood is determined by

$$dL\left(X_i|VS\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2} dx_i \tag{2.6}$$

The parameter values, μ and σ, in the prior are treated independently i.e.

$$dP\left(V|S\right) = dP\left(\mu|S\right) dP\left(\sigma|S\right) \tag{2.7}$$

where the priors on μ and σ are flat in the range of the data and $\log(\sigma)$ respectively and can be written as

$$P\left(\mu|S\right) = \frac{1}{\mu_{max} - \mu_{min}}, \tag{2.8}$$

$$P\left(\sigma|S\right) = \sigma^{-1}\left[\log\frac{\sigma_{max}}{\sigma_{min}}\right]^{-1} \tag{2.9}$$

Multiple Attributes

The cases X_i may have multiple attributes $k \in \mathcal{K}$. The simplest way to deal with them is to treat each of them independently by considering it as a separate problem. The parameter set V consists of sets $V_k = \bigcup_{lk} q_{lk}$ (where $\bigcup_{lk}$ denotes collection across only some of the cases) or $V_k = [\mu_k, \sigma_k]$ depending upon the type of attribute k. The likelihood and the prior are given by

$$L\left(X_i|VS\right) = \prod_{k} L\left(X_{ik}|V_k S\right), \tag{2.10}$$

and

$$dP\left(V|S\right)=\prod_{k}dP\left(V_{k}|S\right) \tag{2.11}$$

respectively.

Multivariate Normal Distribution

Attributes are not always independent of each other but can also exhibit a correlation. The multivariate normal distribution is a standard model to assume a correlation between a set $\mathcal{K}$ of real-valued attributes. In the multivariate normal distribution, s_k^2 and σ_k^2 can be replaced by the data covariance matrix $\mathbf{O}_k$ and model covariance (symmetric) matrix $\mathbf{C}_k$ respectively. The data covariance matrix $\mathbf{O}_k$ is given by

$$\mathbf{O}_{k}=\frac{1}{I}\sum_{i}\left(X_{ik}-\bar{X}_{k}\right)\left(X_{ik}-\bar{X}_{k}\right)^{T} \tag{2.12}$$

where $\left(X_{ik}-\bar{X}_{k}\right)^{T}$ is the transpose of $\left(X_{ik}-\bar{X}_{k}\right)$.
The likelihood of a set of real-valued attributes $\mathcal{K}$ is a multivariate normal in K dimensions and is given by

$$\begin{aligned} dL\left(X_{i}|VS\right) &= dN\left(X_{i},\left\{\mu_{k}\right\},\left\{\mathbf{C}_{k}\right\},K\right) && (2.13)\\ &\equiv \frac{\exp\left(-\frac{1}{2}\left(X_{ik}-\mu_{k}\right)\mathbf{C}_{k}^{-1}\left(X_{ik}-\mu_{k}\right)^{T}\right)}{\left(2\pi\right)^{\frac{K}{2}}\left|\mathbf{C}_{k}\right|^{\frac{1}{2}}}\prod_{k}dx_{k} && (2.14)\end{aligned}$$

where $\left(X_{ik}-\mu_{k}\right)^{T}$ is transpose of $\left(X_{ik}-\mu_{k}\right)$ and $|\mathbf{C}_k|$ is absolute value of the corresponding determinant.
Here also, means and covariance are treated independently at all levels by the prior

$$dP\left(V|S\right)=dP\left(\left\{\mathbf{C}_{k}\right\}|S\right)\prod_{k}\left(\mu_{k}|S\right) \tag{2.15}$$

The prior on means is a simple product of individual priors (as given in equation 2.8) of all the real-valued attributes. However, the prior on $\mathbf{C}_k$ was taken by an inverse Wishart distribution (Mardia et al. 1979).

$$\begin{aligned} dP\left(\left\{\mathbf{C}_{k}\right\}|S\right) &= d\mathcal{W}_{K}^{\text{inv}}\left(\left\{\mathbf{C}_{k}\right\}|\left\{\mathbf{G}_{k}\right\},h\right) && (2.16)\\ &\equiv \frac{\left|\mathbf{G}_{k}\right|^{-\frac{h}{2}}\left|\mathbf{G}_{k}\right|^{\frac{-h-K-1}{2}}e^{-\frac{1}{2}\sum_{k}^{K}\mathbf{C}_{k}^{\text{inv}}\mathbf{G}_{k}^{\text{inv}}}}{2^{\frac{Kh}{2}}\pi^{\frac{K(K-1)}{4}}\prod_{a}^{K}\Gamma\left(\frac{h+1-a}{2}\right)}\prod_{k}^{K}d\mathbf{C}_{k} && (2.17)\end{aligned}$$

where $h = K$ and $\mathbf{G}_k = \mathbf{O}_k$. The chosen h and G_k parameter values make the prior a "conjugate" i.e. the mathematical forms of the resultant posterior $dP(\{\mathbf{C}_k\}|ES)$ and that of the prior are same.

Classification Model

The description of a single class in some space would be the simplest and straightforward application of the above models. However, in practice, one would rather like to model a space S with a mixture of classes. The classical finite mixture model (Titterington et al. 1985, Everitt and Hand 1981) is one such model which allows us to realize a multi-class space built out of single class models. It involves two kinds of parameters: 1) the discrete parameters' $T = [J, \{T_j\}]$ where J is the number of classes, T_j is the probabilistic model of each class j, and 2) the continuous parameters $\vec{V} = [\{\alpha_j\}, \{V_j\}]$ where α_j is the weight of class j and V_j denotes the free parameters, for instance mean and variance or Bernoulli probabilities, of a model for class j. In fact, the classification parameters T and $\vec{V}$ represent a combination of parameters of each class and those of the mixture.

Given a set of data E, the finite mixture model under Bayesian framework starts building a classification with the prior probability distribution $dP\left(\vec{V}T|S\right)$ over the classification parameters where the parameters, J and α_j, are treated as arbitrary priors over integers and a discrete attribute respectively. The prior distribution actually reflects our (*priori*) ignorance about the parameters and that ignorance is overcome by updating the distribution according to the information learned from the data. The posterior probability distribution of parameters gradually gets better as the *priori* ignorance goes away. The end objective of Bayesian classification system is to achieve the most probable set of classification parameters ($\vec{V}$, T) for a given number of classes and the most probable number of classes in the data irrespective of parameters. The obtained classification parameters are then used to calculate the probabilities of individual cases of being into each class.

According to the finite mixture model, the likelihood of classification is given by

$$L\left(E_i|\vec{V}TS\right) = \sum_{j}^{J} \alpha_j L\left(E_i|\vec{V}_j T_j S\right) \tag{2.18}$$

where α_j is the weight of a class j which gives the probability of any case of being into the class j, and $L\left(E_i|\vec{V}_j T_j S\right)$ is the class likelihood which describes how the members of class j are distributed. The likelihood $L\left(E|\vec{V}TS\right)$ is mathematically simple but complex enough to give the joint probability

$$dJ\left(E\vec{V}T|S\right) \equiv L\left(E|\vec{V}TS\right)dP\left(\vec{V}T|S\right). \tag{2.19}$$

The joint probability is rugged and has many local maxima. The ruggedness of $\vec{V}T$ distributions is dealt by breaking the continuous space $\vec{V}$ into small regions R rather than directly normalizing the joint probability as required by Bayes' rule. Each region R supposedly surrounds a sharp peak and no such peak should be spared in an effort to represent the peaks by R regions. A tireless search is performed to find a best combination of RT for which the "marginal" joint

$$M\left(ERT|S\right) \equiv \int_{\vec{V}\in R} dJ\left(E\vec{V}T|S\right) \tag{2.20}$$

is as large as possible.

Expectation Maximization (EM)

An EM algorithm (Dempster et al. 1977, Titterington et al. 1985) is used to find local maxima in regions R of the parameter space $\vec{V}$. To reach a maxima, EM algorithm starts with a random seed and estimates the class parameters $\vec{V}_j$ from the weighted sufficient statistics. Relative likelihood weights

$$w_{ij} = \frac{\alpha_j L\left(E_i|\vec{V}_j T_j S\right)}{L\left(E_i|\vec{V}TS\right)} \tag{2.21}$$

are calculated from the estimated class parameters. The likelihood weights, which satisfy $\sum_j w_{ij} = 1$, are used to calculate the probability that a case i would belong to a class j. The new class data and class-weighted sufficient statistics are created from the likelihood weights w_{ij}. These statistics are then substituted into the previous class likelihood function $L\left(E|\vec{V}_j T_j S\right)$ to have a weighted likelihood $L'\left(E|\vec{V}_j T_j S\right)$. The current estimate of $\vec{V}$ is used to calculate new likelihood weights w_{ij} and then the new weights w_{ij} are used to re-estimate $\vec{V}$. This iteration between two steps stops when they start predicting each other.

Calculating Probabilities

The (intra-class) probability $P(X_i \in C_j \mid \vec{V}, T, S)$ of observing an instance X_i (independent of its attribute vector $\vec{X}_i$) in class C_j is

$$P\left(X_i \in C_j \mid \vec{V}, T, S\right) \equiv \alpha_j \tag{2.22}$$

The parameters $\vec{V}$ have a set of probabilities, also called class weights, $\{\alpha_1, \ldots, \alpha_J\}$ such that $0 < \alpha_j < 1$ and $\sum_j \alpha_j = 1$. The instance attribute vectors $\vec{X}_i$ are distributed independently and identically with respect to the classes. Given an instance X_i belongs to a class C_j, the (inter-class) probability of the instance attribute vector $\vec{X}_i$ is $P\left(\vec{X}_i \mid X_i \in C_j, \vec{V}_j, T_j, S\right)$. The class probability distribution functions (p.d.f.) $\vec{V}_j$, T_j provide a conditional probability which is a product of the distributions modeling conditionally independent attributes k

$$P\left(\vec{X}_i \mid X_i \in C_j, \vec{V}_j, T_j, S\right) = \prod_k P\left(\vec{X}_{ik} \mid X_i \in C_j, \vec{V}_{jk}, T_{jk}, S\right) \tag{2.23}$$

The direct probability that an instance X_i with attribute vector $\vec{X}_i$ is a member of class C_j is obtained by a combination of the interclass and the intraclass probabilities given in equations 2.22 and 2.23 respectively:

$$P\left(\vec{X}_i, X_i \in C_j \mid \vec{V}_j, T_j, \vec{V}, T, S\right) = \alpha_j \prod_k P\left(\vec{X}_{ik} \mid X_i \in C_j, \vec{V}_{jk}, T_{jk}, S\right) \tag{2.24}$$

The probability of an instance X_i without being worried about its class memberships is given by

$$P\left(\vec{X}_i \mid \vec{V}, T, S\right) = \sum_j \left(\alpha_j \prod_k P\left(\vec{X}_{ik} \mid X_i \in C_j, \vec{V}_{jk}, T_{jk}, S\right)\right) \tag{2.25}$$

and the probability of observing a set or database X of instances is given by

$$P\left(X \mid \vec{V}, T, S\right) = \prod_i \left[\sum_j \left(\alpha_j \prod_k P\left(\vec{X}_{ik} \mid X_i \in C_j, \vec{V}_{jk}, T_{jk}, S\right)\right)\right]. \tag{2.26}$$

Applying Probabilistic Framework to Proteins

Representation of Protein Conformations

The score function works with a 5-atoms reduced representation of protein conformations. Each residue of the protein backbone is represented by 5-atoms i.e. N, C_α, C_β, C and O (as shown in figure 2.1). The side chains of protein conformations are considered only upto C_β atoms. The Cartesian coordinates of H atom of the NH group are calculated from those of N atom, preceding C atom and succeeding C_α atom.

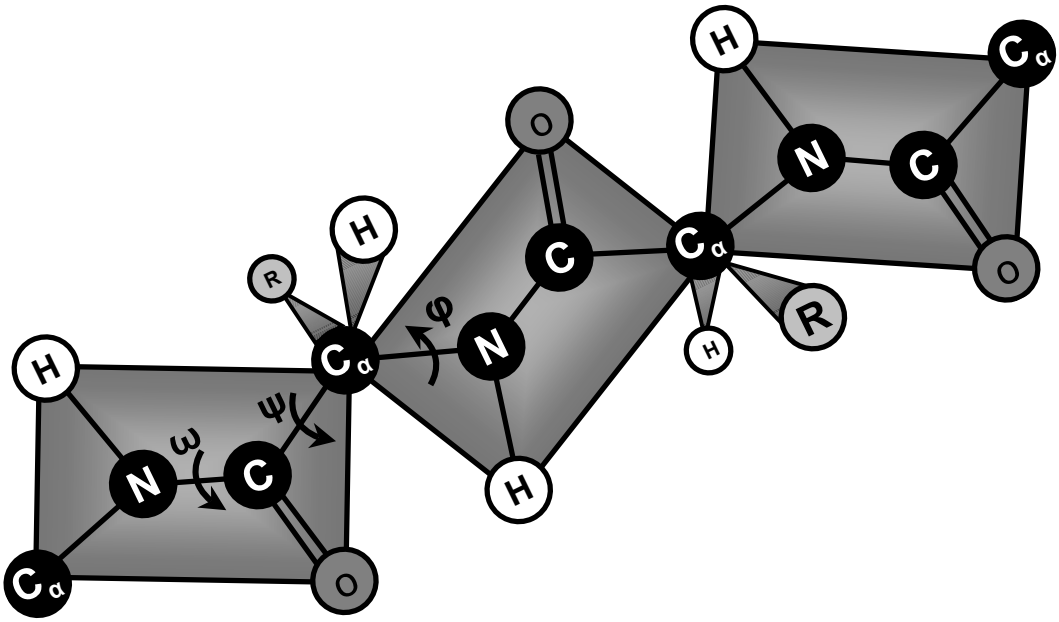


Figure 2.1: *Backbone of a protein conformation. In a Cartesian space, positions of N, H, C_α, C_β, C and O atoms of each residue are represented by their x, y and z coordinates whereas in dihedral space, ω, ϕ and ψ angles are used to define the geometric shape of the protein backbone. ω angle mostly remains close to 180^o and has little effect on the overall conformation whereas ϕ and ψ angles vary and play significant role in the variation of protein conformation. ϕ and ψ angles are defined by backbone atoms: (C, N, C_α, C) and (N, C_α, C, N) respectively. See figure 2.7 for the definition of a dihedral angle.*

Internally, the score function works with two kinds of conformation representations:

1) Cartesian coordinates, and 2) internal coordinates. In a Cartesian coordinates based representation, x, y and z coordinates of N, C_α, C_β, C and O atoms are used to represent the geometric shapes of protein conformations. Whereas in an internal coordinates based representation, bond angles, bond lengths and dihedral angles provide the definitions of conformations. The characteristic values of bond lengths and bond angles depend upon the types of atoms involved and are usually fixed with very little variation (as explained in the first chapter). That is why, bond angles and bond lengths of protein conformations are kept constant in order to improve the computational efficiency by reducing degrees of freedom (Kavraki 2007, Zhang and Kavraki 2002, Choi 2005).

There is an interplay between the two representations of protein conformations. The issue is that some terms of the score function are casted in terms of Cartesian coordinates. Whereas internal coordinates are needed for the probability calculation of conformations. The working of this interplay between two representations will be discussed later in detail.

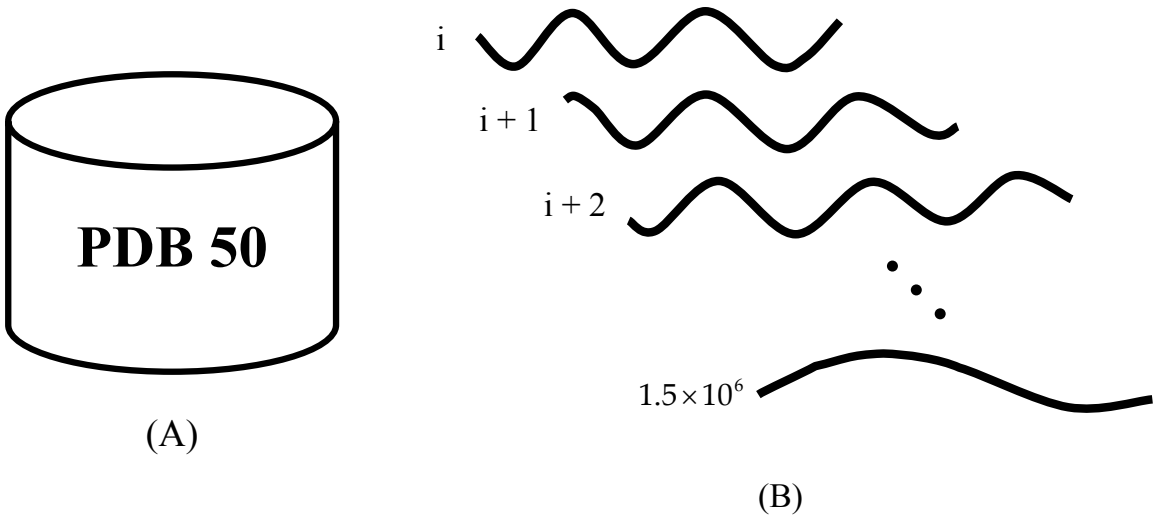
Data Collection

Protein Data Bank (PDB) (Berman et al. 2000) is a storage of very useful information in the form of already solved protein structures. It is often used by different structure prediction methods to make their score functions learn how a protein sequence may evolve into its native structure. We also trained our score function, which is based on a probabilistic classification, on commonly seen information (i.e. sequence and structure) in fragments of the known protein structures (Park and Levitt 1995, Sippl et al. 1992, Bowie and Eisenberg 1994, Jones 1997, Simons et al. 1997).

To avoid redundant information, a set of protein chains was selected from the PDB such that no two members had sequence identity more than 50% (Li et al. 2001). All the chains with less than 40 amino acids and few with unknown sequence were removed from the list of 50% sequence identity chains. From the remaining chains of protein, each possible overlapping fragment of length k was extracted. All those fragments which had any of their bond longer than 2 Å were discarded. The remaining 1.5×10^6 fragments, each of length $k \leq 7$, were assigned their attribute vectors i.e. sequence and structure (ϕ, ψ) features (as shown in figure 2.2).

Descriptors and Models

Sequence and structure (ϕ, ψ) features of protein fragment are the two main descriptors for which our classification system is supposed to find the probability distributions. Therefore, the attribute vector of a protein fragment (as shown in figure 2.2) consists of two sets of information: 1) sequence of the fragment , and 2) dihedral angles (ϕ, ψ)



	Sequence							Structure						
i	A	S	T	C	W	R	I	-3.1 -1.1	-2.0 -0.9	-0.5 -0.7	-1.7 -0.5	-2.0 -0.3	-1.5 -0.8	-2.2 -1.0
i + 1	S	T	C	W	R	I	M	-2.0 -0.9	-0.5 -0.7	-1.7 -0.5	-2.0 -0.3	-1.5 -0.8	-2.2 -1.0	-1.1 -1.1
i + 2	T	C	W	R	I	M	F	-0.5 -0.7	-1.7 -0.5	-2.0 -0.3	-1.5 -0.8	-2.2 -1.0	-1.1 -1.1	-2.1 -0.4
⋮							⋮							
N	P	L	E	N	R	R	V	3.1 1.1	2.0 0.9	1.5 -2.5	1.7 2.3	-2.0 -0.9	-1.5 -1.2	-1.2 -0.8

(C)

Figure 2.2: *(A) Database of protein chains in which no two proteins have sequence identity more than 50%, (B)* 1.5×10^6 *overlapping fragments extracted from 50% sequence identity protein chains and (C) attribute vectors of the fragments shown in (B). Each attribute vector consists of sequence and structure* (ϕ, ψ) *information of the corresponding fragments.*

which define the geometric shape of the fragment. Below, we describe the associated models of these two descriptors:

Sequence

Each of the k residues in each class was modeled by the multi-way discrete distribution discussed earlier in section 2.1.2.

Structure

The structural features of protein fragments are extracted from ϕ and ψ dihedral angles of each of their residues (as illustrated in figures 2.1 and 2.2). For classification, ϕ and ψ angles of fragments were shifted into the periods of 0 to 2π and $-\pi/2$ to $3\pi/2$ respectively and treated as continuous descriptors. To allow correlations between ϕ and ψ angles, they were modeled by the bivariate/multivariate normal distributions of the form given in equation 2.13, where X_i would be a two-dimensional vector of angle pairs (ϕ, ψ) and μ_k be the corresponding vector of means.

Classification

After the identification of appropriate descriptors and the assignment of relevant models, a set of classifications, consisting of 150-300 classes, was generated with varying parameters, e.g. for fragment length $n = 4, 5, 6, 7$. Finding a better classification is the most expensive step of our protein structure prediction method. It often takes several weeks of intensive computations. In each iteration of EM convergence, the class memberships of the fragments are computed from the class parameters and the implied relative likelihoods and then new class members are used for the computation of class statistics and revision of class parameters. These two steps are performed repeatedly until the class memberships and the class parameters stop changing.

Once we have a classification after observing 1.5×10^6 protein fragments, it can be used to make the statistical estimate of unseen (structure) attribute values of those proteins whose structures have not been solved so far.

Sampling Method

Monte Carlo simulated annealing (MCSA) (Kirkpatric et al. 1983) was used as a search method to find the probable conformational arrangements of a given target sequence. The concept of simulated annealing is based on the physical fact that melting and subsequent sudden cooling of a metal makes it very brittle. The property of brittleness shows that the metal structure is trapped in a local minimum energy state making it

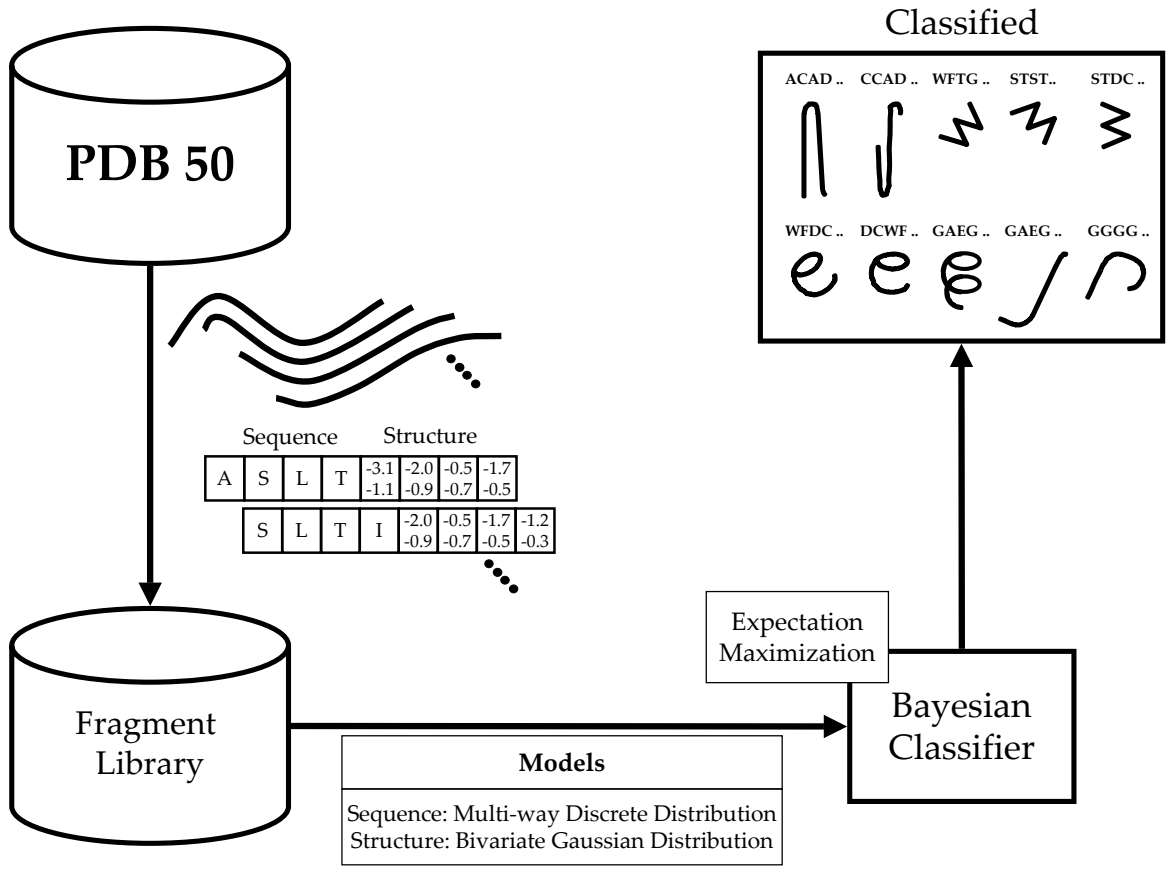


Figure 2.3: *Bayesian classification of* 1.5×10^6 *protein fragments. It finds probabilistic description of the most probable set of classes.*

so unstable. However, gradual and slow cooling of the metal makes it very tough and hard to break. The stable structural arrangement of the metal corresponds to the global energy minimum. Since the prediction of three dimensional protein structure is also a local minima problem (as shown in figure 2.4), the strategy of simulated annealing is often applied to tackle it (Chou and Carlacci 1991).

Interestingly, our search method does not rely on Metropolis Monte Carlo which is often used in molecular simulation techniques to sample over a free energy landscape. It is rather based on a fundamental statistical Monte Carlo which performs an unusual probabilistic sampling in a set of predefined distributions generated by the classification of protein fragments (according to the scheme described in section 2.1.3). Such a search method is supposed to lead to conformations of a protein sequence which are consistent with the predefined distributions.

Under this sampling scheme, there is no direct consideration of Boltzmann statistics in the acceptance criterion. The smoothness of distribution of states $\mathbf{x}_i$ is controlled through an artificial temperature described later in the chapter. The sampling scheme

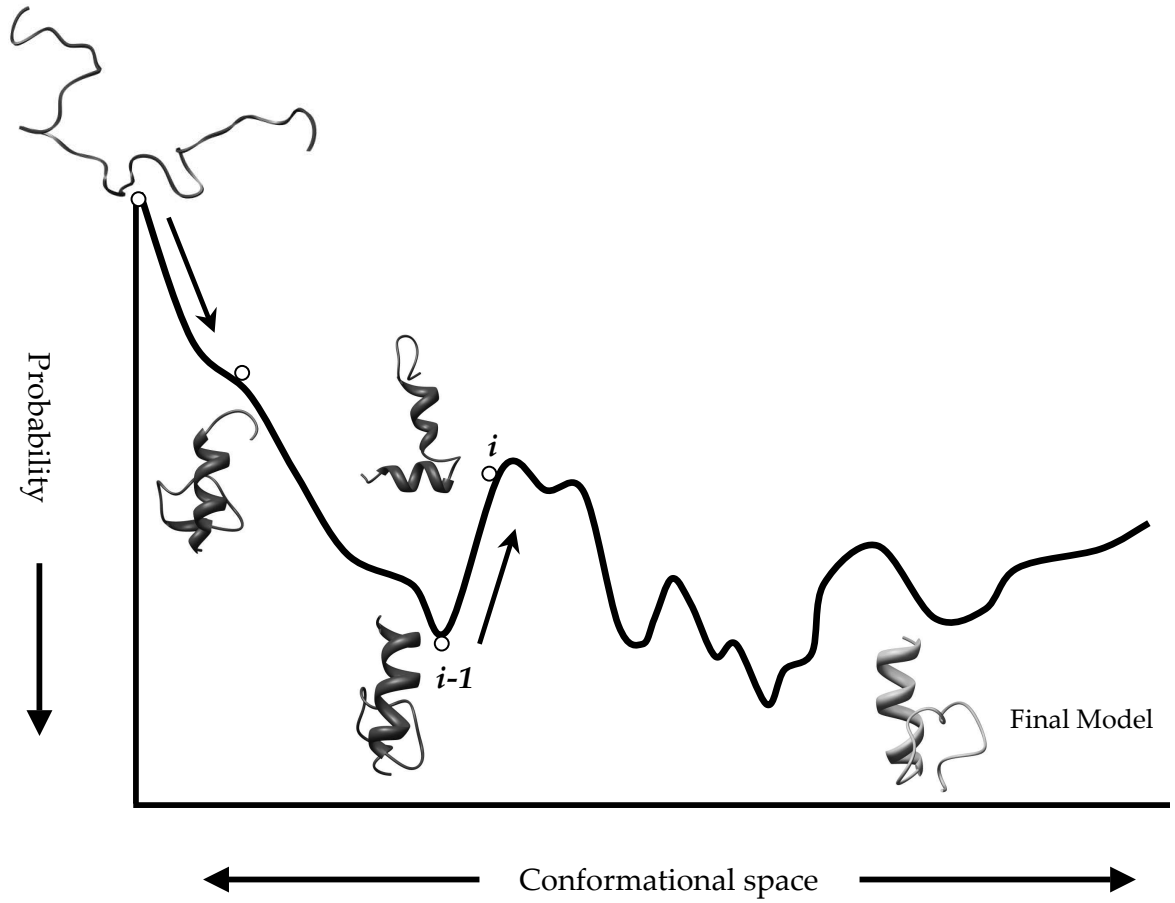


Figure 2.4: *For a given target, the system starts with a randomly generated conformation at high temperature and is gradually cooled down while making (biased or unbiased) moves. Image adapted from (Xu et al. 2006)*

does not have any kind of ensemble. However, there is a little similarity to the micro-canonical ensemble (NVE) but the volume (V) is meaningless and energy $E(\mathbf{x}_i)$ does not exist at all but probabilities $p(\mathbf{x}_i)$, where $\mathbf{x}_i$ denotes the internal coordinates of a protein conformation at ith state of the conformational space $\mathbf{x}$.

The simulated annealing process consists of N steps of random moves in the conformational space $\mathbf{x}$. The moves $\mathbf{x}_i = (i = 1, 2, \cdots, N)$ are accepted or rejected according to the Metropolis prescription (Metropolis et al. 1953) given in equation 1.7. The ratio of the probabilities of the initial conformation $\mathbf{x}_{i-1}$ and the final conformation $\mathbf{x}_i$

$$\mathcal{P}_i = \frac{p\left(\mathbf{x}_i\right)}{p\left(\mathbf{x}_{i-1}\right)} \tag{2.27}$$

is different (from the one in equation 1.7) in a sense that the probabilities $p(\mathbf{x}_{i-1})$ and $p(\mathbf{x}_i)$ are not Boltzmann probabilities but purely probabilistic in all aspects. The accep-

tance criterion can be formulated as

$$\mathbf{x}_i = \left\{ \begin{array}{ll} \mathbf{x}_i, & \text{if } \mathcal{P}_i > 1.0 \\ \mathbf{x}_{i-1}, & \text{if } \mathcal{P}_i > \xi \end{array} \right\} \tag{2.28}$$

where ξ is a random number uniformly distributed over the interval (0, 1). ξ is generated by a random number generator for N steps. Equation 2.28 shows, if $p(\mathbf{x}_i) > p(\mathbf{x}_{i-1})$, $\mathcal{P}_i > 1.0$ is true and hence the final conformation $\mathbf{x}_i$ will be accepted. Otherwise, $\mathbf{x}_i$ is accepted depending upon its probability and the temperature (described in one of the following paragraphs). The acceptance criterion ensures that both downhill and uphill moves are allowed.

Calculation of Probabilities

The probability $p(\mathbf{x}_i)$ of protein conformation $\mathbf{x}_i$ is calculated according to the equation 2.26. To simplify, let F be the number of overlapping fragments of conformation $\mathbf{x}_i$ (as shown in figure 2.5), and J be the number of classes generated by the classification (see figure 2.3), then the probability $p(\mathbf{x}_i)$ is given by:

$$p\left(\mathbf{x}_i\right) \;=\; \left(\prod_{f=1}^{F} \sum_{j=1}^{J} \alpha_{j|\mathcal{F}_f^{seq}} p_j\left(\mathcal{F}_f^{struct}\right) \right)^{1/F} \tag{2.29}$$

where $\alpha_{j|\mathcal{F}_f^{seq}}$ is the weight of a class j given a sequence $\mathcal{F}_f^{seq}$ of fragment f, $\mathcal{F}_f^{struct}$ represents structural (ϕ, ψ) features of a fragment f, and $p_j(\mathcal{F}_f^{struct})$ is the probability of the structural features of a fragment f.

Probability of Acceptance

For applications such as simulated annealing or iterating self-consistent mean field methods to convergence, one needs to be able to impose a temperature on a system. At low temperatures, the system moves towards more probable states. At higher temperatures the distribution between states becomes more even. If we have a Boltzmann distribution, we control this via the temperature. If we do not have a Boltzmann distribution, we can control the smoothness of the distribution in a more artificial way. One usually does not have the absolute probability $p(\mathbf{x}_i)$ of a state $\mathbf{x}_i$, but we will have the relative probabilities of any two states $\mathbf{x}_i$ and $\mathbf{x}_{i-1}$. This ratio $p(\mathbf{x}_i)/p(\mathbf{x}_{i-1})$ is the basis of the acceptance criterion for Monte Carlo.

The best rule for an acceptance probability is to replace $p(\mathbf{x}_i)/p(\mathbf{x}_{i-1})$ by the same probability, raised to a power a. For convenience, let this ratio as $r = p(\mathbf{x}_i)/p(\mathbf{x}_{i-1})$ then

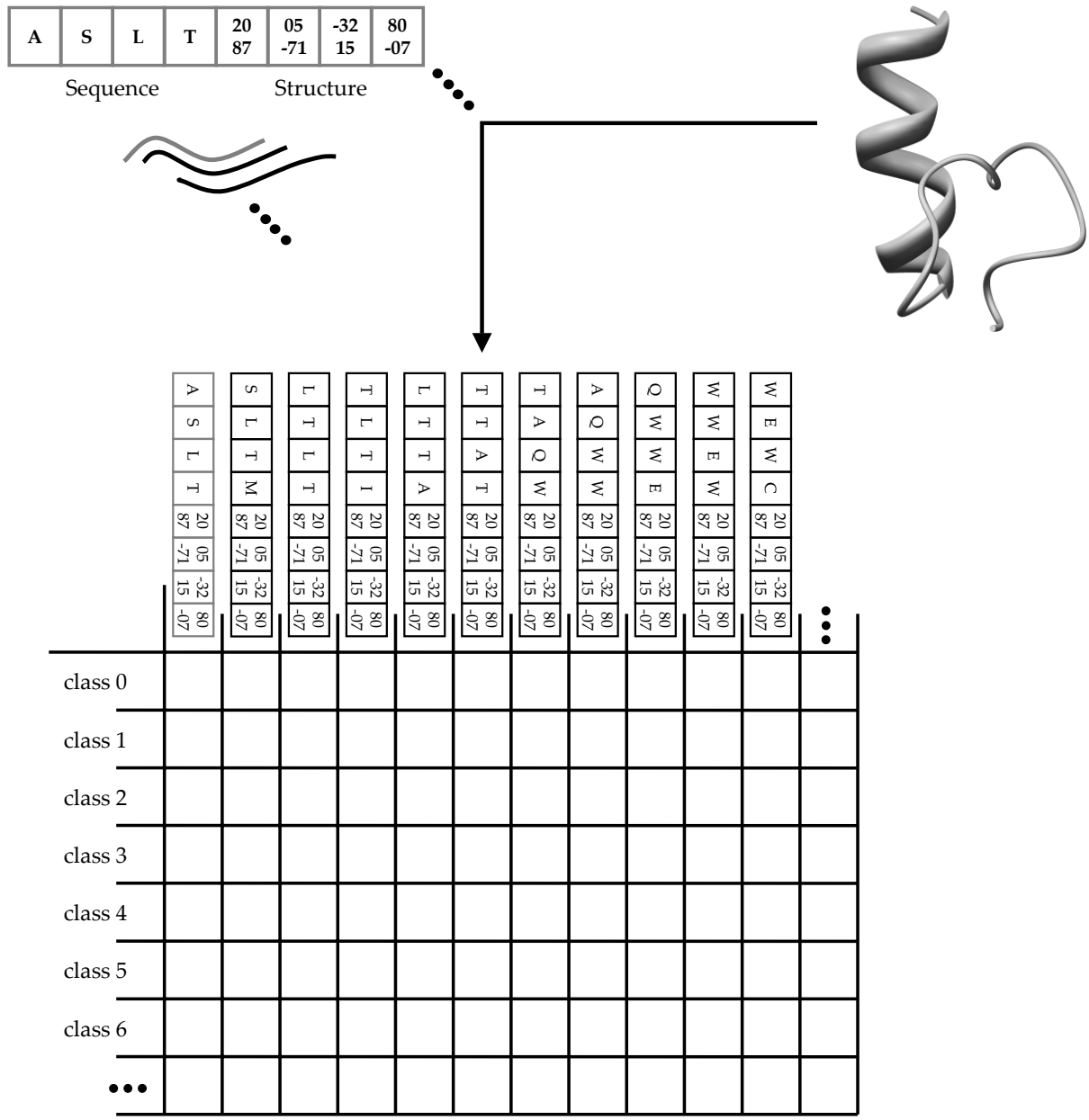


Figure 2.5: *To compute the probability of a given protein conformation of R residues, it is cut into $F = R - L_f$ overlapping fragments, where L_f is length of each fragment. Then probability vectors for all the fragments across the set of classes are computed. Finally, the product of sum of probabilities across all the classes (according to equation 2.29) gives the probability of the conformation.*

$$r^a = \left(\frac{p(\mathbf{x}_i)}{p(\mathbf{x}_{i-1})}\right)^{\frac{1}{a}} \tag{2.30}$$

As $a \to 0$, minima and maxima in the probability distribution become more pronounced. For $a > 1$, the surface is smoothed as it would be with increased temperature. This idea can be rationalized by considering two temperatures T_2 and a reference temperature T_{ref}. Then say

$$r^a = \frac{T_2}{T_2 - T_{ref}} \tag{2.31}$$

and define $\Delta E = E(\mathbf{x}_i) - E(\mathbf{x}_{i-1})$ and in this sign convention,

$$\Delta E = -k_\mathrm{B} T \ln\left(\frac{p(\mathbf{x}_i)}{p(\mathbf{x}_{i-1})}\right) \tag{2.32}$$

and for a specific temperature T_1

$$\begin{aligned} \Delta E &= -k_\mathrm{B} T_1 \ln\left(\frac{p(\mathbf{x}_i)}{p(\mathbf{x}_{i-1})}\right) \\ &= -k_\mathrm{B} T_1 \ln r^{T_1} \end{aligned} \tag{2.33}$$

If we say r^T refers to the population ratio at temperature T, then

$$\begin{aligned} \frac{r^{T_1}}{r^{T_2}} &= \frac{e^{-\Delta E/k_\mathrm{B}T_1}}{e^{-\Delta E/k_\mathrm{B}T_2}} \\ &= exp\left(\frac{-\Delta E}{k_\mathrm{B}T_1} - \frac{-\Delta E}{k_\mathrm{B}T_2}\right) \\ &= exp\left(\frac{-\Delta E}{k_\mathrm{B}}\frac{(T_2 - T_1)}{T_2 T_1}\right) \end{aligned} \tag{2.34}$$

but from equations 2.32,

$$\begin{aligned} \frac{r^{T_1}}{r^{T_2}} &= exp\left(\frac{-k_\mathrm{B} T_1 \ln r^{T_1}}{k_\mathrm{B}}\frac{(T_2 - T_1)}{T_2 T_1}\right) \\ &= exp\left(\frac{(T_2 - T_1)\ln r^{T_1}}{T_2}\right) \\ &= \left(r^{T_1}\right)^{\frac{(T_2 - T_1)}{T_2}} \end{aligned} \tag{2.35}$$

so

$$\begin{aligned} r^{T_2} &= r^{T_1}\left(r^{T_1}\right)^{\frac{(T_1-T_2)}{T_2}} \\ &= \left(r^{T_1}\right)^{1+\frac{(T_1-T_2)}{T_2}} \\ &= \left(r^{T_1}\right)^{T_1/T_2} \end{aligned} \tag{2.36}$$

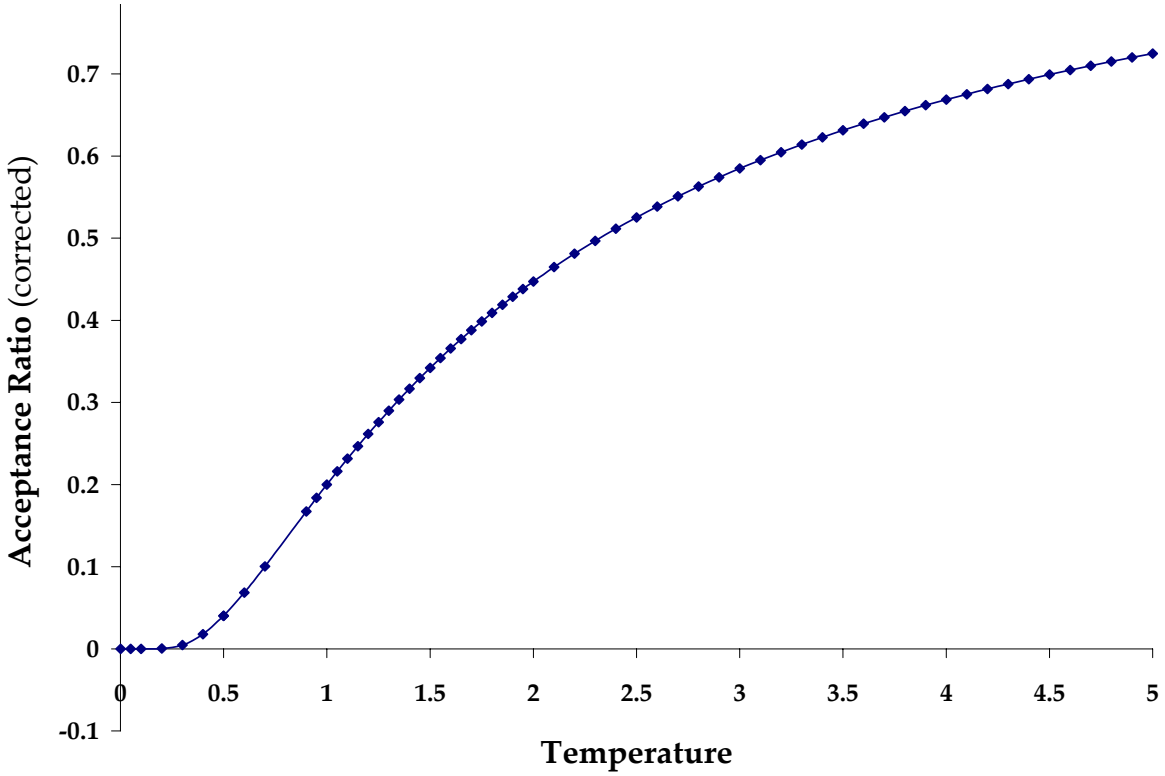


Figure 2.6: *Probability of acceptance*

which has the same effect as equation 2.30. If $T_2 \rightarrow \propto$ the distribution becomes flat and the probability of acceptance approaches 1. As $T_2 \rightarrow 0$, moves to only more likely states are accepted. See figure 2.6. One may note that this is very similar to the method based on (Tsallis 1988) statistics and used by (Andricioaei and Straub 1996).

Move Sets

Rotation about dihedral angles is the most important internal degree of freedom. The move set of our search method is based on rotations or changes in one or more dihedral angles of protein conformation. A dihedral angle is the smallest angle between the two planes P_1 and P_2 where each plane is defined by the four consecutive atoms of the protein backbone. See figure 2.7.

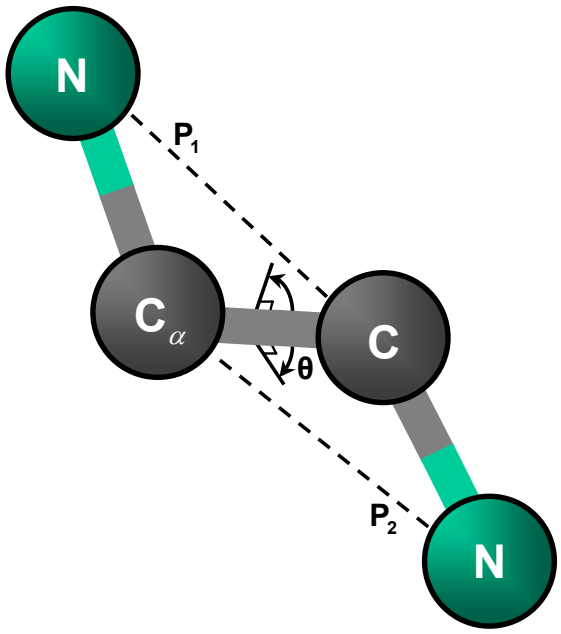


Figure 2.7: *Dihedral angle θ is the smallest angle between two planes P_1 and P_2. Plane P_1 is defined by N, C_α and C atoms whereas P_2 is defined by C_α, C and N.*

Interplay of Two Conformation Representations

The interconversion of two representations, the Cartesian coordinates and the internal coordinates of a protein conformation, is an essential component for the working of our score function. The internal coordinates (i.e. dihedral angles: ϕ, ψ) are needed for the calculation of structural features whereas the Cartesian coordinates are needed for the visualization of conformations, the calculation of hydrogen bonding energies, and partly for the computation of solvation features described in chapter 3. A search move, no matter whether its is biased or un-biased, is made by perturbing dihedral angles at a randomly selected residue position of a conformation. In order to express the changes in dihedral angles, one needs to update the Cartesian coordinates of the conformation.

There are three methods to update the Cartesian coordinates of a conformation from its dihedral angles: the simple rotations, the Denavit-Hartenberg local frames, and the atom-group local frames (Choi 2005). The simple rotations method applies a sequence of rotations, each determined by two points and an angle, to update all atom positions. The order of updates and some bookkeeping of the atom positions is necessary. The Denavit-Hartenberg local frames method uses local frames at the bonds and a series of matrix multiplications is applied to update atom positions. There is no bookkeeping in the Denavit-Hartenberg local frames method but multiple local frames are needed for bonds which have more than one child. The atom-group local frames method is based

on the concept of *atomgroups*. All the connected atoms are considered in one *atomgroup* if none of the bonds between them rotates. Only one local frame is required for each atomgroup regardless of number of children *atomgroups* (Zhang and Kavraki 2002).

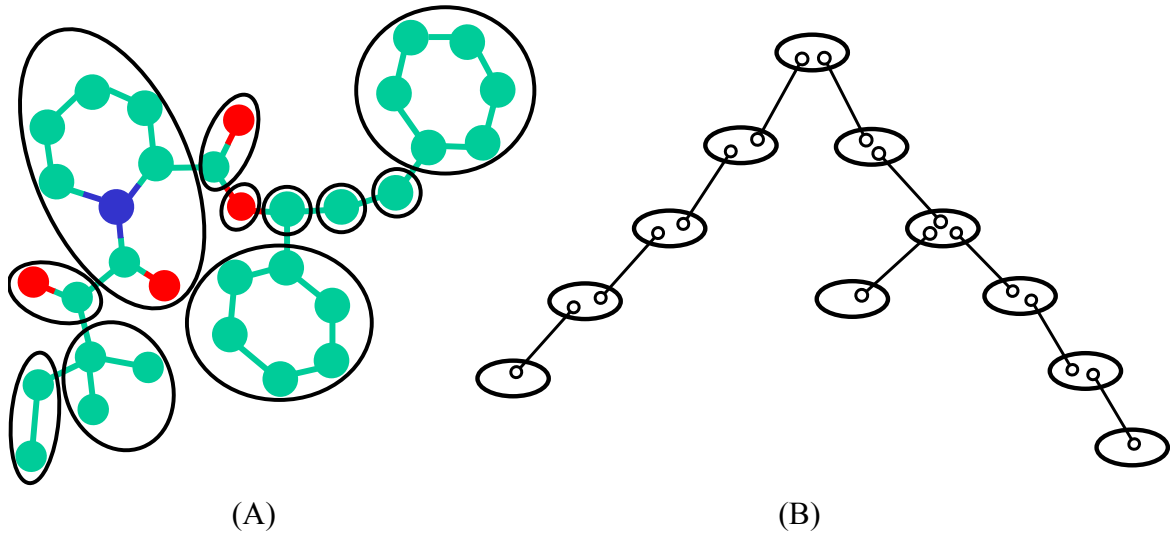


Figure 2.8: *(A) A conformation is represented as a set of rigid fragments in circles where each fragment consists of a group of atoms connected through non-rotatable bonds and (B) Rooted tree representation of fragments from (A) where each vertex represents a rigid fragment and edge a rotatable bond.*

In principle, the simple rotations method is a global frame method whereas the other two are local frame methods. To update specific atoms of a molecule, a local frame method might outperform the simple rotations method. However, when updating an entire conformation, there is no significant difference between both methods. In this work, we have used the improved simple rotations method (Choi 2005).

According to the improved simple rotations method, a protein conformation or molecule is divided into rigid fragments. A group of atoms connected through the non-rotatable bonds is considered as a rigid fragment. By choosing one of the fragments as a root, the entire conformation can be represented as a rooted tree (see figure 2.8). Each rotatable bond b_i is assigned its parent and child atoms denoted as P_i and Q_i respectively. The rigid motion along the rotatable bonds is a transformation that rotates all the atoms around a fixed origin while keeping distances between them preserved. Such a motion can easily be expressed by a rotation followed by a translation.

The transformation to rotate a bond b_i by angle Θ_i, which also rotates all descendants of Q_i by angle θ_i, is given by:

$$\mathbf{M}_i \;=\; [\mathbf{R}_i, Q_i - \mathbf{R}_i\,(Q_i)] \tag{2.37}$$

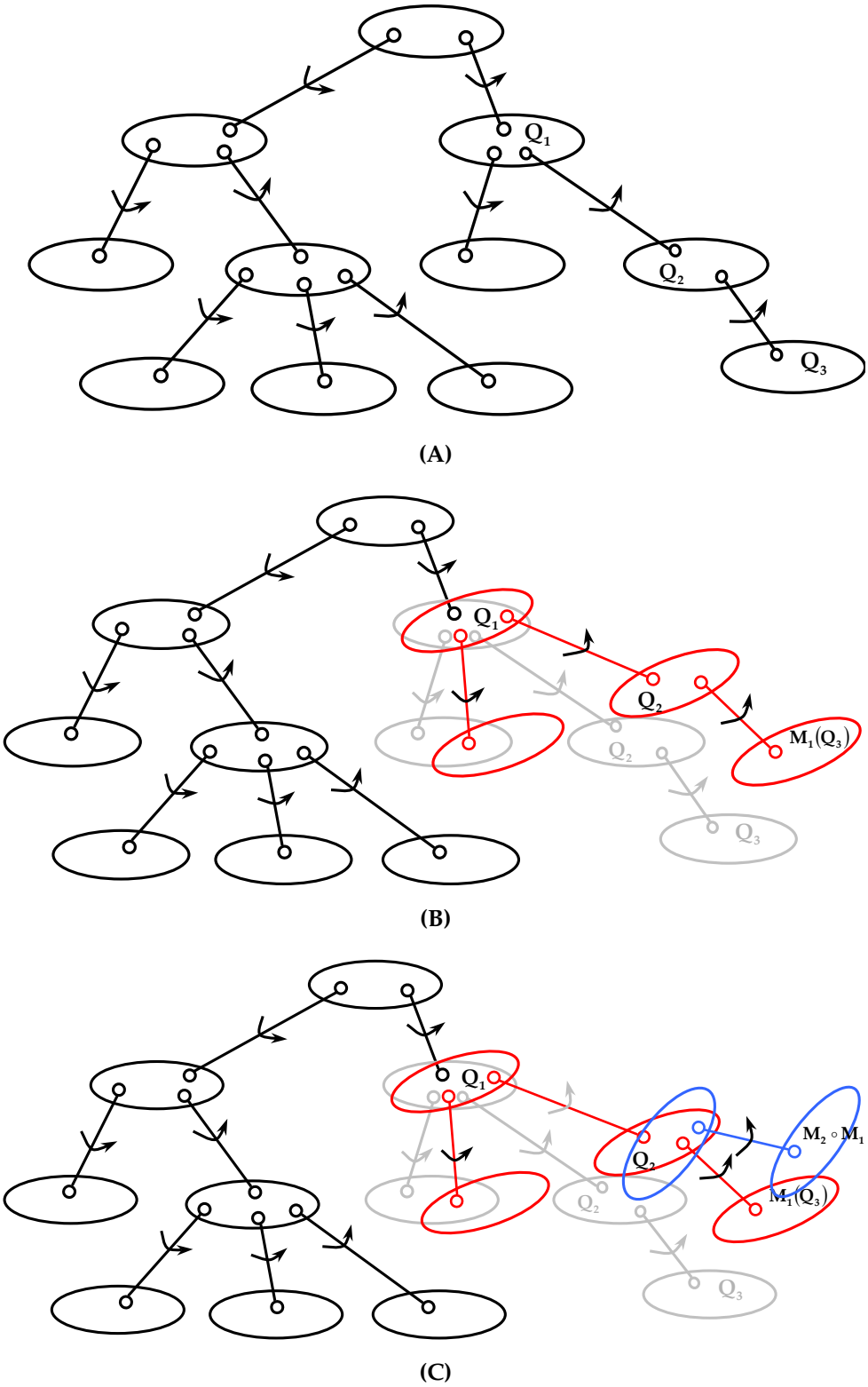


Figure 2.9: *(A) Path of rotation consists of child (rigid) fragments Q_1, Q_2 and Q_3. (B) Descendant fragments (atoms) of Q_1 are transformed by $\boldsymbol{M}_1$. (C) Descendant atoms of Q_2 are transformed by $\boldsymbol{M}_2$ i.e. $Q_3 = \boldsymbol{M}_2 \times \boldsymbol{M}_1(Q_3)$*

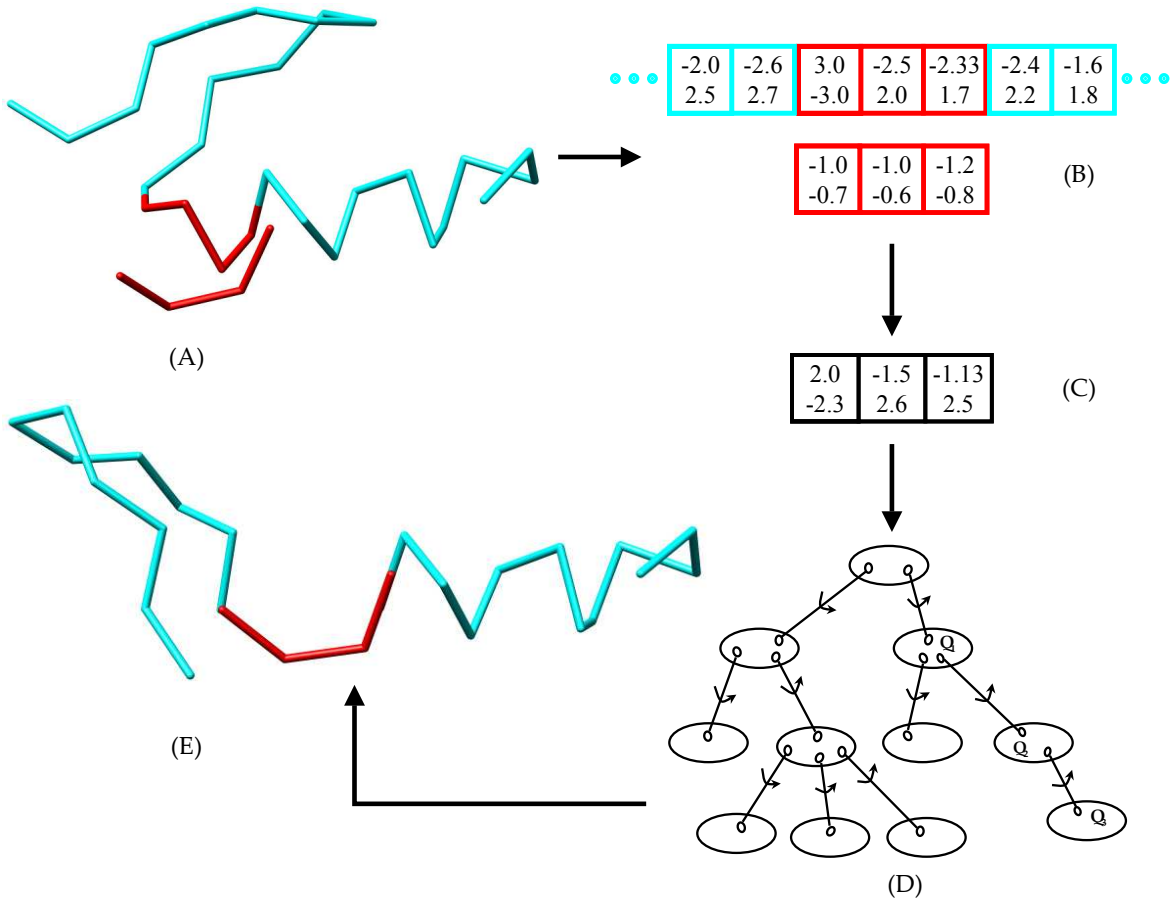


Figure 2.10: *(A) Protein conformation and a fragment to be inserted into the former at location marked with red color, (B) ϕ, ψ dihedral angles of the conformation and those of the fragment, (C) dihedral angle difference between the fragment and the part of the conformation to be replaced with the former, (D) a rooted tree construction of the conformation (in accordance with the figures 2.8 and 2.9) to apply rotations according to the dihedral angle difference calculated in (C) and (E) the newly updated conformation after an insertion of the fragment.*

where $\mathbf{R}_i$ is rotation applied to Q_i after translating it to the origin.

Given a sequence of bonds $b_1, \ldots, b_{i-1}$, b_i for a path from atom Q_1 to atom Q_i and a transformation $\mathbf{M}_k$ to make rotation about a bond b_k by angle Θ_k, for $k = 1, \ldots, i$, new position of atom Q_i, after rotating about bonds $b_1, \ldots, b_{i-1}$ by angles $\Theta_1, \ldots, \Theta_{i-1}$ respectively, is given by: $Q_i' = \mathbf{M}_{i-1} \times \ldots \times \mathbf{M}_2 \times \mathbf{M}_1(Q_i)$. For illustration, see figure 2.9. The individual rigid motions, for $i \geq 1$, can be treated as a single accumulated rigid motion $\mathbf{N}_i = \mathbf{M}_i \times \ldots \times \mathbf{M}_2 \times \mathbf{M}_1$ or $\mathbf{N}_i = [\mathbf{S}_i, \mathbf{T}_i]$ where $\mathbf{S}_i$ is a rotation and $\mathbf{T}_i$ is a translation. Since there is a common rotation between $\mathbf{M}_i$ and $\mathbf{N}_{i-1}$, therefore the only required rigid motions for computations are $\mathbf{N}_i$'s and $\mathbf{M}_i$ can be ignored. This gives us

$$\left\{ \begin{array}{l} \mathbf{N}_1 = \mathbf{M}_1 \text{ if } i = 1 \\ \mathbf{N}_i = \mathbf{M}_i \times \mathbf{N}_{i-1} = [\mathbf{R}_i \times \mathbf{S}_{i-1}, \mathbf{R}_i (\mathbf{T}_{i-1}) + Q_i - \mathbf{R}_i (Q_i)] \text{ if } i > 1 \\ \quad = [\mathbf{R}_i \times \mathbf{S}_{i-1}, \mathbf{R}_i (\mathbf{T}_{i-1} - Q_i) + Q_i (Q_i)] \end{array} \right\} \tag{2.38}$$

There are several representations for a rotation: Euler-angle, angle-axis, unit quaternion etc. In our implementation, we used quaternion representation of rotation. Unit quaternion representation of a rotation is $q = (q_0, q_x, q_y, q_z)$, where $q_0 = cos(\theta/2)$ and $(q_x, q_y, q_z) = sin(\theta/2)\mathbf{v}$, θ is rotation angle, and $\mathbf{v}$ is the unit vector along rotation axis (through the origin). After a rotation q, new position p' of a point $p \in R^3$ is given by $b' = qb\tilde{q}$, where $\tilde{q} = (q_0, -q_x, -q_y, -q_z)$ is the conjugate of q and $b = (0, x, y, z)$ for any $b = (x, y, z) \in R^3$. Multiplication of quaternions gives us corresponding rotation matrix $\mathbf{Q}$

$$\mathbf{Q} = \begin{pmatrix} 2(q_0^2 + q_x^2) - 1 & 2(q_x q_y + q_0 q_z) & 2(q_x q_z + q_0 q_y) \\ 2(q_x q_y + q_0 q_z) & 2(q_0^2 + q_y^2) - 1 & 2(q_y q_z + q_0 q_x) \\ 2(q_x q_z + q_0 q_y) & 2(q_y q_z + q_0 q_x) & 2(q_0^2 + q_z^2) - 1 \end{pmatrix} \tag{2.39}$$

Using quaternions, the equation 2.38 becomes

$$\left\{ \begin{array}{l} \mathbf{N}_1 = \mathbf{M}_1 = \left[q_1, Q_1 - q_1 \mathring{Q}_1 \tilde{q}_1\right] \text{ if } i = 1 \\ \mathbf{N}_i = \mathbf{M}_i \times \mathbf{N}_{i-1} = \left[q_i s_{i-1}, q_i (\mathbf{T}_{i-1}\mathring{} - Q_i)\tilde{q}_i + Q_i\right] \text{ if } i > 1 \end{array} \right\} \tag{2.40}$$

where q_i and s_i are unit quaternions to represent $\mathbf{R}_i$ and $\mathbf{S}_i$ respectively.

To predict structure of a target sequence, the simulation starts with a randomly generated structure at a high temperature (as shown in figures 2.4 and 2.11). Subsequently, the temperature is gradually lowered to cool the system down. As the temperature of the system is lowered, the search method makes two kinds of moves: biased, and unbiased to maximize the probability which corresponds to the minimization of energy in the traditional Monte Carlo simulations.

Biased Moves

Biased moves were introduced with an intention that the search method could spend most of the computational time exploring likely regions of the conformational space by avoiding very unlikely ones. In rigorous Monte Carlo, there is a way to correct such a bias by making the acceptance criterion harder to accept moves in more probable regions of the conformational space. One can correct this bias by making moves correspondingly less likely to be accepted.

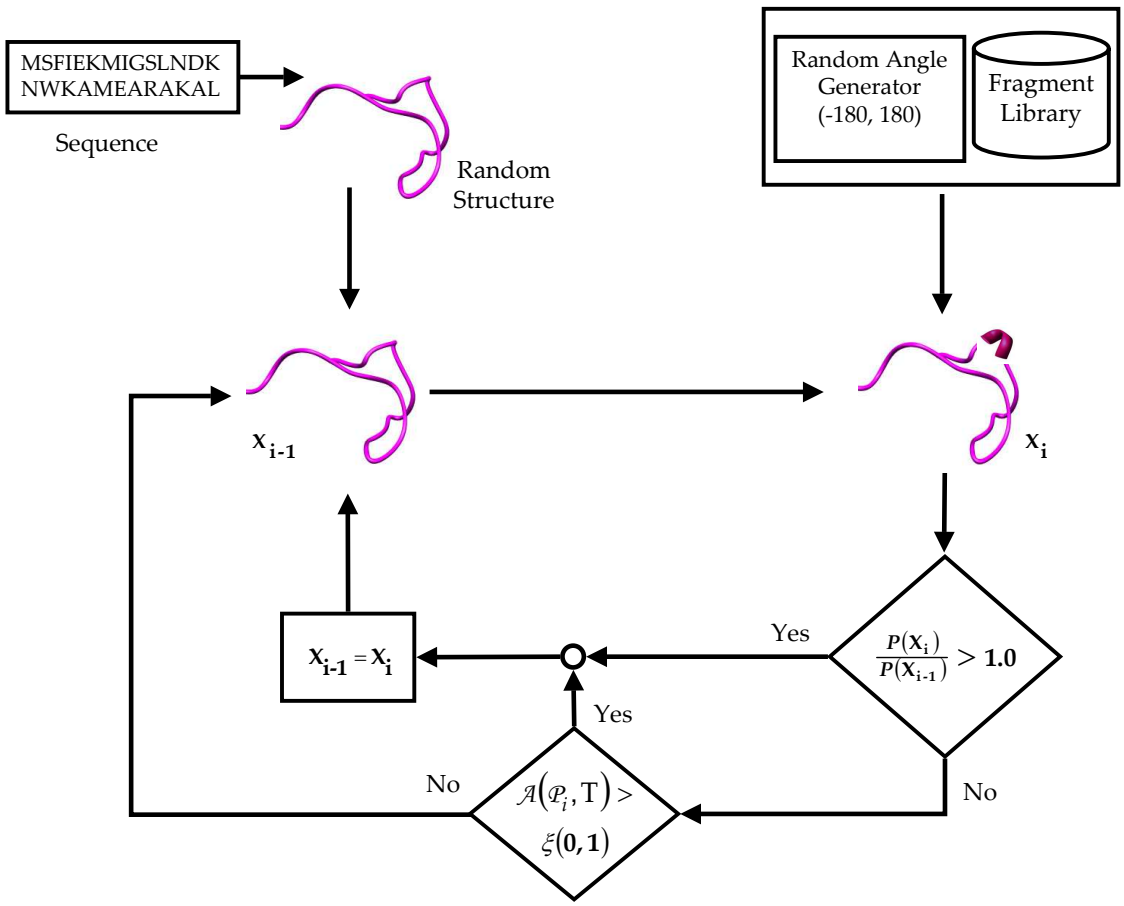


Figure 2.11: *Given a target sequence, protein structure prediction simulation starts with a random conformation. Then, at each search (biased or unbiased) move step, the probabilities $p(\mathbf{x}_{i-1})$ and $p(\mathbf{x}_i)$ of old and proposed conformations, respectively, are calculated. If the ratio of the probabilities $\mathcal{P}_i$ is greater than 1.0, the move step is accepted by retaining the new conformation otherwise the acceptability is decided according to the acceptance criterion $\mathcal{A}$.*

Biased moves are made by randomly drawing a fragment from a fragment library that was generated from the known protein structures in the PDB (see figure 2.12). The drawn fragment is then inserted at a randomly chosen residue position of the current conformation. The fragment insertion and conformation update procedure is shown in the figure 2.10. The length of fragments used in biased moves ranges from 1 to 5 residues.

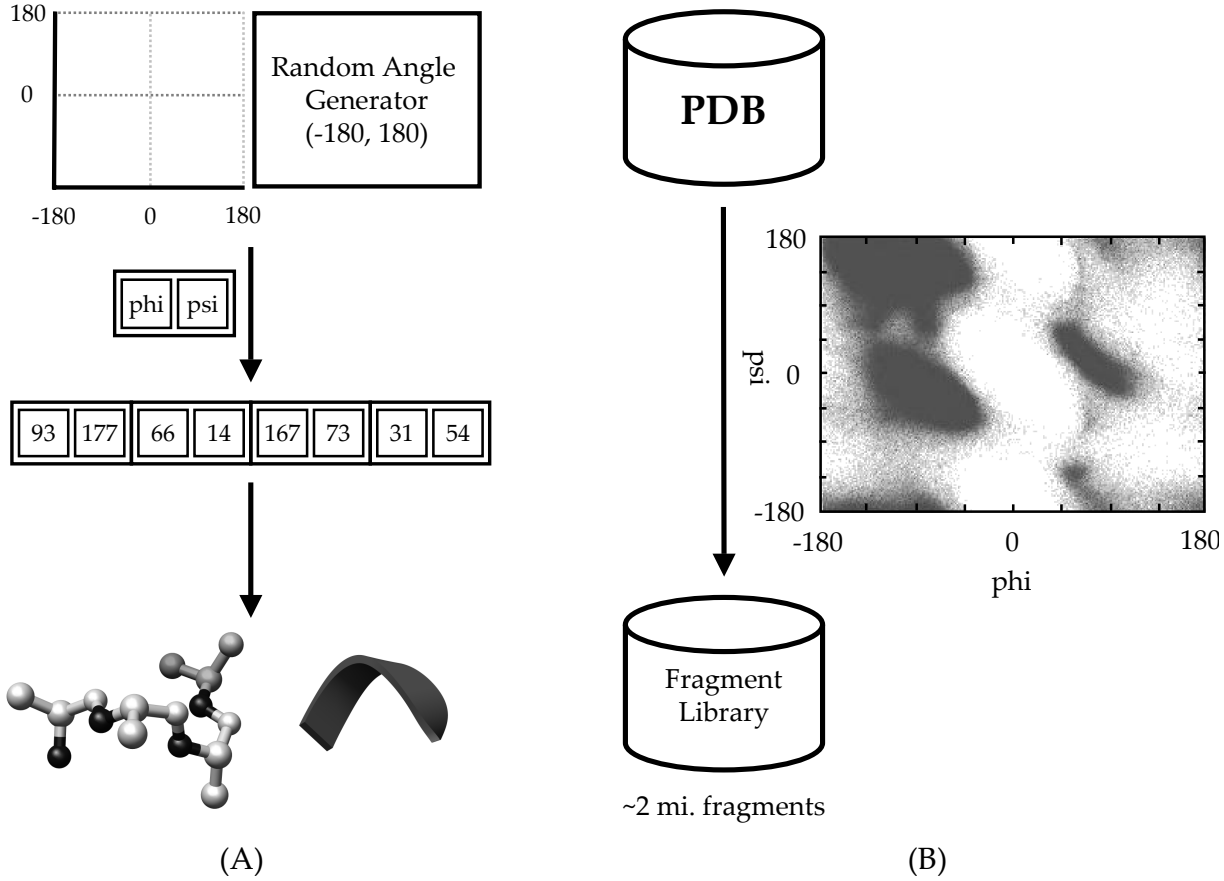


Figure 2.12: *(A) Fragments for unbiased moves are generated by drawing ϕ, ψ dihedral angles at random over the interval (-180, 180), and (B) fragment library for biased moves consists of about $\sim$ 2 million fragments (of length 1-5 residues) generated from the known proteins in the PDB.*

Statistically Unbiased Moves

Statistically unbiased moves are made by inserting random fragments at randomly chosen residue positions of the current conformation. A random fragment is generated by picking each pair of (ϕ, ψ) angles of the fragment randomly over the interval (-180, 180) as shown in figure 2.12. A simulation run with unbiased moves may take more than usual computational time to produce native-like conformations of target sequences.

Results

Method performance was evaluated by predicting the three-dimensional structures of carefully selected protein sequences. In the following, we have described (sequence) test set, assessment measures and the results.

Table 2.1: *Details of the targets used for the evaluation of prediction method*

PDB ID	**CASP7**[1] **ID**	**CASP7 Category**	**No. of Residues**	**Seq** [2] (%)	**DSSP**[3]	
					Helix(%)	Beta[6](%)
2GZV	T288	TBM[4]	93	45	13	27
2H40	T309	FM[5]	76	28	7	26
2HD3	T306	TBM	95	20	9	44
2HE4	T340	TBM	90	58	14	36
2HF1	T348	FM	68	41	14	32
2HFV	T349	TBM	75	30	26	16
2HFQ	T353	FM	85	25	22	22
2HJJ	T358	TBM	87	16	21	27
2IWN	T359	TBM	97	44	14	40
2HJ1	T363	TBM	97	19	12	32

Target Sequences:

Protein sequences in the test set were selected from a sequence list released during 7^{th} biannual competition, also called 7^{th} *Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP)*, held in 2007 by the protein structure prediction center (Center 2007). The CASP sequences, often called targets, are the sequences of those proteins whose structure are either expected to be solved shortly or have been solved but not made public yet.

Broadly speaking, there are two kinds of targets released by the CASP organizers: 1) the ones which have high similarity to the sequences of already known structures, and 2) those which have very low or almost no similarity to the sequences of known

[1] 7^{th} *Community Wide Experiment on* Critical Assessment of Techniques for Protein Structure Prediction
[2] Sequence identity to the known protein structures
[3] Database of secondary structure in proteins
[4] Template-based modeling
[5] Free modeling
[6] Beta sheet

structures. Three-dimensional structures of the former and the latter are essentially predicted by template-based and free modeling methods respectively.

Test set includes rather smaller targets with size of their sequences ranging from 68 to 97 residues. Sequence similarity of the targets to the known structures is between 16% and 58%. Three out of ten targets are purely free modeling targets whereas the rest of the targets are supposedly candidates of template-based modeling. Natives of all the targets have 7% to 26% helices and 16% to 40% beta sheets, according to DSSP secondary structure assignment (Kabsch and Sander 1983). These details about the data set are also given the table 2.1.

Prediction Parameters

Temperature, number and type of moves and the size of fragments for the moves are some of the significant parameters. These parameters influence prediction results to a great extent. The search method, simulated annealing Monte Carlo, always starts with an initial randomly generated conformation at a very high temperature i.e. 20.0. The unit of temperature is not given as it is not a real temperature but an arbitrary temperature. In subsequent move steps, the system cools down slowly while temperature is gradually lowered to ≈ 0.0. The speed of cooling process mostly depends on the number of move steps. In our studies, the search method attempted $200,000$ to $500,000$ move steps in predicting structures of different targets. The choice between biased and unbiased moves (discussed earlier in sections 2.3.3) severely affects the search in conformational space. These moves may use fragments of size ranging from 1 to 5 residues.

With different combinations of parameter values i.e. fragment size (1-mer, 2-mer, 3-mer, 4-mer and 5-mer), varying number of move steps, and biased or unbiased moves, several simulation runs with different initial random conformations were conducted for all the targets. Finally, probabilistic scores of the generated conformations allowed us to select appropriate models of the targets.

Model Assessment

Root mean square deviation (RMSD) score is often used to assess the quality of a model. RMSD score is an average distance between the backbones of native and model structures. Our models at this point in time are not as compact as native structures mainly due to the absence of medium to long-range interactions. This makes evaluations of a model rather harder using RMSD score only. Therefore, we had to use of other quality assessment measure such as the radius of gyration and plots between the dihedral angles of native and model structures. The radius of gyration is an indicator of structure compactness. It is defined as root mean square distance of backbone atoms of a protein structure from the center of gravity.

In the following, we describe some of the models and the parameters used to generate these models.

2HFQ

Model for the target 2HFQ shown at the top left in the figure 2.13 was generated by a prediction run attempting 200,000 totally unbiased move steps. These unbiased moves used 3-mer fragments generated by picking their dihedral angles (ϕ, ψ) randomly. This lead to a pretty sensible model structure with its dihedral angles closer to those of the native structure as shown at the bottom right in the figure 2.13.

Table 2.2: *The radii of gyration of random, model and native structures of the targets.*

Target ID		**Radius of Gyration (Å)**		
PDB ID	**CASP ID**	**Random**[1]	**Model**	**Native**
2GZV	T288	423.5	805.9	369.5
2HD3	T306	532.2	1022.5	420.6
2HF1	T348	231.3	384.5	229.3
2HFQ	T353	397.5	502.3	338.2
2HJ1	T363	541.6	447.6	336.3
*2HF1[2]	T348	196.9	478.2	229.3

On secondary structure level, two helices and beta strands in 2HFQ model (colored yellow and green respectively) were predicted quite well. The dihedral angle plot in the figure 2.13 also demonstrates a convergence from an initially randomness conformation to the one which is somewhat related to the native structure in terms of distribution of its dihedral angles. However, 2HFQ model has very large radius of gyration of 502.3 Å which is far from perfect compaction of its native. The radius of gyration of native structure is equal to 338.2 Å.

2HF1

2HF1 is one of the interesting targets because helix makes only 14% in its native structure and the rest is either beta sheets or turns.

The same set of parameters given above for target 2HF1 (i.e. 200,000 moves, 3-mer unbiased fragments) came up with a model for the target 2HF1. Both model and native

[1]Randomly generated conformation of a target sequence.
[2]Simulation run involved biased moves.

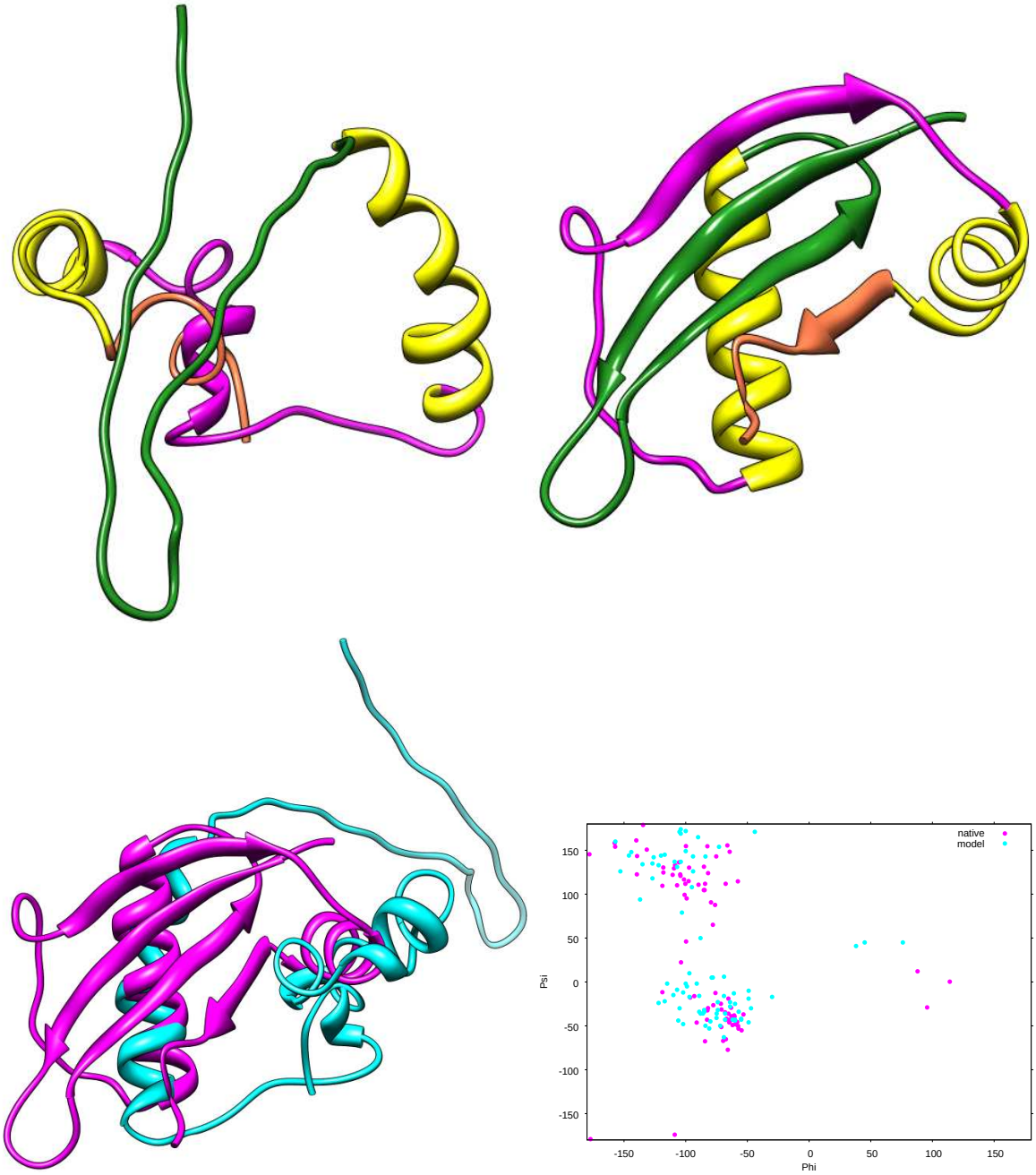


Figure 2.13: *Target 2HFQ - Top right: native, top left: model, bottom left: superimposition of native and model structures (colored magenta and cyan respectively), and bottom right: dihedral angles (ϕ, ψ) of native and model. Radius of gyration: native = 338.2 Å, and model = 502.3 Å. Simulation parameters: 200,000 unbiased move steps using 3-residues fragments.*

of 2HF1 are shown at the top in figure 2.14. The overall organization of this model has a weak similarity to that of the native structure but former is still not very compact. The radii of gyration model and native are 384.5 Å and 229.3 Å respectively. Secondary structures of model and native also looks similar except beta sheets colored with magenta color. Likewise, the dihedral angles of model and native plotted at the bottom right in figure 2.14 relate to each other.

It is worth mentioning that both of the targets (i.e. 2HFQ and 2HF1) whose predicted models described above have been categorized as free modeling targets by the CASP7 organizers. For free modeling targets, it is very unlikely to produce their models using contemporary homology or threading methods.

2HD3

2HD3 is a target with 20% sequence similarity to the known structures and helix making only 9% of its native structure. These two characteristics make it an interesting target for structure prediction. The predicted model for target 2HD3 (given at the top left of figure 2.15) involved 500,000 attempted moves by the search method. The rest of parameters are same as those of targets 2HFQ and 2HF1.

In 2HD3 model, the prediction of a single helix and beta sheet strands (colored yellow, green and cyan respectively) looks quite accurate. Had the score function had some terms for medium to long range interaction, the misplaced strands of this model could have made a proper beta sheets. The extended nature of model structure also contributed to an increased radius of gyration of 1022.5 Å which is almost twice of that of native structure i.e. 420.6 Å.

2GZV

2GZV may be ranked as a target of an average difficulty. It has 45% sequence similarity to the known structure. Helix and beta sheet are 13% and 27% respectively. 2GZV model (given at the top left in figure 2.16) has all the strands of its beta sheets, and helices correctly predicted. In this case also, secondary structures of model could not collapse quite well. As a consequence, the radius of gyration of this model is higher (i.e. 805.9 Å) than that of native (i.e. 369.5 Å). However, the dihedral angles of model (given at bottom right in figure 2.16) calculated out of model (secondary structures) are in agreement with those of native. For this prediction, the search method attempted 200,000 unbiased move steps with 1-mer fragments.

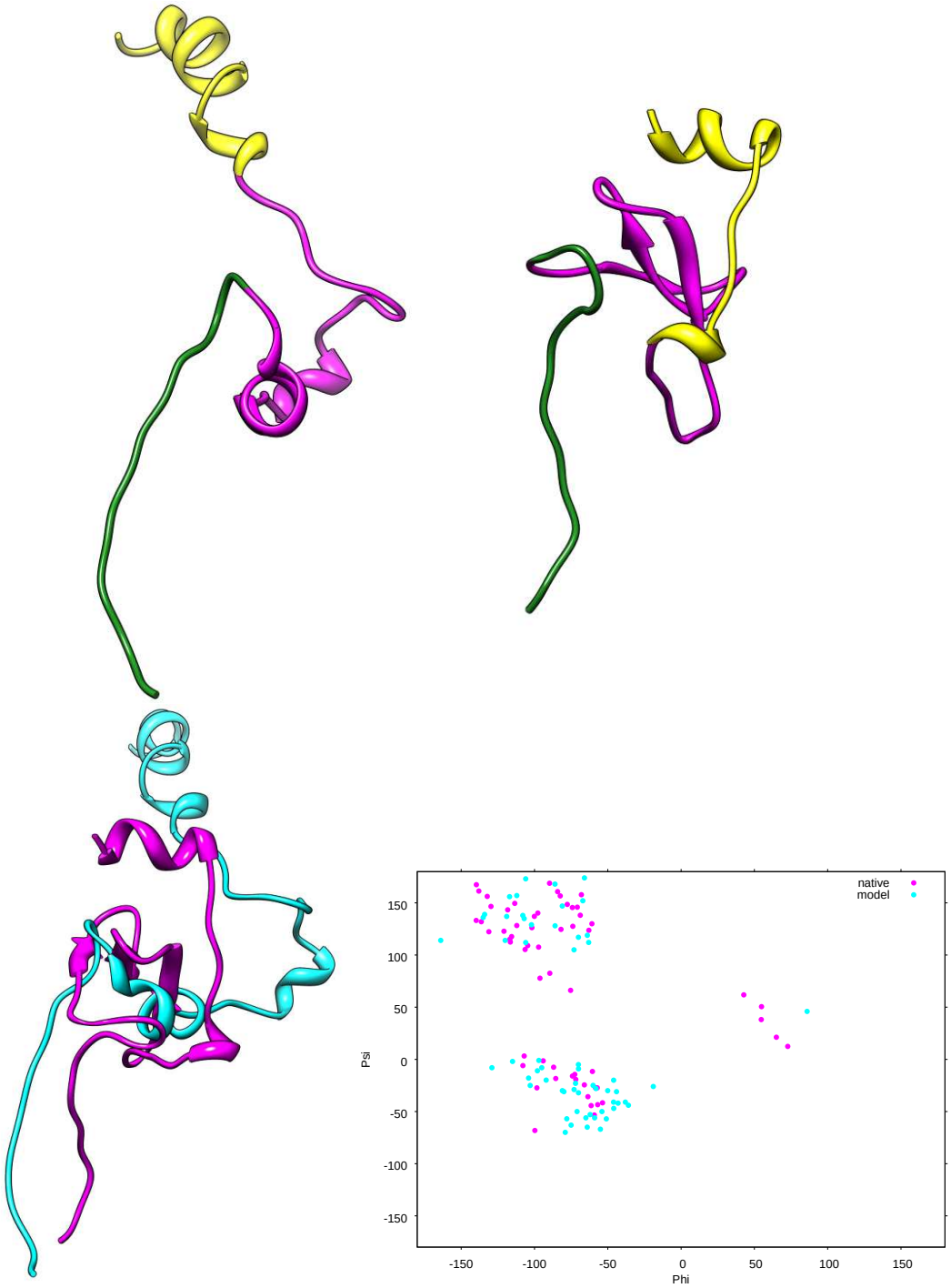


Figure 2.14: *Target 2HF1 - Top right: native, top left: model, bottom left: superimposition of native and model structures (colored magenta and cyan respectively), and bottom right: dihedral angles (ϕ, ψ) of native and model. Radius of gyration: model = 384.5 Å, and native = 229.3 Å. Simulation parameters: 200,000 unbiased move steps using 3-residues fragments.*

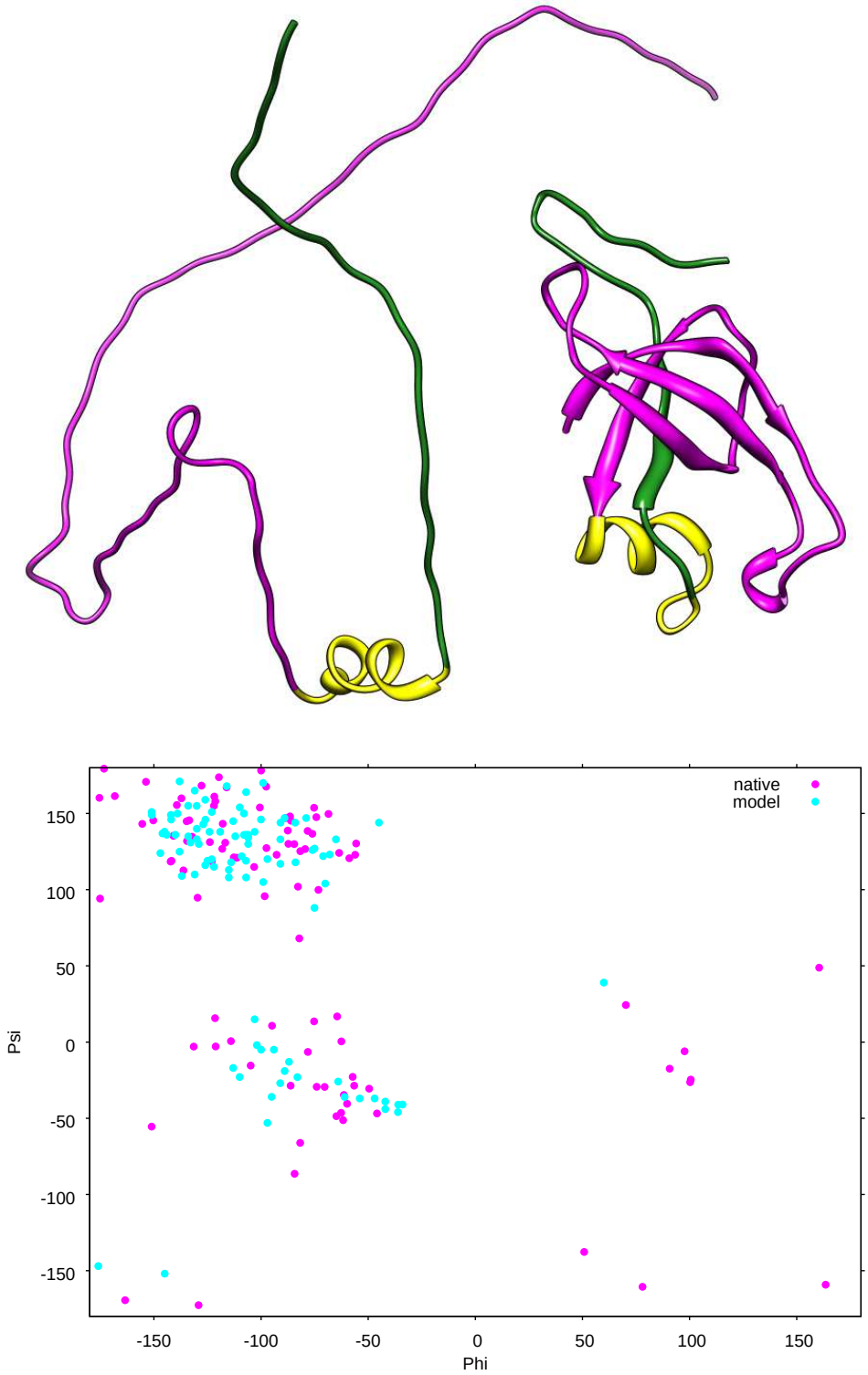


Figure 2.15: *Target 2HD3 - Top right: native, top left: model, and bottom :dihedral angles* (ϕ, ψ) *of native and model. Radius of gyration: model = 1022.5 Å, and native = 420.6 Å. Simulation parameters: 500,000 unbiased move steps using 3-residues fragments.*

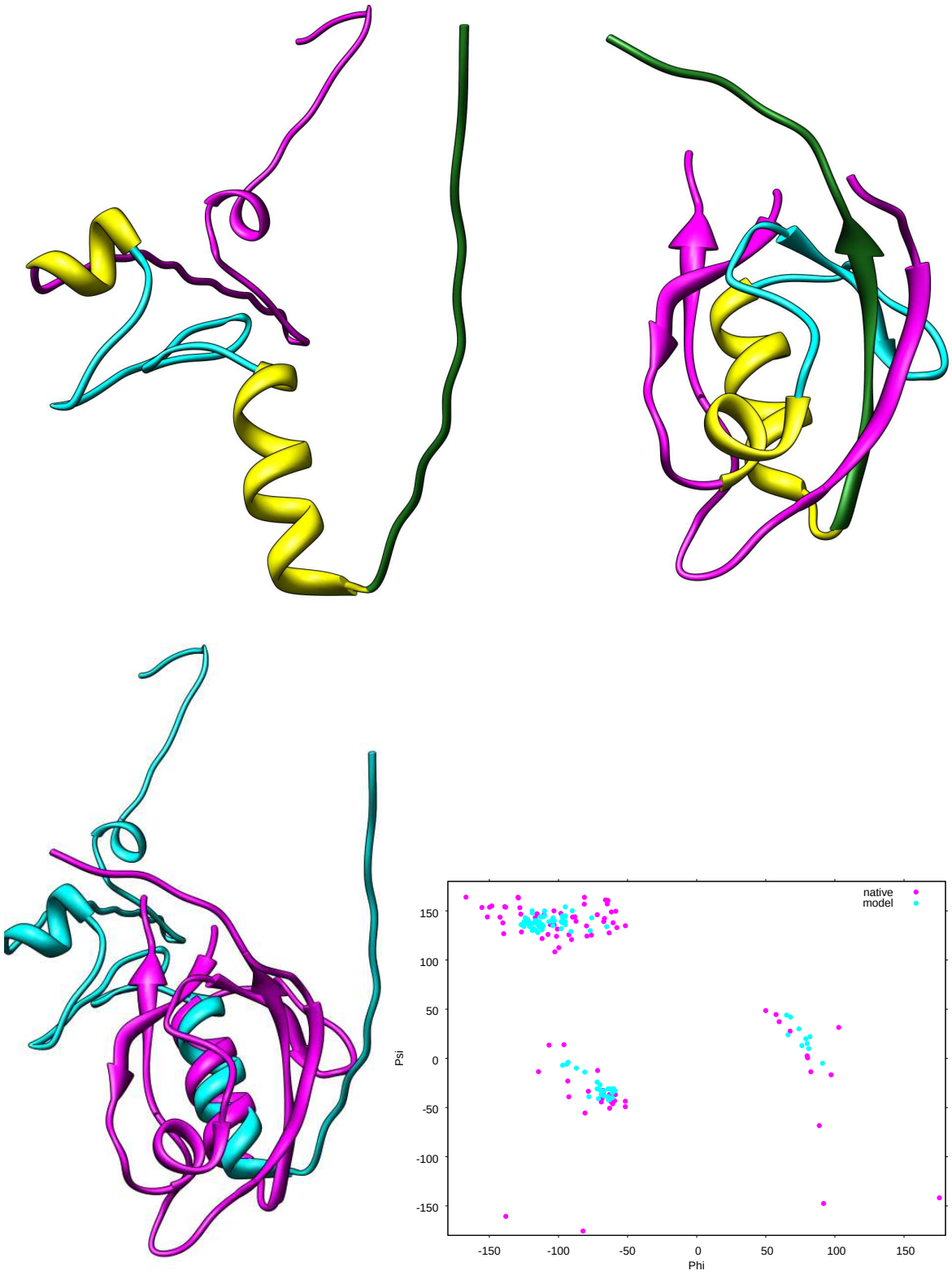


Figure 2.16: *Target 2GZV - Top right: native, top left: model, bottom left: superimposition of native and model structures (colored magenta and cyan respectively), and bottom right: dihedral angles (ϕ, ψ) of native and model. Radius of gyration: model = 805.9 Å, and native = 369.5 Å. Simulation parameters: 200,000 unbiased move steps using 1-residue fragments for moves.*

2HJ1

Model for the target 2HJ1 (given at the top left in figure 2.17) is the most compact structure among all the models presented before. The radius of gyration for this model is equal to 447.6 Å. Among secondary structures, only helix (colored yellow) and strands of beta sheets (colored green and cyan) have correct predictions. Model's dihedral angles (given at the bottom right in figure 2.17) also show close relationship to those of native. This model's prediction involves 200,000 attempted moves with 5-mer fragments by the search method.

All models discussed earlier were generated through unbiased moves of the search method's move set. Biased moves showed very strong biased towards helices. Every target whether it was made of helices or beta sheet turned into a helical structure. Model of 2HF1 target show in figure 2.18 represents an example of such predictions using biased moves.

In a typical prediction simulation, the probabilistic score of an evolving conformation continuously improves while the search method makes biased/unbiased moves through the conformational space. The final (model) conformation often has a probabilistic score higher than 1.0 (as shown in probability versus annealing temperature plot at the bottom in figures 2.17 and 2.18). The score is higher than 1.0 because the acceptance criterion requires non-normalized probabilities of overlapping fragments of the conformations.

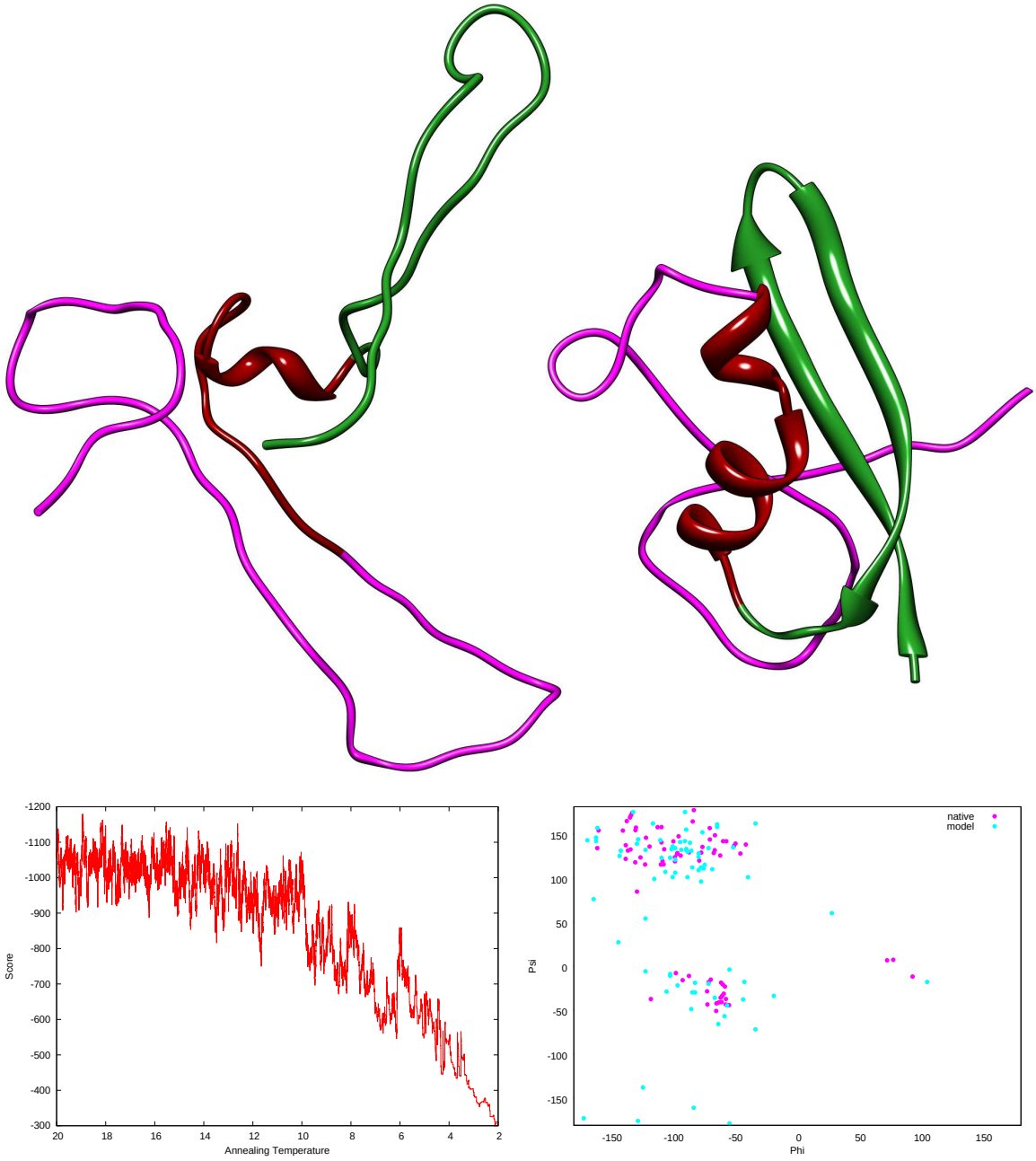


Figure 2.17: *Target 2HJ1 - Top right: native and top left: model, bottom left: probability of model structure versus annealing temperature, and bottom right: and bottom right: dihedral angles (ϕ, ψ) of native and model. Radius of gyration: model = 447.6 Å, and native = 336.3 Å. Simulation parameters: 200,000 unbiased move steps using 5-residues fragments.*

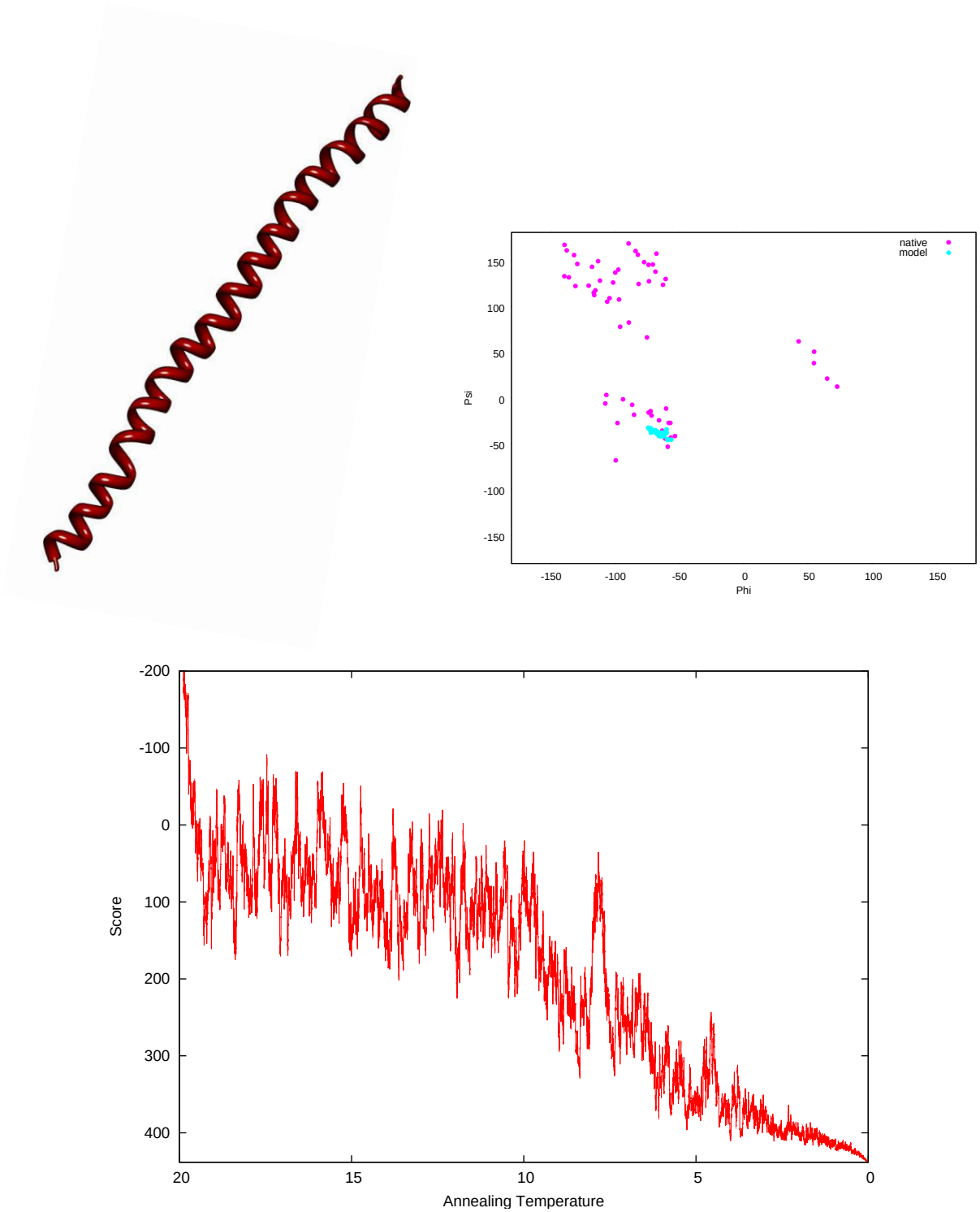


Figure 2.18: *Target 2HF1 - Top left: model, and top right: dihedral angles of native (shown at top right of figure 2.14) and model, and bottom: probability of model structure versus annealing temperature. Radius of gyration: model = 478.2 Å, and native = 229.3 Å. Simulation parameters: 200,000 biased move steps using 1-residue fragments.*

Discussions

In this part of work, an *ab initio* structure prediction method was set up. It relies on a score function which is purely probabilistic and has nothing to do with the Boltzmann statistics. Previously, this score function has successfully been used for protein sequence-structure and structure-structure alignment (Schenk et al. 2008) and protein threading (Torda et al. 2004). Two main terms i.e. sequence and structure of the score function are entirely dependent on probability distributions generated by Bayesian classification of protein fragments. The search method, simulated annealing Monte Carlo, has an acceptance criterion entirely based on the conformational probabilities. Initially, move set of the search method consisted of two kinds of moves: 1) biased and 2) unbiased moves. The prediction models presented in the previous section demonstrate that this purely probabilistic score function with simulated annealing Monte Carlo as a search method has an ability to build three-dimensional structures of target sequences from scratch. Although the generated models are not very close to their natives, but they essentially demonstrate the strengths and the weaknesses of the score function and the search method.

The most encouraging aspect which one learns from the given results is that the score function can guide its search method towards states where conformations look like protein structures. Obviously, the generated models are far from being perfect but they have good predictions about secondary structures. As the score function is built through the classifications of N-mer protein fragments of the known structures in a set of classes. In the found classes, each descriptor has its own distribution and the probability of an unknown fragment is computed as a mixture of the probabilities of those descriptors across all classes of the classification. Since each proposed conformation is considered as a set of overlapping fragments by the score function, it has very good understanding of the secondary structures through the local interactions of those fragments. The length of constituent (overlapping) fragments depends upon the length of fragments used for classification. Different classifications built with 4-mer, 5-mer and 6-mer fragments were used to generated models for the targets. It was observed that the classifications built with large fragments produce better models. It is worth mentioning that the computational cost to build a classification is considerably increased with an increase in the length of the fragments.

One of the main weaknesses which one can notice from the generated models is lack of compactness in their structures. Almost all models have extended conformations because the score function could not guide the search by distinguishing between compact and less compact states of the system. This behavior was expected to some extent as the

score function had no mechanism in it to figure out the hydrophobic and hydrophilic features of the constituent fragments of a protein conformation. The other issue with the score function has been the the method for (conformation) probability calculation (see figure 2.5). In this method, probabilities of the adjacent fragments are too much dependent on each other and the fragments suspectedly are over-influenced by the occurrence of their neighbor fragments. Therefore, there was a need to introduce some new balance methods for probability calculations.

Both incorporation of sovlation feature into the score function (section 3.1) and the implementation of new probability calculation methods (section 3.2.2) were achieved in second part of this work described in chapter 3.

The search method, simulated annealing Monte Carlo, has equally important role in producing the protein-like models of the targets (given in result section). It generates these models by not taking the Boltzmann statistics into account and relying its acceptance criterion entirely on a ratio of the probabilities derived from probabilistic distributions of the selected descriptors (i.e. sequence and structure). As the score function does not involve any kind of physics of protein structures directly, the search method has no real temperature but an artificial scheme to control the smoothness of conformational states. This scheme has an arbitrary temperature. For the generated models shown in result section, the search method started with a high temperature of 20.0 and was gradually lowered down to ≈ 0.0 while cooling the system down. The acceptance criterion is designed in such a way that at high temperature, it has more even distribution of states and at lower temperature it prefers more probable conformational states.

Both biased and unbiased moves in the search method's move set were used to make predictions against all the targets in the data set but only unbiased moves could produce better prediction models whereas biased moves always end up into helical models (as shown in figure 2.18). As the helical protein structures are more abundant among the known protein structures than non-helical ones, this fact provides the reason of bias towards helices (in the fragment library for biased moves and to some extent in Bayesian classification as well). Therefore, any prediction which starts with an initial random conformation is quickly pushed to fold into a helical structure by assigning relatively higher probabilities to the frequently extracted helical fragments from the fragment library. This kind of behavior is evident from figure 2.18 (at the bottom) where probabilistic score of a random conformation went up quickly after starting few moves by the search method. The rate at which an initial random conformation gets folded into a helical structure is also influenced by the size of fragment used by biased moves. The moves through larger fragments lead more quickly to the helical conformation.

On the search method side, the second part of this work (described chapter 3) was

set to accomplish two main objective: 1) removal of bias in the fragment library driven moves to make them nearly ergodic, and 2) the extension of move set so that the conformational space could be explored more efficiently with an upgraded score function.

Chapter 3

Introducing Solvation and Hydrogen Bonding

Solvation

Water (solvent) plays an active role in folding, stability and function of protein structure (Dill 1990, Sharp et al. 1991, Finney 1996, Bogan and Thorn 1998, Hummer et al. 2000, Covalt et al. 2001, Cheung et al. 2002, Pratt and Pohorille 2002, Harano and Kinoshita 2004). In protein folding, water has a fundamental role in defining hydrophobic attractions (Kauzmann 1959, Dill 1990) and bringing the hydrophobic residues together (Levy and Onuchic 2004). Patterns of hydrophobicity exhibited by the residues are given a special consideration in the prediction of protein structures from their sequence data (Eisenberg et al. 1984).

The effect of water molecules may be treated explicitly in a detailed microscopic force field. However, in practice, such a treatment which involves interactions between solvent and solute in a high dimensional configuration space cannot be realized due to lack of required computational power (Tadashi et al. 2002, Jaramillo and Wodak 2005, Sagui and Darden 1999). That is why, models have been developed which incorporate the influence of solvent implicitly (Edelsbrunner and Koehl 2005). There are two basic types of implicit solvent models:

1) empirical models in which solvation free energy is assumed as a sum of atom or group contributions. Individual contributions are functions of either solvent accessible surface area (Lee and Richards 1971, Eisenberg and McLachlan 1986, Ooi et al. 1987, Wesson and Eisenberg 1992) or volume of a solvent shell (Gibson and Scheraga 1967, Kang et al. 1988, Colonna-Cesari and Sander 1990). These models incorporate only hydrophobic and electrostatic aspects of solvation.

2) continuum electrostatics based models where the solvent and solute interior are defined by different dielectric constants and the solvation free energy is computed by solving the Poisson-Boltzmann equations (Jaramillo and Wodak 2005). The continuum electrostatic models also take into account screening effects of interactions between charges.

As the score functions for protein structure prediction consist of a combination of terms, for instance terms for sequence, structure, etc., an additional solvation term may be introduced to make better predictions about native-like structures by improving the accuracy of medium to long range interactions. The protein molecules contain both hydrophobic and hydrophilic parts residing at their surface and interior, respectively, (Kauzmann 1959). The solvation terms are usually based on some knowledge-based potential functions with main focus on non-polar effect, also referred as hydrophobic effect, of water (Papoian et al. 2004, Levy and Onuchic 2006, Baker and Sali 2001).

Solvation in Bayesian Framework

In order to incorporate a solvation term into a probabilistic score function, the preliminary task one would have to do is to figure out a measure of hydrophobic effect. Such a measure should be simple as well as statistically consistent with the existing terms of the score function. The quantification of hydrophobic effect by the concept of solvent-accessible surface (Lee and Richards 1971, Eisenberg and McLachlan 1986) shows that the hydrophobic atoms (or residues) have lesser accessible area than hydrophilic atoms. It has also been shown earlier that the number of residues observed in a sphere of certain radius around a reference residue is related to the exposure of latter to solvent (Melo et al. 2002).

To introduce solvation in our probabilistic score function, we also wanted to know whether a measure of solvation we are interested in could be used as an effective statistical descriptor in Bayesian framework. For that purpose, the solvation effect of 20 residues of protein was averaged out from the known protein structures in PDB. Each of 20 residues in protein structures was put under solvation spheres of radii: 8, 10, 12 Å by fixing their centers at the C_β atoms of residues and then counting the number of neighbor C_β atoms falling inside each solvation sphere. Figure 3.1 illustrates how solvation spheres of different radii were used to count neighbor C_β atoms of a residue. The number of C_β atoms (except one at the center of solvation sphere) were used to generate histograms of each of 20 residue types (shown in figures 3.19, 3.20, 3.21 and 3.22).

Histograms of neighbor C_β atoms of 20 residue types (shown in figures 3.19, 3.20, 3.21 and 3.22) apparently exhibit nearly normal distributions. Means and standard deviations over the count of neighbor C_β atoms (within solvation spheres of radii: 8, 10, and 12 Å) of all 20 residues are given in table 3.3. According to the statistics given in table 3.3, hydrophilic residues: Asp, Glu, Asn, Pro, Gln, and Lys have the lowest mean of neighbor C_β count inside 10 Å solvation sphere i.e. 8-9 (also see figure 3.19). On the other hand, mean of C_β count (within 10 Å solvation sphere) for hydrophobic residues: Ala, Cys, Phe, Leu, Ile, Met, Val, Trp, and Tyr was recorded to be 13-16 (histograms given in figures 3.21 and 3.22). Similarly, Gly, Arg, Thr, Ser and His exhibited 10-12

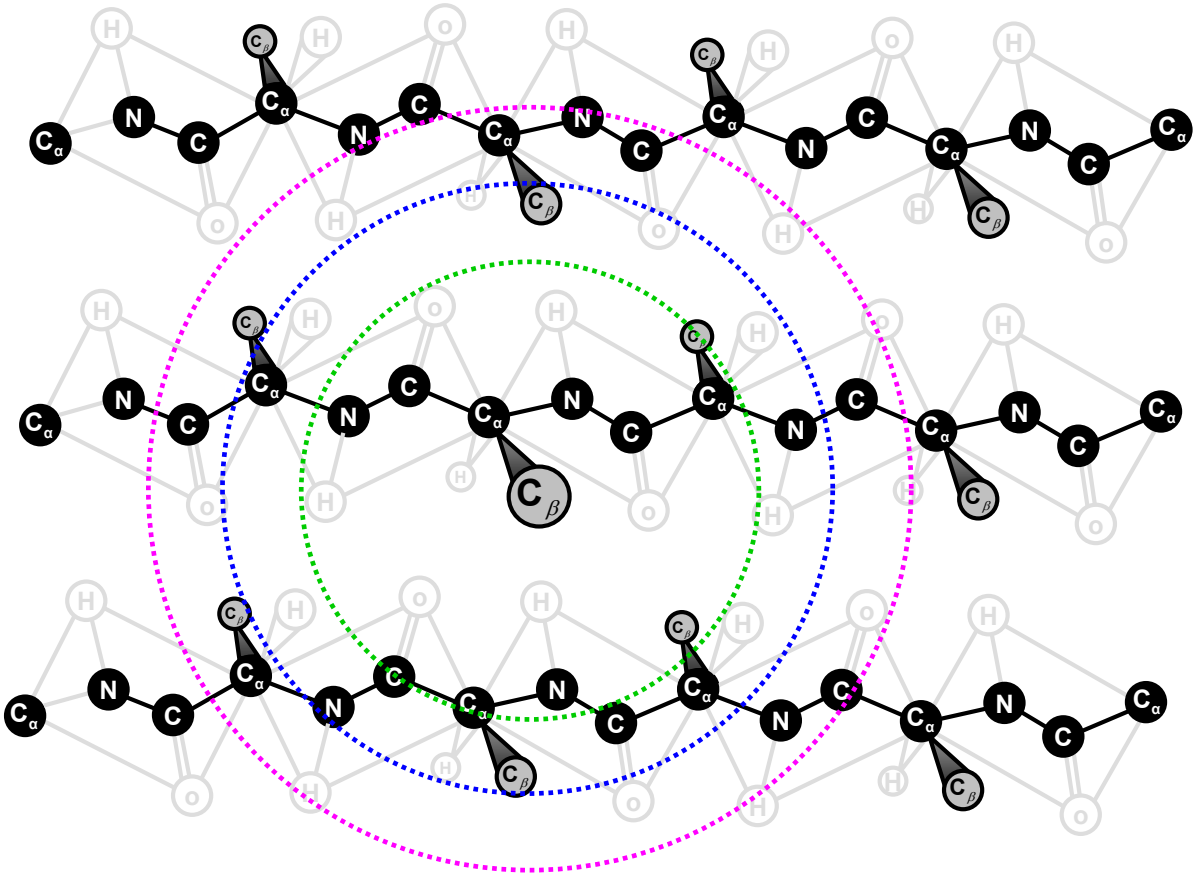


Figure 3.1: *Solvation sphere: the solvation features of each residue of a fragment were taken into account by counting the number of neighbor C_β atoms around its own C_β atom. To determine a meaningful spread of neighborhood, neighbor C_β atoms of each of 20 residues (of protein structures) in the PDB were counted within distances of 8, 10 and 12 Å from their own C_β atoms. The solvation spheres, colored green, blue, and magenta, basically represent the cutoff distances of 8, 10 and 12 Å, respectively. The histograms of the C_β neighbor counts of 20 residues are shown in the figures 3.19, 3.20, 3.21, and 3.22. For all 20 residues, means and standard deviations over the count of neighbor C_β atoms are given in table 3.3.*

(mean) C_β neighbors in 10 Å solvation sphere (see histograms in figure 3.20).

Clearly, these findings provide an opportunity to incorporate solvation effect, purely on statistical basis, into our probabilistic score function.

Data Collection

Again, the solvation enabled score function was trained over fragments of the known protein structures present in the PDB (Berman et al. 2000) but this time, in addition to sequence and structure (ϕ, ψ), the commonly seen information of fragments also includes their solvation features. Same 1.5×10^6 fragments generated from already known protein structure in the PDB (and described in section 2.2.2) were used to build a new classification but this time the attribute vectors of these fragments were assigned an additional set of values representing the measurement of solvation features of the fragments. See figure 3.2. The solvation measurement values of an attribute vector are the number of neighboring C_β atoms each residue of a fragment has within a solvation sphere of a certain radius (e.g. 8, 10, 12 Å).

Model

Sequence and structural features of protein were modeled by multi-way discrete and bivariate Gaussian distributions, respectively, as described in section 2.2.3. Solvation statistics of 20 residue types (given in figures 3.19, 3.20, 3.21, and 3.22 and table 3.3) generated under a solvation sphere of 10 Å radius suggest that solvation effect can be modeled with simple Gaussian distributions of the form given in equation 2.6. In equation 2.6, x_i, μ, and σ are the vectors consisting of solvation values of each residue of a fragment, mean and standard deviation, respectively.

Re-Classification

A number of classification were built with varying parameters (e.g. different lengths of overlapping fragments and solvation spheres). Though 10 Å solvation sphere seems to be appropriate for measurement of solvation features, classifications with solvation spheres of 8 and 12 Å radii were also built to conduct a comparative analysis of all of them and to see whether the obtained results are in agreement with the preliminary statistics (given in figures 3.19, 3.20, 3.21, and 3.22 and table 3.3).

The iterative steps (of class memberships and class parameters calculations) of EM convergence to build these classification took longer than the one described in section 2.2.4 due to addition of an extra descriptor (about solvation features of protein fragments) to the attribute vectors. The resulted probabilistic classifications (built by observing sequence, structural and solvation features of 1.5×10^6 fragments of the known

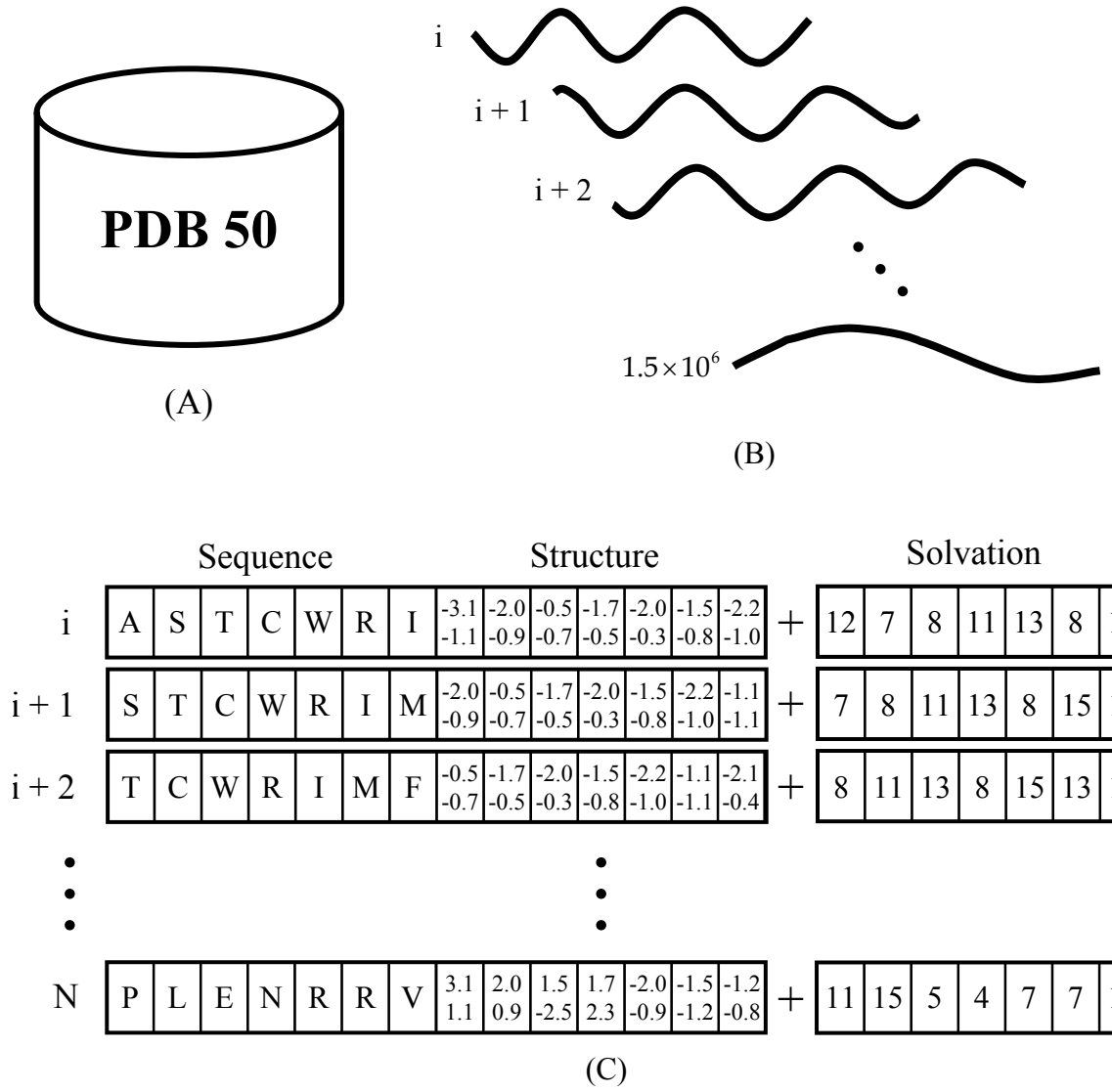


Figure 3.2: *(A) A database of protein chains in which no two proteins have sequence identity of more than 50%, (B)* 1.5×10^6 *overlapping fragments extracted from 50% sequence identity protein chains and (C) attribute vectors of the fragments shown in (B). Each attribute vector consists of sequence, structure* (ϕ, ψ) *and solvation information of the corresponding fragments. The solvation descriptor consists of the number of neighbor* C_β *atoms which each residue of a fragment has within a solvation sphere of a certain radius (e.g. 10 Å).*

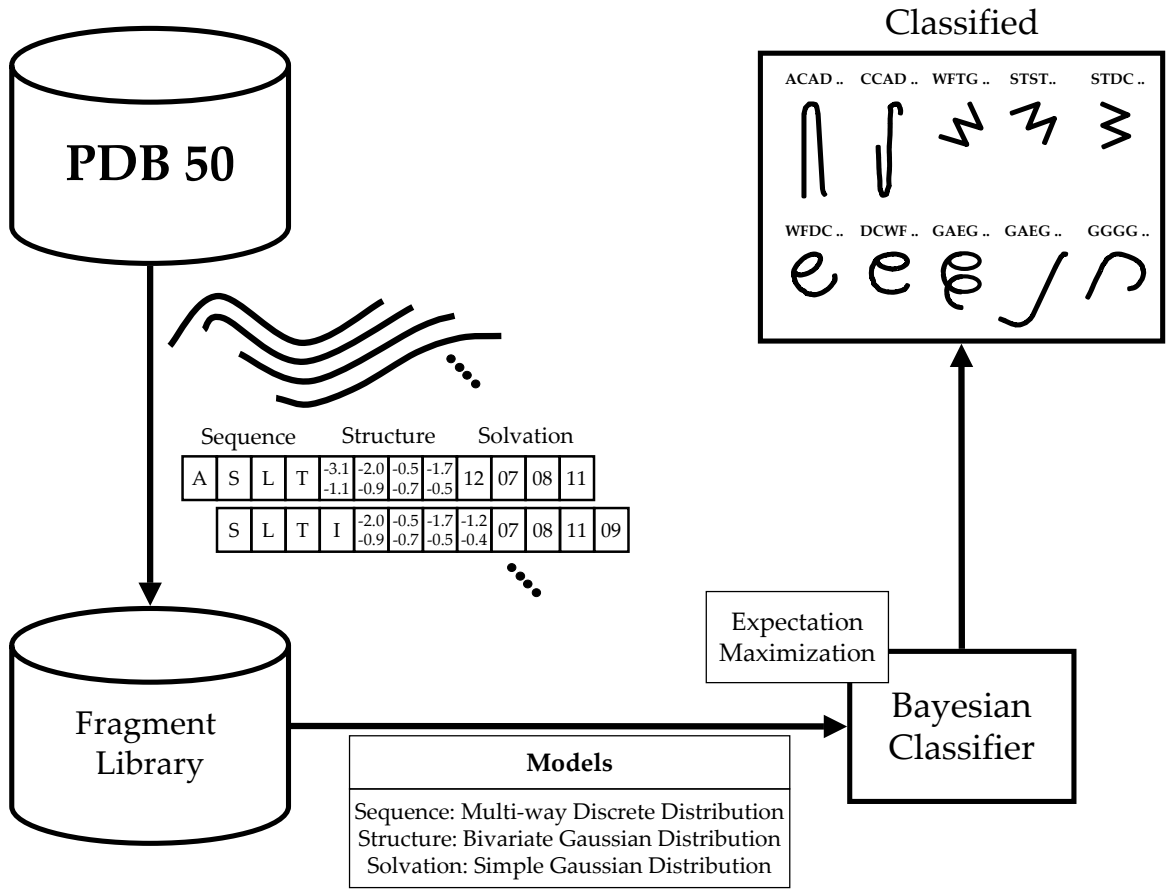


Figure 3.3: *Bayesian classification of* 1.5×10^6 *protein fragments. The fragments are represented by attribute vectors consisting of sequence, structure (*ϕ*,* ψ*) and solvation information of the fragments. The classification finds a probabilistic description of the most probable set of classes.*

protein structure) allow us to calculate the statistical estimate of structural and solvation features of protein sequences whose structures have not been solved yet.

Hydrogen Bonding

Theory

Hydrogen bonds (Huggins 1971) are weaker but one of the most important inter-atomic interactions. In protein folding, hydrogen bond patterns are thought to play an essential role in the formation of secondary structure elements. An ideal hydrogen bond, both theoretically and experimentally, is a bond consisting of four atoms: donor heavy atom (N), the hydrogen (H), the acceptor lone electron pair (O) and the acceptor (C) lie in the same line (McDonald and Thornton 1994) (as shown in figure 3.4).

In proteins, most of the hydrogen bonds are made between main-chain NH and

CO. The bonds between main-chain and side chains are of least interest because in our reduced representation of protein structure the side chains beyond C_β atoms are not entertained. In the following, we have described an electrostatic model that was used in our code to calculate hydrogen bonds of protein conformations.

Electrostatic Model

Two pairs of hydrogen bonding atoms (C, O) and (N, H) have partial charges (+q_1, -q_1) and (-q_2, +q_2), respectively. The electrostatic interaction energy of a bond between the two pairs of atoms is calculated by

$$E_{hb} = q_1 q_2 \left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right) \cdot f \tag{3.1}$$

where r(AB) is an inter-atomic distance between atoms A and B, and f is a dimensional factor to convert energy E_{hb} into kcal/mol and its value is 332. The values of partial charges q_1 and q_2 are $0.42e$ and $0.20e$, respectively. e is the unit electron charge.

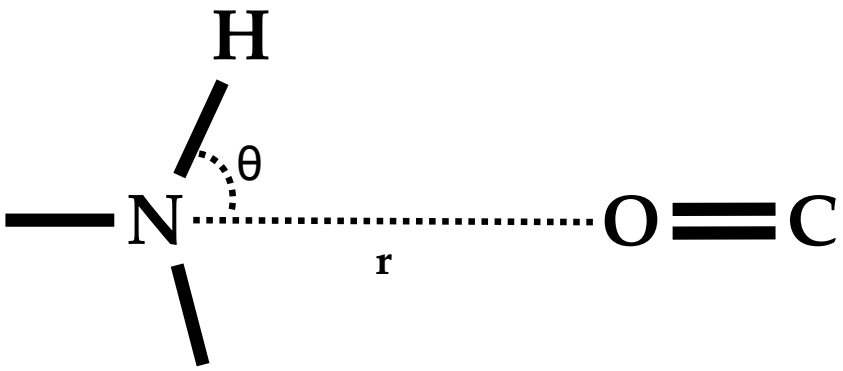


Figure 3.4: *Geometry of hydrogen bond: an ideal hydrogen bond has distance r = 2.9 Å, $\theta = 0°$, and electrostatic interaction energy E_{hb} = -3.0 kcal/mol. Generally, a bond with E_{hb} = -0.50 kcal/mol is considered a meaningful hydrogen bond. In simple hydrogen bonds, N-O distance up to r = 5.2 Å and misalignment of up to $\theta = 63°$ are allowed.*

As shown in figure 3.4, energy E_{hb} is determined by H-N-O angle (θ) and N-O distance (r). An ideal hydrogen bond with energy E_{hb} equal to -3.0 kcal/mol corresponds to distance r = 2.9 Å and $\theta = 0°$. However, it is justifiable to keep the cutoff energy of -0.50 kcal/mol (corresponding to θ below $63°$ and distance r equal to 2.5 Å or below 5.2 Å for $0°$ H-N-O alignment) to also take weaker or slightly misalign hydrogen bonds into consideration (Lifson et al. 1979, Margulis et al. 2002).

Sampling with Solvation enabled Score Function

Sampling in probabilistic space $\mathbf{x}$ after addition of dimensions of solvation is still performed according to Monte Carlo scheme outlined earlier in section 2.3. The N random moves $\mathbf{x}_i = \{i = 1, 2, \cdots, N\}$ in space $\mathbf{x}$ are accepted or rejected according to Metropolis criteria (given in equation 2.28). The ratio of probabilities of conformations $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$ is given by

$$\mathcal{P}_i = \frac{p\left(\mathbf{x}_i\right)}{p\left(\mathbf{x}_{i-1}\right)} \cdot \exp\left(-\frac{\Delta E_{hb}}{k_{\mathrm{B}}T}\right) w \tag{3.2}$$

where $\exp(-\Delta E_{hb}/k_{\mathrm{B}}T)$ is the hydrogen bonding term, ΔE_{hb} is the hydrogen bond energy difference, k_{B} is the Boltzmann constant and T is the annealing temperature, and w is a weighting factor for hydrogen bonding.

Let F be the number of overlapping fragments of conformation $\mathbf{x}_i$ (as shown in figure 2.5), and J be the number of classes generated by classification, then in equation 3.2 the probability $p(\mathbf{x}_i)$ of conformation $\mathbf{x}_i$ is given by

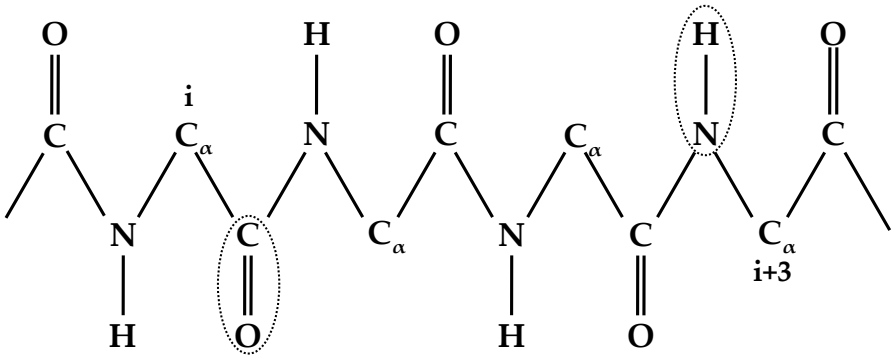
$$p\left(\mathbf{x}_i\right) \;=\; \left(\prod_{f=1}^{F}\sum_{j=1}^{J} \alpha_{j|\mathcal{F}_f^{seq}} p_j\left(\mathcal{F}_f^{struct,solv}\right)\right)^{1/F} \tag{3.3}$$

where $\alpha_{j|\mathcal{F}_f^{seq}}$ is the weight of class j given the sequence $\mathcal{F}_f^{seq}$ of fragment f, $\mathcal{F}_f^{struct,solv}$ represents structural (ϕ, ψ) and solvation (neighbor C_β atoms) features of the fragment f, and $p_j(\mathcal{F}_f^{struct,solv})$ is the probability of the structural arrangement of fragment f with sequence $\mathcal{F}_f^{seq}$.

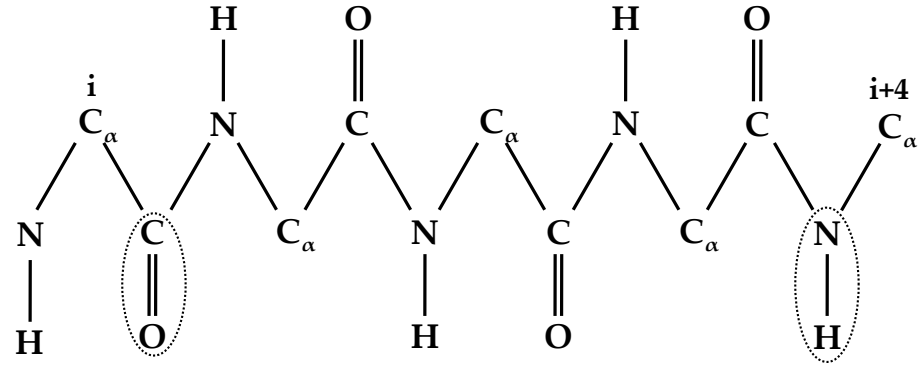
Filtering Hydrogen Bonds

According to the electrostatic model described in section 3.1.2, the straight forward approach to calculate the hydrogen bond energy of a protein conformation would involve two steps: 1) evaluation of individual hydrogen bonds of each residue of conformation with the rest of residues, and 2) then sum of energies of valid hydrogen bonds. This approach of hydrogen bond energy can possibly introduce structural clashes into a conformation by the formation of lone or/and random hydrogen bonds between any two residues of conformation. That is why, we had to take elementary hydrogen bond patterns into account for the calculation of hydrogen bond energy of protein conformations.

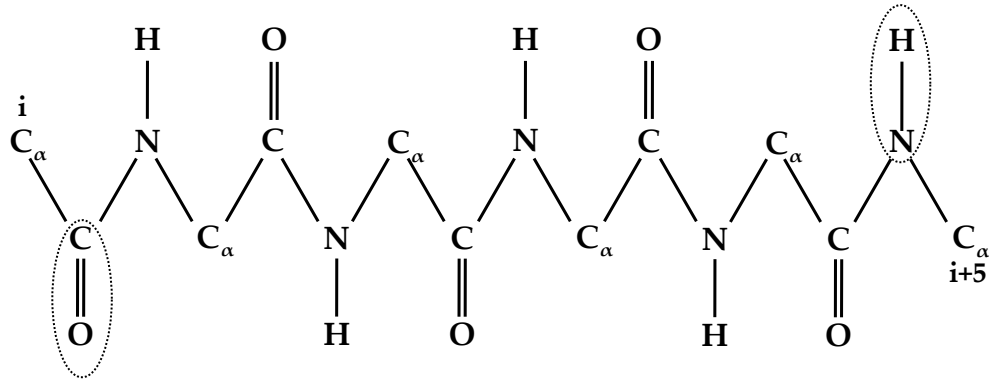
Our approach to calculate hydrogen bond energy of a protein conformation works in two stages: In first stage, elementary hydrogen bond patterns, n-turns and bridges, are identified. An n-turn of type $(i, i+n)$ is a simple hydrogen bond from $CO(i)$ to $NH(i+n)$ where $n = 3, 4, 5$ (see figure 3.5). A parallel or anti-parallel bridge, depending



(A) 3-turn

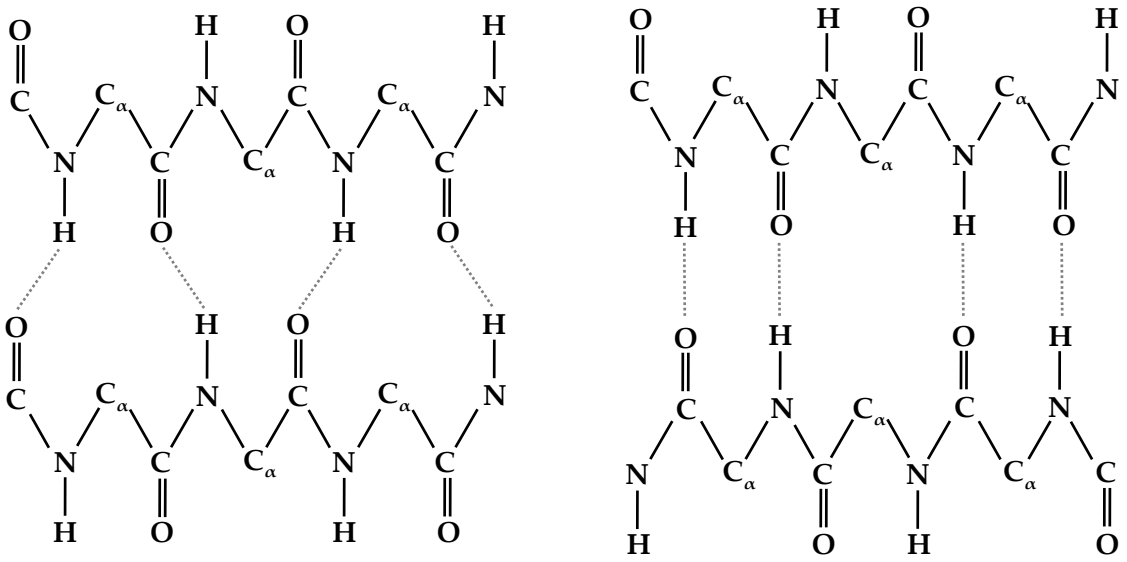


(B) 4-turn



(C) 5-turn

Figure 3.5: *n-turns, elementary hydrogen bond patterns. (A) 3-turn: hydrogen bond from $CO(i)$ to $NH(i+3)$, (B) 4-turn: hydrogen bond from $CO(i)$ to $NH(i+4)$, and (C) 5-turn: hydrogen bond from $CO(i)$ to $NH(i+5)$.*



(A) Parallel bridge **(B) Anti-parallel bridge**

Figure 3.6: *Bridges, elementary hydrogen bond patterns. (A) parallel bridge: between two 4-residues long non-overlapping parallel stretches, and (B) anti-parallel bridge: between two 4-residues long non-overlapping anti-parallel stretches.*

upon basic patterns, is formed between two 3-residues long non-overlapping stretches as shown in figure 3.6. In second stage, cooperative hydrogen bond patterns, helices, β-ladders and β-sheets, were defined from the elementary patterns of the first stage. At least two consecutive n-turns are required to define a minimal helix. The overlaps of minimal helices are used to defined longer ones. A set of one or more bridges is used to define a ladder whereas a sheet is defined by one or more ladders connected through some shared residues (Kabsch and Sander 1983). The total hydrogen bond energy of a conformation is the sum of energies of those hydrogen bonds which are part of helices and sheets of that conformation.

Probabilities Calculation Methods

A straight forward method for the probability calculation of a conformation $\mathbf{x}_i$ proposed by the search method is to convert conformation $\mathbf{x}_i$ into $F = (R - L_f)$ overlapping fragments where L_f is the length of each fragment and R is the number of residues of conformation $\mathbf{x}_i$, and then to calculate a sum of the probabilities of each fragment across all the classes of probabilistic classification. The product of F sums gives an overall probability of the conformation $\mathbf{x}_i$. We call this method 'crude' probability calculation method. Graphical illustration of this method is depicted in figures 2.5 and 3.7, whereas its mathematical description is given in equation 3.3.

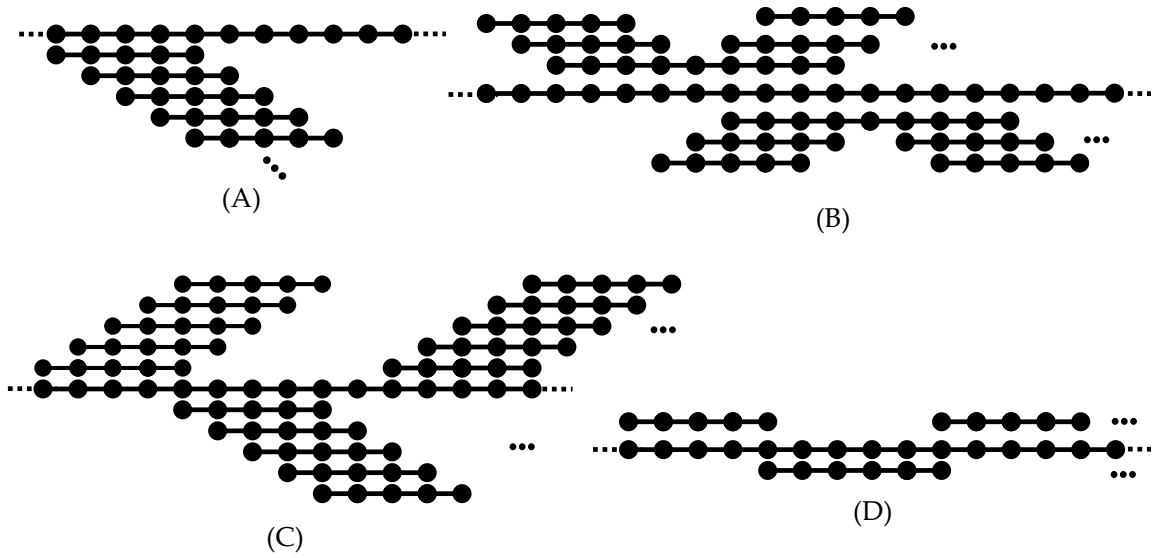


Figure 3.7: *(A) crude, (B) centered average, (C) average, and (D) simple.*

After the extension of probabilistic space by addition of solvation dimensions to it, the probability of the conformation $\mathbf{x}_i$ calculated by 'crude' method is often so different from that of $\mathbf{x}_{i-1}$ that it is very unlikely to find native-like conformations of any protein sequence. This difference is caused by the increased interdependence of neighboring fragments. The insertion of a new fragment at some location of the existing conformation does not only assigns neighbors to that fragment but the neighbors of the (existing) neighboring fragments are also changed depending upon the shape of the new fragment. To deal with this problem, new methods of probability calculation were introduced: centered-average, average, and simple. While exploration of the conformational space (for a native-like conformations) by the search method, these methods keep conformational probabilities (between any two consecutive steps) smooth.

In 'centered-average' probability calculation method (shown in figure 3.7:B), the probability of each of $F = (R/L_f)$ non-overlapping fragments of conformation $\mathbf{x}_i$ is calculated by taking an average of the probabilities of two overlapping fragments (generated earlier in the crude method) in forward direction, two overlapping fragments in backward direction and the non-overlapping fragment itself. The product of centered-average probabilities of F fragments gives an overall probability of a conformation.

'Average' probability (figure 3.7:C) of conformation $\mathbf{x}_i$ is calculated by dividing it into $F = (R/L_f)$ non-overlapping fragments, and determining the probability of each fragment as an average of the probabilities of overlapping fragments. The corresponding overlapping fragments are generated according to the scheme described in crude probability calculation method. The probability of a fragment j spanning from residue

k to residue $(k + L_f)$ is an average of the probabilities of fragments from residue k to residue $(k + L_f)$, from $(k + 1)$ to $((k + 1) + L_f)$, $\cdots$, and fragment from residue $(k + L_f)$ to $((k + L_f) + L_f)$. The product of average probabilities of F non-overlapping fragments gives an overall probability of conformation $\mathbf{x}_i$.

Similarly, in 'simple' probability calculation method (shown in 3.7:D), a conformation $\mathbf{x}_i$ is divided into $F = (R/L_f)$ non-overlap fragments and the overall probability is simply a product of the probabilities of F non-overlapping fragments alone.

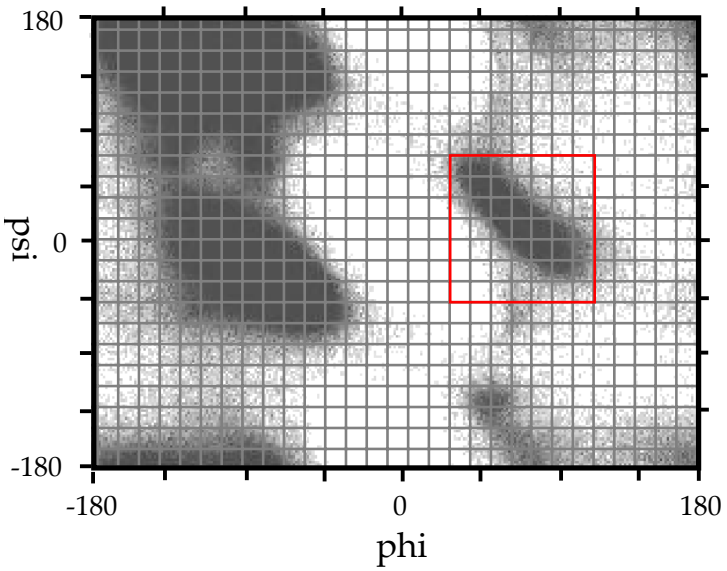
Biased Moves: Liking the Unlikely

Biased search moves described in section 2.3.3 are not completely random because dihedral angles to update conformations are derived from fragments drawn from a library and that library is generated from fragments of the known protein structures in the PDB. As a result, dihedral angles of those local structures, for instance helices, which are more abundant in protein structures are preferred over those of less abundant ones. In the following, this bias has been fixed to a great extent through a scheme which makes the acceptance of moves made with the less abundant dihedral angles rather easier than of those with abundant ones.

Dihedral angles (ϕ, ψ) of the known protein structures in the PDB can be used to draw a plot (shown in figure 1.2). The superimposition of an $n \times n$ grid on such a plot allows us to assign a box membership to each dihedral angle pair (ϕ, ψ) on the plot. Each box of the grid contains certain number of dihedral angle pairs as its members. Internally, we have a one-dimensional array M, consisting of $n \times n$ elements, where each element of it corresponds to a box of the grid and its value to the number of members (i.e. dihedral angles) in that box. Another one dimensional array P with the same number of elements as those of M and each element of P contains probability of having a dihedral angle pair (ϕ, ψ) in the corresponding grid box. An array A is used for the storage of accumulative probabilities. First element of A has probability of first box of the grid, second has sum of probabilities of first and second box of the grid, . . ., and last element is sum of probabilities of all boxes of the grid (i.e. 1.0) and it corresponds to $n \times n$th box. Figure 3.8 illustrates above described steps.

To make a search move, a random number r, uniformly distributed over the interval (0, 1), is generated and the index i of array A is propagated to the element where its values is greater than r. The corresponding index j of array P tells about the relevant box $b[j]$ of the grid and the probability of fragments which belong to the box b. A fragment f_j is randomly selected from box $b[j]$ to update dihedral angles of conformation $\mathbf{x}_{i-1}$ with those of fragment f_j. This gives us a new conformation $\mathbf{x}_i$.

To adjust the likelihood of acceptance, the probability of box $b[j]$ given by $p[j]$ is multiplied to the usual probability of acceptance. Hence, the equation 3.2 becomes



(A)

153	174	120	54	27	15	8
115	280	121	112	121	64	31
63	63	333	297	287	78	59
0	34	80	254	300	300	192
0	12	29	387	224	121	204
0	0	22	43	51	180	162
0	0	10	22	19	30	32

(B)

k	M
1	300
2	789
3	653
⋮	⋮
n-1	465
N	501

(C)

j	P
1	1.8e-4
2	4.7e-4
3	3.9e-4
⋮	⋮
n-1	2.8e-4
N	3.0e-4

(D)

i	A
1	1.8e-4
2	6.5e-4
3	1.0e-43
⋮	⋮
n-1	9.9e-1
N	1.0

(E)

Random Number Generator (0, 1)

Figure 3.8: *(A) Every dihedral (ϕ, ψ) angle pair of the existing protein structures in PDB has got a position on ϕ, ψ plot. The superimposition of an $n \times n$ grid on the plot gives a box membership to each angle pair, (B) part of grid to show that each box of the grids has a certain number of angle pairs as its members, (C) M is a one-dimensional array of size $N = n \times n$ where each element corresponds to a box of the grid and its value to the number of members that box has, (D) P is another 1-D array of the same size as that of M and each element of it is probability of having an angle pair in the corresponding grid box and (E) array A has accumulative probability where the first element of it has the probability of the first box of the grid, the second has the sum of the probabilities of the first and the second box of the grid, . . ., and the last element is the sum of the probabilities of all the boxes of the grid (i.e. 1.0) and it corresponds to $n \times n$th box.*

$$\mathcal{P}_i = \frac{p\left(\mathbf{x}_i\right)}{p\left(\mathbf{x}_{i-1}\right)} \cdot \exp\left(-\frac{\Delta E_{hb}}{k_{\mathrm{B}}T}\right) w \cdot p\left(\phi_{f_j}, \psi_{f_j}\right) \tag{3.4}$$

where $p(\phi_{f_j}, \psi_{f_j})$ is the probability of dihedral angles of fragment f_j.

'Controlled' moves

The challenges posed to the search method by increased degrees of freedom after incorporation of solvation effect into the score function led us to extend move set of the search method in order to better explore solvation-rich probabilistic space for native-like conformations of given targets.

When one inserts a (unbiased or biased) fragment into a conformation, maybe at the inner of the conformation, it does not only change the dihedral angles of that part of the conformation but adapts solvation features from its environment and also changes the solvation features (i.e. number of neighbor C_β atoms) of neighboring fragments. Without solvation effect, the insertion of a fragment was used to transmit only a smooth wave-like effect to adjacent fragments, basically because of overlapping character of probability calculation method. The score function with solvation term additionally causes a strong turbulent effect to probabilities of the neighboring fragments by changing their solvation features. Consequently, the overall probability of the conformation is drastically changed from one move step to another as the simulation proceeds.

To make the effect of a fragment insertion smoother, a new type of search moves, called 'controlled' moves, was introduced. In these moves, the dihedral angles (ϕ, ψ) of randomly selected residue position of a conformation are not updated through a biased or unbiased fragment but each of them is rotated up to $3 - 5^\circ$. Direction of rotation (clockwise or anti-clockwise) is decided by drawing a random number. These moves cause only a smooth and gradual change in probabilities of the neighboring fragments of a conformation. Since 'controlled' move steps are smaller, Monte Carlo simulation, of course, has to run for a longer period of time to find the probable conformation of given target sequence.

Results

Target Sequences

In addition to the target data set used in chapter 2 (table 2.1), some targets from the 8^{th} biannual CASP competition (CASP8) held in 2008 were also included. Table 3.1 has all the details of the targets selected from the CASP8 target list. These targets consist of both template-based or/and free modeling and purely template-based modeling candidates.

Table 3.1: *Details of the targets selected from the CASP8 target list.*

PDB ID	**CASP8 ID**	**CASP8 Category**	**No. of Residues**	**Seq**[1] (%)	**DSSP**[2]	
					Helix(%)	Beta[3](%)
2K3I	T437	TBM[4]	99	36	29	22
3D4R	T397	FM[5]	153	46	04	45
3D3Q	T416	FM	332	39	48	10
3DEE	T443	FM/TBM	248	42	43	11
3DED	T453	TBM	91	45	16	27
2K4N	T460	FM	111	25	28	27
3DFD	T465	FM	157	0	34	08
2K53	T473	TBM	68	33	61	02
2K5C	T476	FM/TBM	108	25	31	12
2K4X	T480	TBM	55	38	00	12
2K4V	T482	FM	120	30	28	32
3DO9	T496	FM	178	31	48	13
3DOA	T510	FM	228	24	26	26
3DUP	T513	FM	292	32	28	29

Out of total fourteen CASP8 target, eight targets (3D4R, 3D3Q, 2K4N, 3DFD, 2K4V, 3DO9, 3DOA, and 3DUP) are purely free modeling candidates, four targets (2K3I, 3DED, 2K53, and 2K4X) are template-based modeling candidates, and two targets (3DEE, and 2K5C) are candidates for both template-based and free modeling. The size and sequence similarity to the know structures of these targets ranges between 55-332 residues and 0-46%, respectively. According to DSSP secondary structure assignment, native structures

[1]Sequence similarity to the known protein structures
[2]Database of secondary structure in proteins
[3]Beta sheets
[4]Template-based modeling
[5]Free modeling

of these targets have 4-61% and 2-45% of helix and beta sheet, respectively.

In addition to CASP7 and CASP8 targets, data set also includes few non-CASP target sequences (1AGT, 1FSV, 2HEP, 1ZMQ). These target sequences are rather smaller and their native structures have both helices (e.g. 1FSV, 2HEP, 1AGT) and beta sheets (e.g. 1ZMQ, 1AGT).

Table 3.2: *The radius of gyration of native and model and random structures.*

Target ID		**Radius of Gyration (Å)**		
PDB ID	**CASP ID**	**Random**[1]	**Model**	**Native**
2HF1	T348	252.3	221.0	229.3
2K4X	T480	280.7	183.2	284.1
2K53	T473	351.9	243.9	253.6
2K4N 1[2]	T460	1109.8	503.4	634.0
2K4N 2	T460	1010.5	511.4	634.0
2K5C 1	T476	634.7	448.0	375.8
2K5C 2	T476	889.5	612.1	375.8
2K5C 3	T476	1015.2	515.3	375.8
3DFD	T465	1815.7	800.5	651.8
1FSV	NA	91.6	69.9	75.4
2HEP	NA	227.7	129.1	142.2
1AGT	NA	257.1	109.1	102.2

Predictions: Non-CASP Targets

1FSV

1FSV is a relatively smaller non-free modeling target. One may call it as an easy target. Model for this target (shown at the top left of figure 3.9) is pretty close to its native structure with root mean square distance (RMSD) score equal to 3.514 Å. The radii of gyration for model and native structure were calculated to be 69.9 Å and 75.4 Å, respectively. See table 3.2) for radii of gyration. Model 's lower radius of gyration is an indication of the over-compactness of its structure.

To generate this model, the search method had to attempt 200,000 biased moves with 3-mer fragments. Hydrogen bond term of the score function remained enabled during this prediction.

[1]Randomly generated conformation of a target sequence.
[2]Model 1

2HEP

Target 2HEP has a helical hairpin like native structure (as show at the top right in figure 3.10). Model for target 2HEP (given at the top left in figure 3.10) has two of its helices predicted and placed correctly except a little loop connecting them. The score function 's tendency toward over-compactness possibly resulted into the overreaching of the two helices and the misplacement of loop between them. RMSD score calculated between model and native structures is equal to 4.04 Å. The radii of gyration of model and native were calculated to be 129.1 Å and 142.2 Å, respectively. Lower value of the radius of gyration indicates over-compaction of 2HEP model.

For 2HEP model, the search method had to make 500,000 attempted move steps using 3-mer residues fragments while keeping the hydrogen bond term of the score function switched on.

1AGT

The two non-CASP targets presented above are mainly helical in their structures. 1AGT is an interesting target because its structure consists of both helix and beta sheets. Model for 1AGT (shown at the top left in figure 3.11) has prediction of beta sheets at right place but these beta sheets are parallel than anti-parallel. The helix could not be prediction except one turn of it. The formation of beta sheets is significant when the score function does not have very precise term for long-range interactions. Radius gyration of 1AGT model (i.e. 109.1 Å) is comparable to that of native (i.e. 102.2 Å). The structural alignment of model and native, and dihedral angle plot are shown at the bottom in figure 3.11.

The prediction of 1AGT model involves rather longer simulation run of 1,500,000 attempted move steps by using 1-mer fragments. Hydrogen bond term of the score function was kept enabled during this prediction.

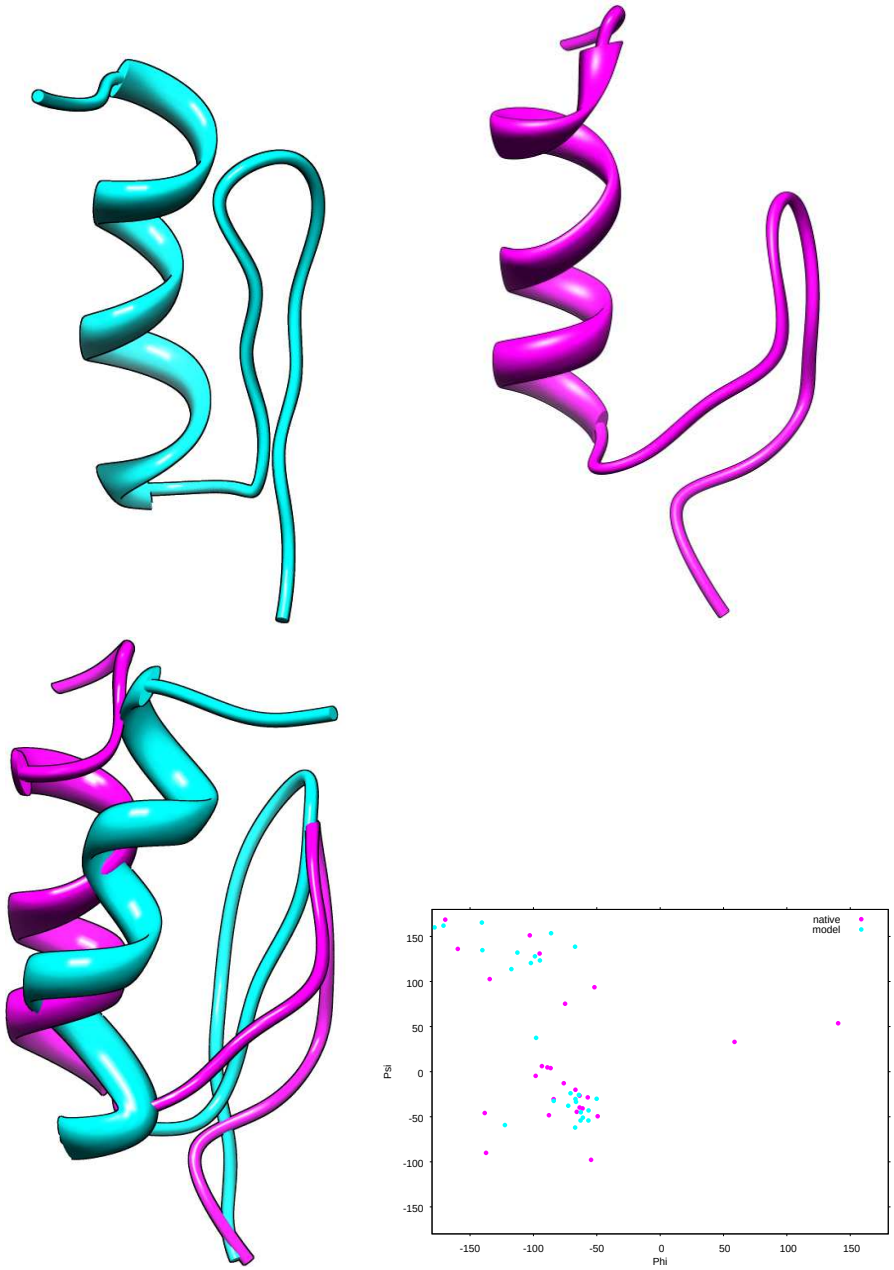


Figure 3.9: *Target 1FSV - top left: model, top right: native, bottom left: superimposition of native and model structures (colored magenta and cyan, respectively) with rmsd = 3.514 Å, and bottom right: dihedral angles of native and model. Radius of gyration: model = 69.9 Å, and native = 75.4 Å. Simulation parameters: 200,000 biased move steps with 3-mer fragments, and hydrogen bonding term of the score function was kept enabled.*

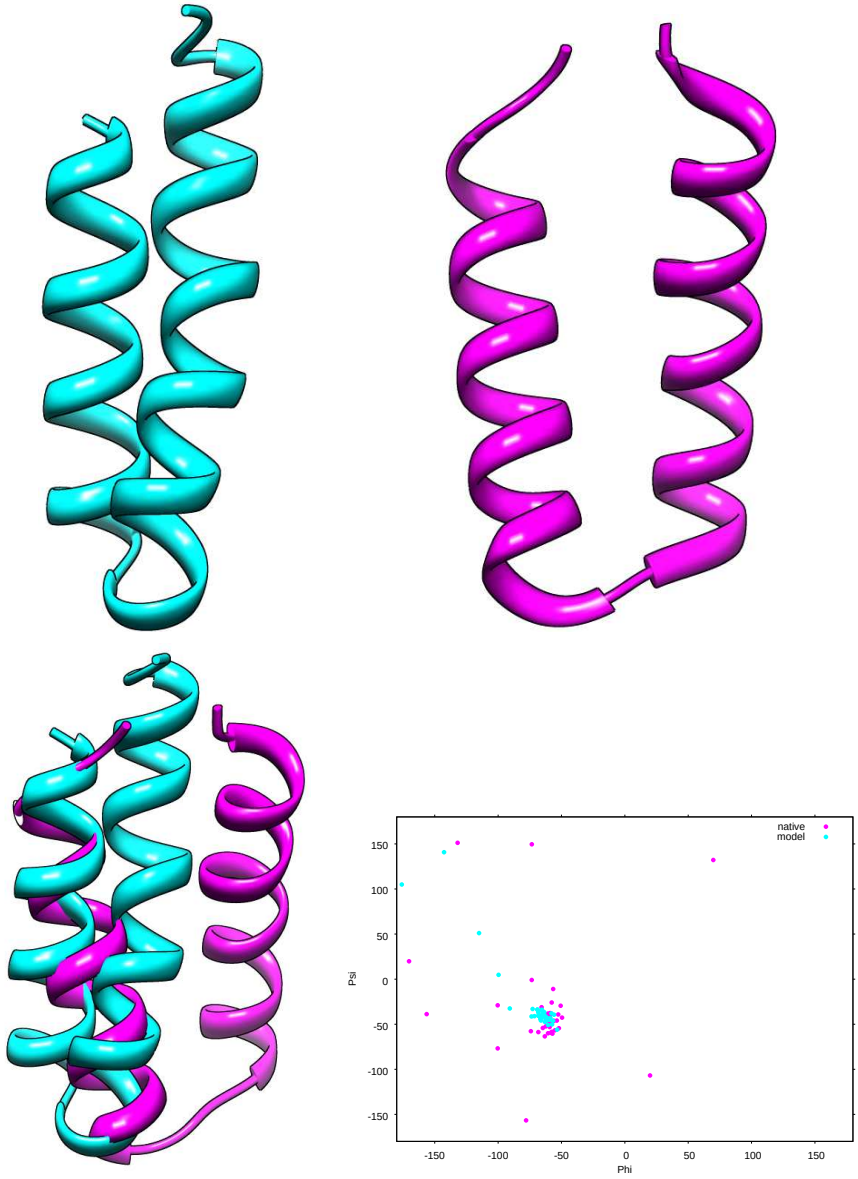


Figure 3.10: *Target 2HEP - top left: model, top right: native, bottom left: superimposition of native and model structures (colored magenta and cyan, respectively) with rmsd = 4.04 Å, and bottom right: dihedral angles of native and model. Radius of gyration: model = 129.1 Å, native = 142.2 Å. Simulation parameters: 500000 biased move steps using 3-residues fragments, and hydrogen bonding term of the score function enabled.*

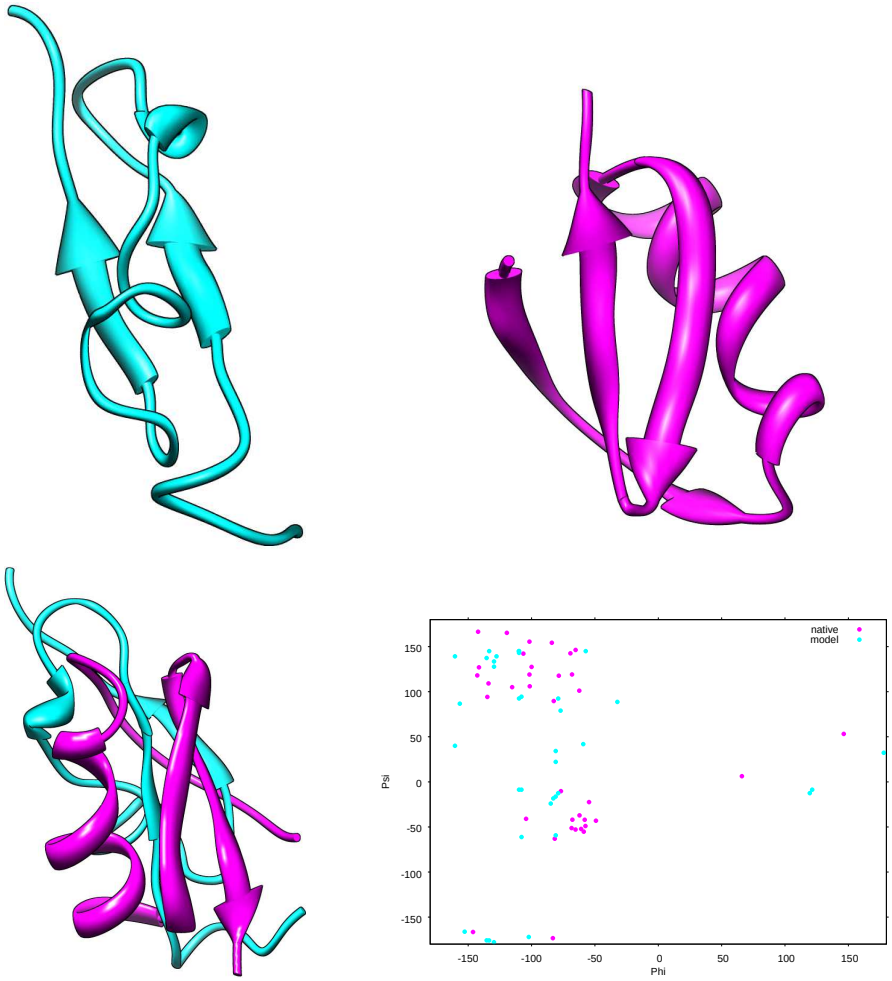


Figure 3.11: *Target 1AGT - top left: model, top right: native, bottom left: superimposition of native and model structures (colored magenta and cyan, respectively), and bottom right: dihedral angles of native and model. Radius of gyration: model = 109.1 Å, and native = 102.2 Å. Simulation parameters: 1500000 biased move steps using 1-residue fragments, and hydrogen bonding term of the score function enabled.*

With few predictions on different targets, it was learned that the search with the solvation enabled score function is more effective when 1-mer fragments are used to make the moves than large fragments. This effect is a direct consequence of an increase in the degrees of freedom due to the incorporation of an additional descriptor of solvation. On insertion of a 3-mer or larger fragments, there is a big change in the overall probability of the conformation due to change in the probabilities of the neighboring fragments. This change in the probabilities of the neighboring fragments occurs because of the dramatic change in their sovlation features (depending upon the shape of inserted fragment). That is why, 1-mer fragments were preferred to produce the following results.

In the following, we describe some models of relatively hard targets. Basically, the search method used biased and 'controlled' moves (with 1-mer fragments) to generate these models and the score function did not take any role of hydrogen bond term into account.

Predictions: CASP Targets

2HF1

2HF1 is a CASP7 free modeling target. Model predicted for this target (shown at the top left of figure 3.12) demonstrates the real strengths and weaknesses of the prediction method. In fact, it has right predictions of a little helical part and of beta sheets. Not only that but the secondary structures roughly have positioned themselves at the places where they should be (as show by superimposition of native and model structures at the bottom right in figure 3.12). The radii of gyration for model and native structures are 221.0 Å and 229.3 Å, respectively. Model reflects a lack of precise long-range interactions between the strands of beta sheets.
The prediction of 2HF1 model involved 500,000 attempted move steps (with 1-mer fragments) by the search method.

The following results are different from the ones described earlier in terms of move set of the search method to find the probable 3D conformation of a model. The search method made use of so called 'controlled' moves to generated these models. Each of the simulation runs to generate these models made 500,000 attempted moves.

2K4X

2K4X is a template-based modeling target from CASP8 target list. Its native structure (given at the top right in figure 3.13) has 0% helix and 12% beta sheet. Model for 2K4X

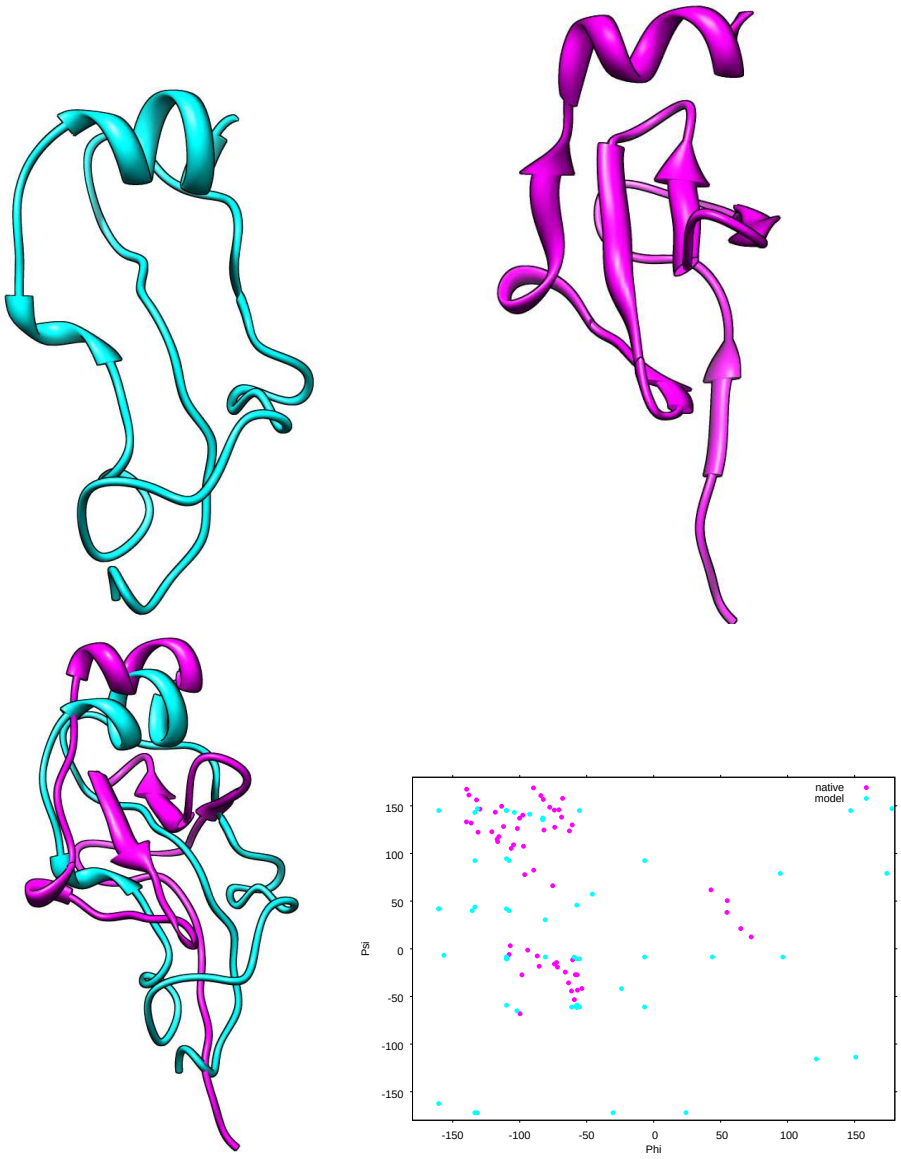


Figure 3.12: *Target 2HF1 - top left: model, top right: native, bottom left: superimposition of native and model structures (colored magenta and cyan, respectively), and bottom right: dihedral angles of native and model. Radius of gyration: model = 221.0 Å, and native = 229.3 Å. Simulation parameters: 500000 biased move steps using 1-residue fragments, and hydrogen bonding term of the score function was kept disabled.*

(given at the top left in figure 3.13) shows right prediction of a substantial part (colored green in both native and model at the top in figure 3.13) of it. Particularly, the two beta sheets roughly positioned themselves at the right place. However, an elongated strand (colored brown) at one end of 2K4X turned to be a helix in model structure. That is why, model looks over-compact in terms of the radius of gyration of model (i.e. 183.2 Å). Whereas the radius of gyration of native structure is 284.1 Å.

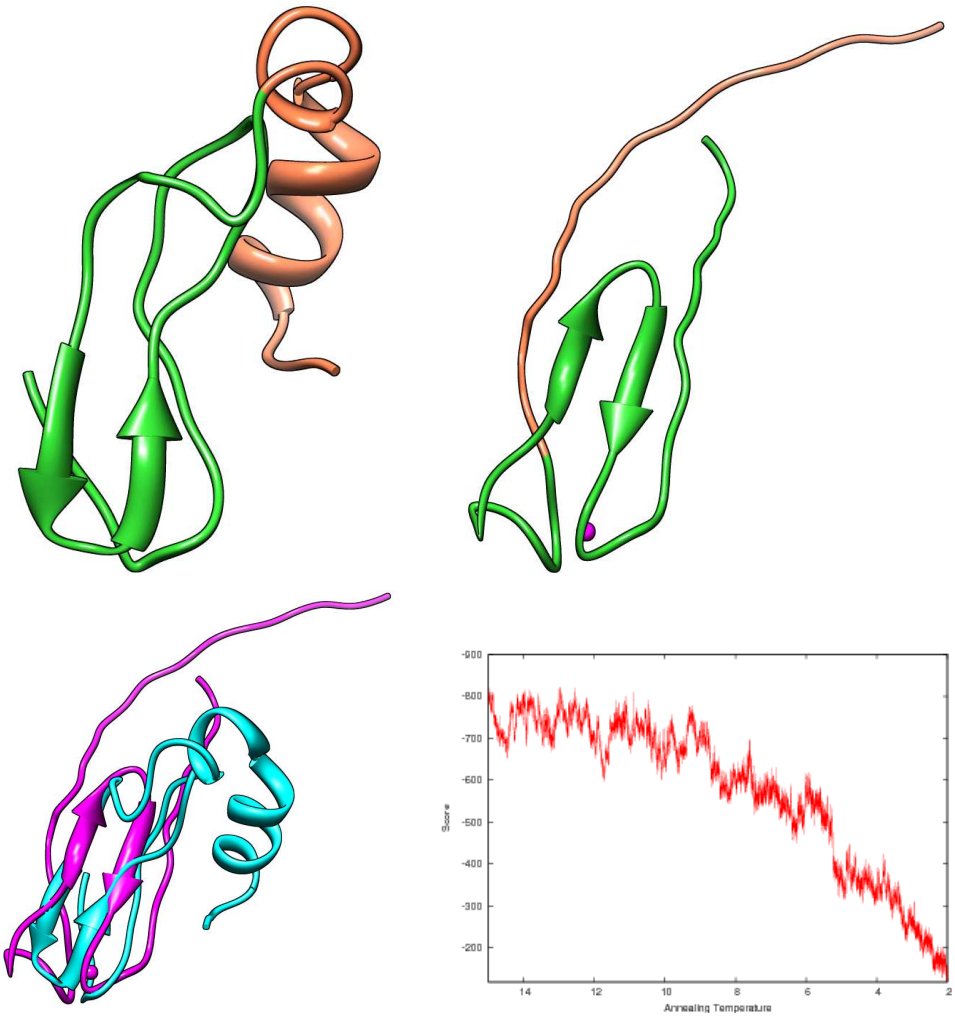


Figure 3.13: *Target 2K4X - top left: model, top right: native, bottom left: superimposition of native and model structures (colored magenta and cyan), and bottom right: probabilistic score versus annealing temperature. Radius of gyration: model = 183.2 Å, and native = 284.1 Å. Simulation parameters: 500000 'controlled' move steps, and hydrogen bonding term of the score function disabled.*

2K53

is a template-based modeling target from the CASP8 target list. Its native structure (given at the top right in figure 3.14) has one long and 3 rather short helices.
2K53 model (given at the top left in figure 3.14) has got good prediction and placement of the longer helix (colored green). Among other three helices, one at the top of other helices (two making a 'V' like shape) is roughly close to the one in native structure. Whereas the prediction of helices which make a 'V' like shape went wrong.
The radii of gyration of model and native structures are 243.9 Å, and 253.6 Å, respectively.

2K4N

2K4N is CASP8 free modeling target consisting of 111 residues. Native structure of 2K4N has 28% helix and 27% beta sheet (according to DSSP secondary structure assignment method). Figure 3.15 (at top) shows two models for 2K4N. Both models have correct prediction of a helix lying at their right. Among beta sheets, model 1 (at the top left in figure 3.15) has good prediction of beta sheet lying below two helices whereas model 2 (at the top right in figure 3.15) went entirely wrong in prediction of this beta sheet. However, model 2 has one of the helices (hanging down on the left) predicted correctly but it is rather misplaced. The two long beta strands (colored brown) in both models went totally wrong.
The dihedral angle plot (at the top in figure 3.17) shows that the angles of both models are somewhat converged to helix and beta sheet regions. The radii of gyration of model 1, model 2 and native were calculated to be 503.4 Å, 511.4 Å, and 634.0 Å, respectively.

2K5C

2K5C is a CASP8 free modeling target of size 108 residues. Its native structure (at the bottom right in figure 3.16) has 31% helix and 12% beta sheet.
Figure 3.16 shows three models for 2K5C with some interesting predictions of various parts of it. Two of the models at the left column in figure 3.16 have close prediction of a helix and a preceding strand (colored brown) at one end. There is a similar helix (colored cyan) succeeded by another strand (colored dark red) at the other end of structures. In this substructure, models at the left column of figure 3.16 have correct prediction of strand whereas model at the top right went entirely wrong. Models at the top right and the bottom left in figure 3.16 could predict correctly but this helix is forward shifted in third model at top left in figure 3.16. Among rest of the structure, a helix (colored green) has right answer in models on the left column in figure 3.16. Some little perturbations

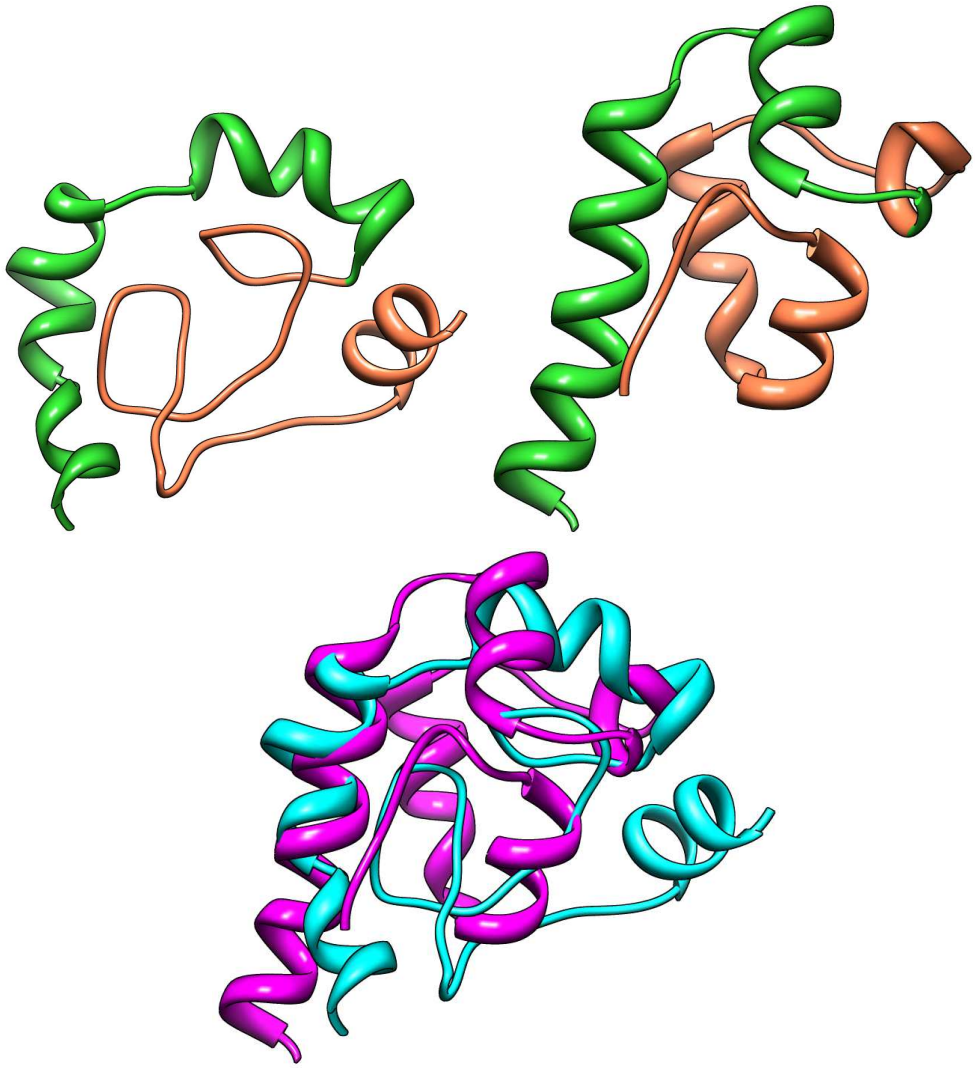


Figure 3.14: *Target 2K53 - top left: model, top right: native, and bottom: superimposition of native and model structures (colored magenta and cyan). Radius of gyration: model = 243.9 Å, and native = 253.6 Å. Simulation parameters: 500,000 'controlled' move steps, and hydrogen bonding term of the score function was kept disabled.*

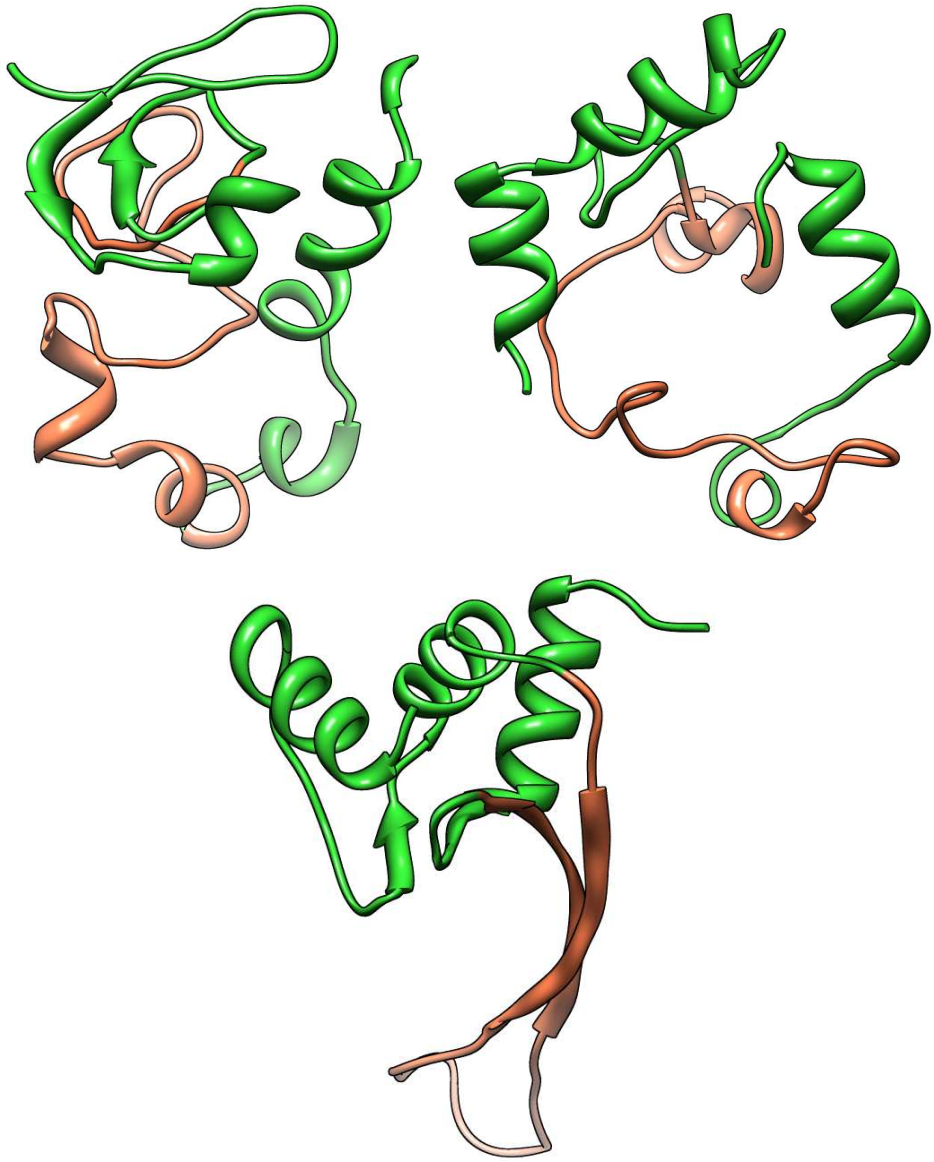


Figure 3.15: *Target 2K4N - top left: model 1, top right: model 2, and bottom: native. Radius of gyration: model1 = 503.4 Å, model2 = 511.4 Å, and native = 634.0 Å. Simulation parameters: 500,000 'controlled' move steps, and hydrogen bonding term of the score function was kept disabled.*

would have led to proper beta sheets (colored cyan and green) in models at the top right and the bottom left in figure 3.16.

Dihedral angle plot of three models and native are given at the bottom in figure 3.17. The radii of gyration of model 1 (top left in figure 3.16), model 2 (top right in figure 3.16), model 3 (bottom left in figure 3.16), and native (bottom right in figure 3.16) were calculated to be 448.0 Å, 612.1 Å, 515.3 Å, and 375.8 Å, respectively.

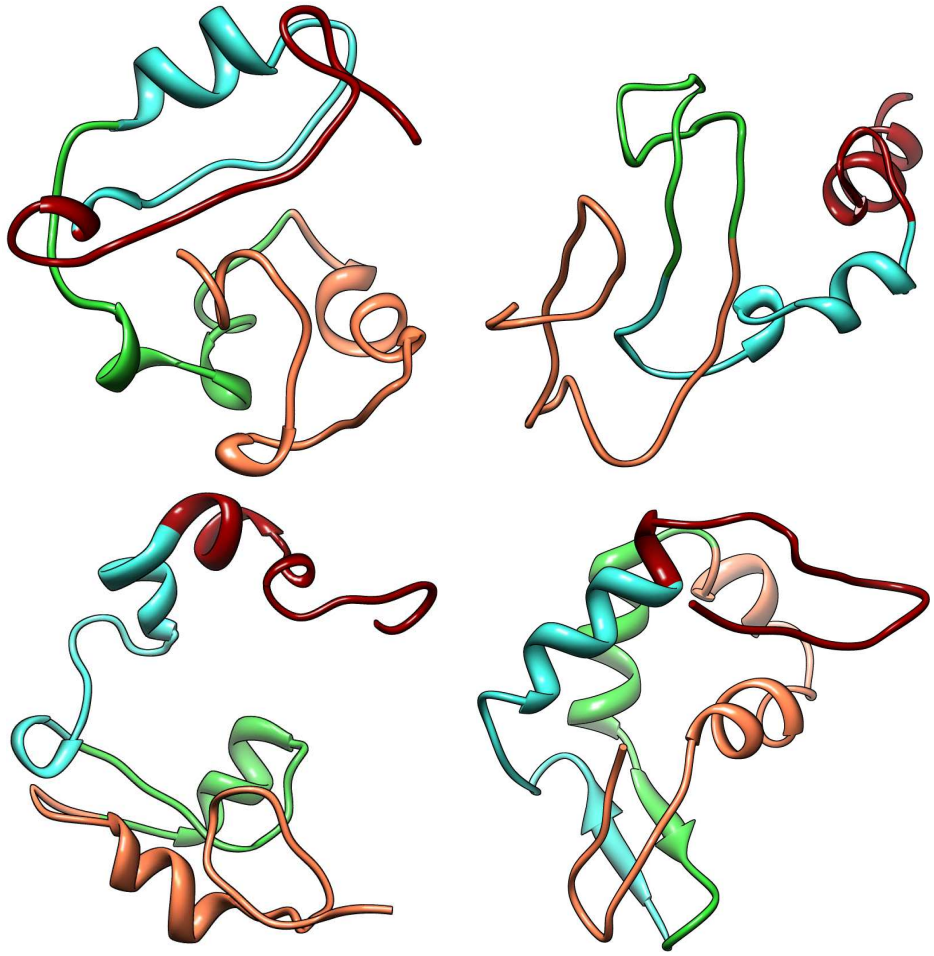


Figure 3.16: *Target 2K5C - top left: model 1, top right: model 2, and bottom left: model 3, and bottom right: native. Radius of gyration: model1 = 448.0 Å, model2 = 612.1 Å, model3 = 515.3 Å, and native = 375.8 Å. Simulation parameters: 500000 'controlled' move steps, and hydrogen bonding term disabled.*

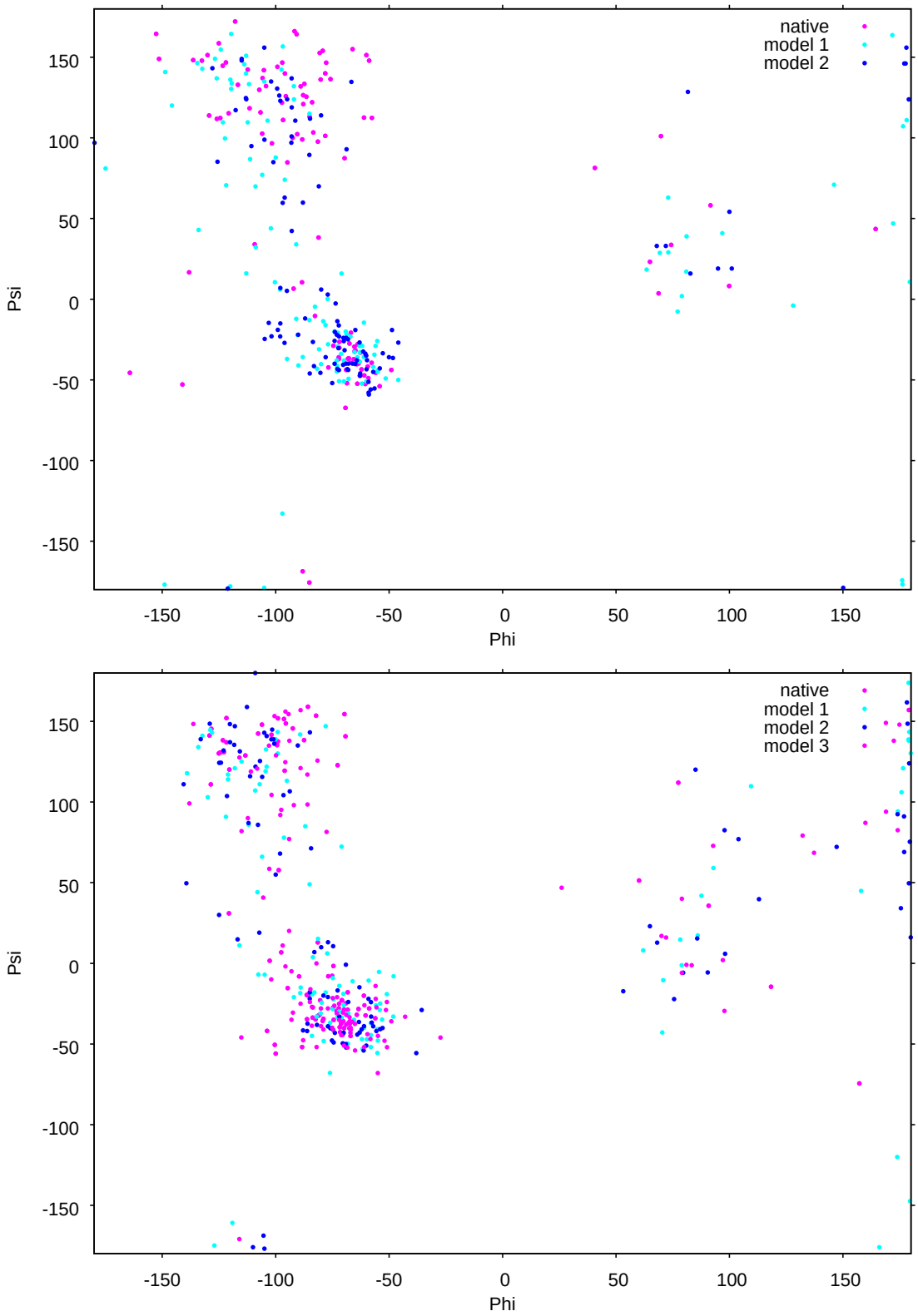


Figure 3.17: *Top: dihedral angles of target 2K4N and two of its models (shown in 3.15), and bottom: dihedral angles of target 2K5C and three of its models (shown in 3.16).*

3DFD

3DFD is rather large (157-residues) free modeling target from CASP8 target list. Helix and beta sheet make 34% and 8% of its native structure. No known structure exists with a sequence slightly similar to that of 3DFD. This makes it a good test for any *ab initio* structure prediction method.
Model for 3DFD (given at the top left in figure 3.18) shows that a considerable part (colored cyan) making helices at the one end was predicted correctly. Beta sheets (colored dark red) are messed up as it has been so due to inability of the score function to recognize long-range interactions precisely.
Dihedral angle plot (given at the bottom in figure 3.18) shows a shift in model angles to the upper right region out of proportion. These shifted angles might be of broken beta sheets. The radii of gyration of 3DFD model and the native were calculated to be 800.5 Å and 651 Å, respectively. The different in the radii of gyration clearly reflects non-compactness of model.

For all the above presented results, the score function used probabilistic classifications built either on 5- or 6-mer fragments. Solvation descriptor for these classifications was calculated by using 10 Å solvation sphere.The classifications with 3, 4 and 7 fragments and 8 or 12 Å solvation spheres could not produce better results.

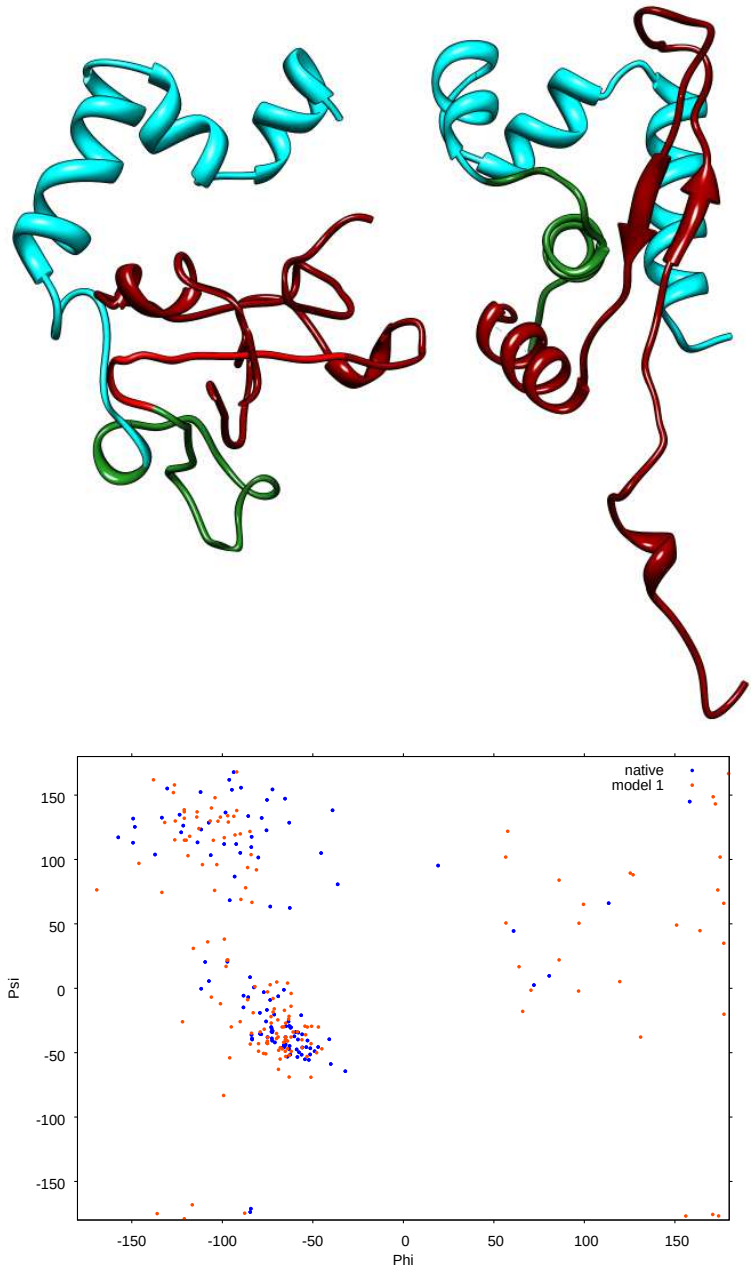


Figure 3.18: *Target 3DFD - top left: model, top right: native, and bottom: dihedral angles of native and model. Radius of gyration: model = 800.5 Å, and native = 651 Å. Simulation parameters: 500,000 'controlled' move steps, and hydrogen bonding term disabled.*

Table 3.3: *From known protein structures in PDB, means and standard deviations of number of neighbor C_β atoms calculated within 8 Å, 10 Å and 12 Å radii around C_β atoms of each of 20 residue types (see figure 3.1).*

Residue	**8 Å SS**[1]		**10 Å SS**		**12 Å SS**	
	Mean	**Std. Dev.** [2]	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Ala	6	3	13	7	24	9
Arg	5	3	10	5	20	8
Asn	4	3	10	6	19	10
Asp	4	3	9	5	18	9
Cys	7	3	16	5	27	9
Gln	4	3	10	5	19	9
Glu	4	3	8	5	17	8
Gly	5	4	11	7	21	11
His	5	3	12	6	22	10
Ile	7	3	15	5	28	10
Leu	7	3	14	5	27	9
Lys	4	2	9	5	17	9
Met	6	3	14	6	26	10
Phe	7	3	15	5	27	9
Pro	5	3	10	6	19	10
Ser	5	3	11	6	20	11
Thr	5	3	12	6	22	10
Trp	6	3	14	5	26	9
Tyr	6	3	14	5	25	9
Val	7	3	15	6	28	10

[1]Solvation sphere
[2]Standard deviation

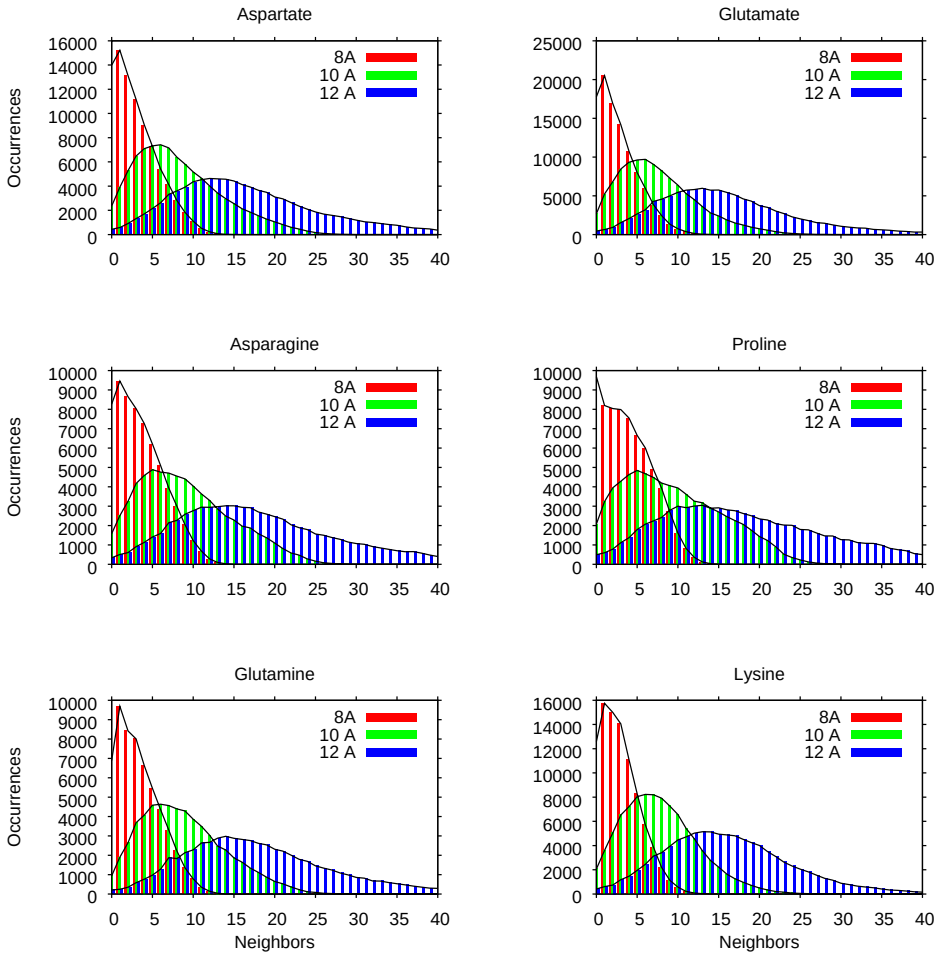


Figure 3.19: *Histograms representing the count of neighbor C_β atoms of the residues: Asp, Glu, Asn, Pro, Gln, and Lys in PDB. Red, green and blue colored histograms represent the count of neighboring C_β atoms within radii of 8, 10 and 12 Å, respectively.*

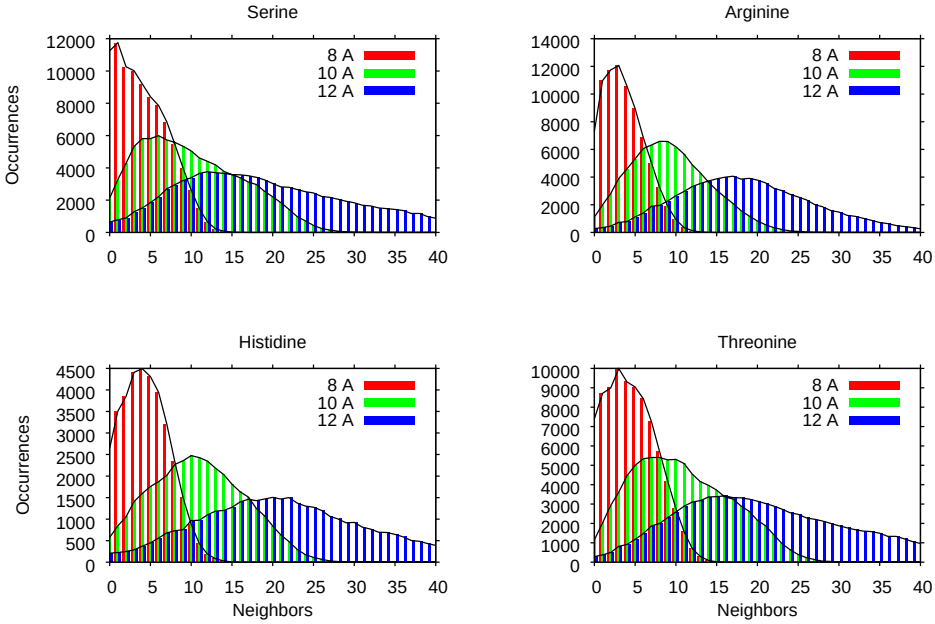


Figure 3.20: *Histograms representing the count of neighbor C_β atoms of the residues: Arg, Thr, Ser, and His in PDB. Red, green and blue colored histograms represent the count of neighboring C_β atoms within radii of 8, 10 and 12 Å, respectively.*

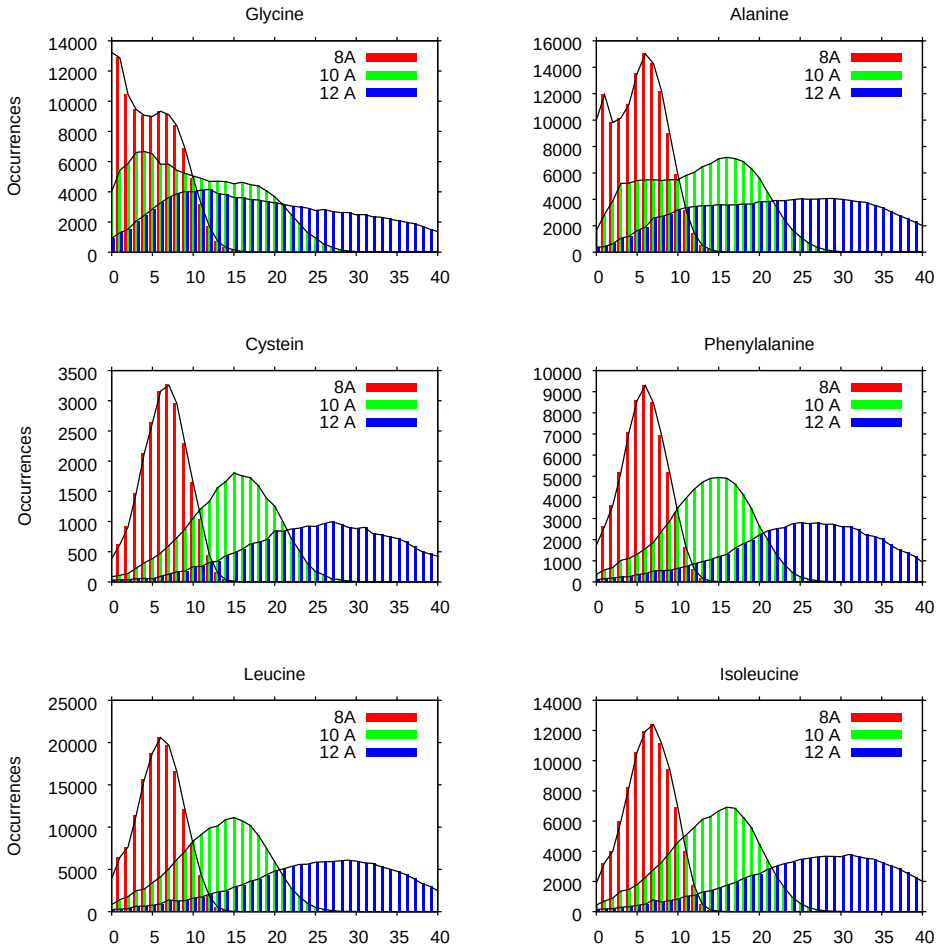


Figure 3.21: *Histograms representing the count of neighbor C_β atoms of the residues: Gly, Ala, Cys, Phe, Leu, and Ile in PDB. Red, green and blue colored histograms represent the count of neighboring C_β atoms within radii of 8, 10 and 12, Å respectively.*

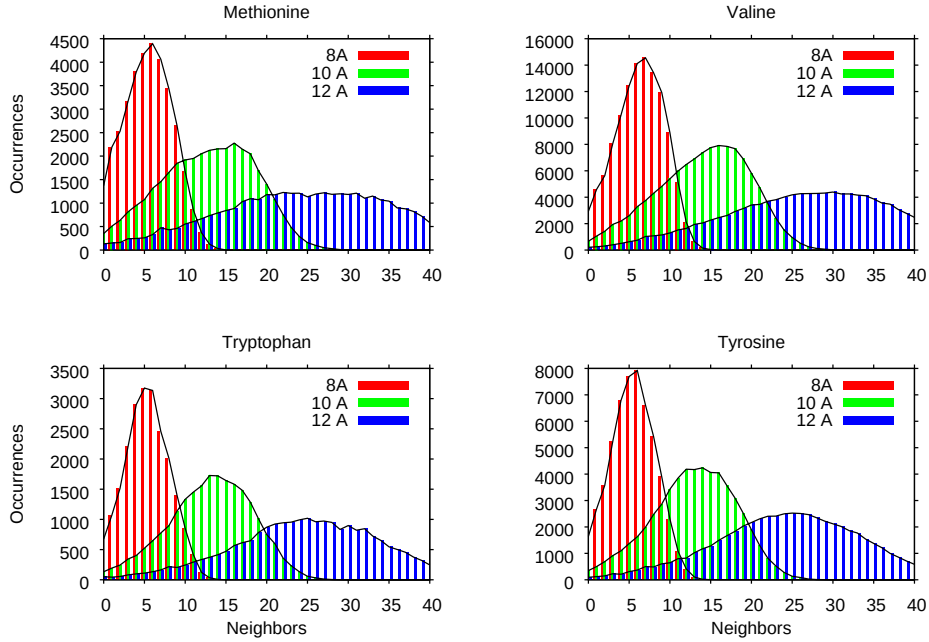


Figure 3.22: *Histograms representing the count of neighbor C_β atoms of the residues: Met, Val, Trp, and Try in PDB. Red, green and blue colored histograms represent the count of neighboring C_β atoms within radii of 8, 10 and 12 Å, respectively.*

Discussions

In general, the idea of introducing solvation effect into the score function has been helpful in the improvement of predictions. It did not only improved the compaction of models (see table 3.2 and 2.2) but also had a positive effect on the accuracy of the predictions as a whole. The extended score function with an improved search method could produce native-like structures for few targets, for example, models for targets like 1FSV, 2HEP and 2HF1. On the other hand, the compactness of models for smaller targets was improved a lot, for example, models for targets 1FSV, 2HEP, and 2K53. However, models for the large targets have been relatively less compact probably because of a very much need but still missing term in the score function for precise long-range interactions.

Model for 2HF1 target (see figure 3.12) perhaps states the true story of how much progress our method has made so far and what possible improvements could be made in the future. 2HF1 model (and most of the other models presented in results section) were generated by using classifications built with 6-mer fragments. In addition to correct predictions of secondary structures, 2HF1 model has a compact 3D structure with an overall arrangement of secondary structures similar to that of its native structure. What is missing is a nice formation of beta sheets. This could have been achieved, if the score function precisely knew about the long-range interactions between loosely lying strands of beta sheet. One may ask why the overall arrangement of secondary structures is not so good in models of rather large size. The reason is our sovlation term can consider medium-range interactions (in smaller structure) to some extent but the real long-range interactions between the distant parts of the large structures are beyond its scope. That 's why, models of the large targets are not well organized and lack in compaction too.

As a short term solution to long-range interactions, an ad hoc hydrogen bonding term based on an electrostatic model (equation 3.2) was introduced. Obviously, this term was not consistent with other (probabilistic) terms of the score function. The main objective of hydrogen bond term was to improve the formation of beta sheets and compactness further by taking hydrogen bonding networks into account. Unfortunately, this effort has not been very successful in achieving the set goal. There are two reasons of this failure: 1) hydrogen bonding term was not consistent with the rest of three terms in the score function (i.e. sequence, structure (ϕ, ψ), and solvation), and 2) the difficulty to determinate an appropriate weight factor w (see equation 3.2) for hydrogen bonding term.

These findings clearly demonstrate the role which a score function consistent hydrogen bonding term can play to make nice predictions. In future, one would need to introduce such a term which should be statistical in nature and acceptable to the existing terms of the score function. This could be achieved through a descriptor based on hydrogen bonding patterns of the fragments to build a Bayesian classification. Such a descriptor can easily be modeled by one of the existing models in the Bayesian framework.

To address the weakness and meet the challenges posed by the upgraded score function (after introduction of solvation term), the search method was extended by: 1) adjustment of the bias in the library-driven moves to make it nearly ergodic, 2) introduction of a third type of moves, called 'controlled' moves, to the move set, and implementation of new methods for conformation probability calculation.

As biased moves had a bias towards helices (as demonstrated and described in sections 2.4 and 2.3.3 and figure 2.18 of chapter 2), this bias was fixed through a scheme (given in section 3.2.3) to make it nearly ergodic. The new nearly ergodic (formerly called biased) moves made with 3-mer fragments seemingly worked better particularly for prediction of small targets, for example, 1FSV and 2HEP. However, in case of large targets, 3-mer or larger fragments were not found to be suitable for these moves even after making them nearly ergodic. This reason is insertion of a such fragment into a large protein conformation does not only replace the dihedral angles and assigns new neighbors (solvation) to that fragment but it also changes the neighbors (solvation) of the neighboring fragments. This multi-dimensional effect causes huge turbulence in the probabilities of the comprising fragments of a protein conformation. On the other hand, the conformation changes caused by the moves made through short fragments are smoother and easier to deal with by the score function. In addition to solvation, the other reason in this extra-ordinary change in the probabilities on insertion of fragments is the way probabilities of conformations are calculated. Every proposed conformation is considered as a set of overlapping fragments. This makes the probability of a fragment too dependent on its preceding and succeeding fragments. This was addressed by implementing three other methods for calculation of probabilities. In fact, these methods lessen the dependence of the constituent fragments of a conformation.

Due to similar reasons mentioned in case of nearly ergodic (formerly biased) moves, the unbiased moves made with large fragments do not produce better prediction models. Even with 1-mer fragments it is hard for the score function to find a probable conformation of any target. The property of complete randomness in unbiased moves leads to strange inconsistent shifts in dihedral angles and consequently to upset in neighbors (solvation) of comprising fragments of a conformation.

On realization of the limitations in biased and unbiased moves, a new type of moves, called 'controlled' move, was introduced to the move set of the search method. The main idea behind these moves was to keep probabilistic changes smooth during conformational search. It was achieved by putting some control over the rotations caused by the change(s) in the dihedral angles. Models presented in section 3.3.3 show that these moves have had positive effect on the predictions.

In future, one can think of two improvements in the search method:

1. Biased moves should not always rely on drawing of N-mer fragments from a library randomly and inserting it at random locations in conformations. Such a scheme actually wastes its time by trying most of the time irrelevant fragments against some sequence fragment/segment of the conformation. It would be nice to save search time by selecting homologous fragments and build a target specific fragment library. Once we have such a library, the search method could make moves with more probable fragments.

2. Having built a target-specific library of N-mer homologous fragments where $N > 1$, the next challenge could be how to utilize these fragments during conformational search by keeping the changes in probabilities smooth and easy for the score function. One possibility could be to draw a homologous fragment randomly and update the dihedral angles of the location of insertion through 'controlled' move steps. That means, if the difference in the dihedral angles of the fragment and those of the conformation is $(40^\circ, -40^\circ)$, this difference should be filled in by making 10 'controlled' move steps. If a 'controlled' move step is rejected in between, either that rejected move should be retained only or all the preceding (accepted) moves (related to the drawn fragment) should also be retained.

Chapter 4

Summary - Zusammenfassung

Protein structure prediction has been the most important scientific problem in the field of computational biology to fill the ever widening gap between protein sequence and structure databases. Experimental methods for protein structure determination are slow and expensive. Only 1% structures of the available protein sequences have been solved experimentally so far. That is why, computational methods to predict three-dimensional structures of protein sequences are absolutely inevitable. As comparative modeling methods rely on the known protein structures, the structure prediction of protein sequences with low sequence similarity to the known protein structures, also called free modeling targets, is difficult for these methods. *Ab initio* structure prediction methods are specifically designed to build the structures of free modeling targets. Like any *ab initio* structure prediction method, there are two aspects of our method: 1) score function, and 2) search method. Although Monte Carlo is frequently used as a search method by the prediction methods, we have uniquely used it with a purely probabilistic score function making no use of Boltzmann statistics.

In first part of this work (described in chapter 2), probabilistic score function based on sequence to structure compatibility functions (Schenk et al. 2008) (from our previous work for protein threading (Torda et al. 2004)) was initially used for *ab initio* structure prediction. The score function consists of a sequence and a structure term modeled by multi-way Bernoulli and bivariate Gaussian distributions respectively. A Bayesian classification of protein fragments (generated from the known protein structures in the Protein Data Bank (PDB)) was built to have the most probable set of classes in the observed data. The parameterized probabilistic descriptions of the found set of classes allows us to calculate the probabilities of the proposed conformations of any target sequence. Unlike Metropolis Monte Carlo, the acceptance criterion is directly based on the ratio of conformational probabilities. The working of the score function also involves an interplay between Cartesian and internal coordinates of protein conformations. Since we do not have a Boltzmann distribution of conformational states, the smoothness of the distribution is controlled through an artificial scheme in our simulated annealing Monte Carlo. There is an arbitrary temperature which ensures that the system moves

towards more probable states at lower temperature and has more even distribution of the states at higher temperature. While the temperature is lowered to cool the system down, the search method makes either biased moves by drawing a fragment from a fragment library or unbiased moves by picking each of the two dihedral values over the interval $(-\pi, \pi)$ to get to the probable state of the system.

The benchmark results of first part demonstrate: 1) the unusual coupling of Monte Carlo with an entirely and purely probabilistic score function works and it can generate protein-like conformations, 2) secondary structures of the target sequences are often predicted at the right positions, 3) the generated models are not properly compact due to the absence of solvation term and the long-range interactions in the score function, and 4) biased moves always lead to straight helical structures for most of the targets. In short, the performance of the method developed this far was up to the expectations and good enough to persuade us to improve it further by: 1) addition of solvation and the long-range interactions to the score function, and 2) extension of move set in order to better explore the conformational space.

In second part of the work (described in chapter 3), the score function was extended by incorporating a solvation term. For this term, a solvation sphere was used to measure the effect of solvation. To calculate solvation effect of a residue, solvation sphere of a certain radius, for example 10 Å, was fixed on that residue and the neighboring C_β atoms within the sphere were calculated. A re-classification of protein fragments was performed to get a new set of classes and their probability distribution parameters. During re-classification, the third term of solvation was modeled by simple normal distribution. A hydrogen bonding effect was temporarily introduced through an electrostatic model. This hydrogen bonding effect in the score function is limited by a weight factor w. To coup with the increased degrees of freedom after introduction of solvation, the move set was also improved through bias correction of the biased moves and addition of 'controlled' moves. Furthermore, new methods for the calculation of conformational probability: average, center average, and simple were implemented to reduce very high interdependency of the constituent fragments of a conformation.

The benchmark results with CASP7, CASP8 and non-CASP targets show a considerable improvement over the solvation-less score function in first part of the work. Models generated for easy non-CASP targets are too close to their native structures, e.g. RMSD of 1FSV native and its model is 3.5 Å. Models of hard and slightly large CASP7 and CASP8 targets (generated without inconsistent hydrogen bonding) are rather compact and sometimes impressive in secondary structure predictions. In future, one would need to incorporate a (probabilistic) hydrogen bonding term consistent with the score function. Such term could help in packing and refinement of models by taking their

long-range interactions into account.

Proteinstrukturvorhersage ist seit einiger Zeit das wichtigste Problem im Bereich der Bioinformatik da sich die Schere zwischen verfuegbaren Sequenz-, und Strukturinformationen immer weiter oeffnet. Experimentelle Methoden zur Strukturbestimmung von Proteinen sind zeitaufwendig und teuer. Fuer nur 1Proteinsequenzen sind die Strukturen bekannt. Um das zu aendern sind rechnergestuetzte Methoden zur Vorhersage der dreidimensionalen Struktur von Proteinen mit bekannter Sequenz unvermeidbar. Weil "Comparative Modelling" Ansaetze von bekannten Strukturen mit aehnlicher Sequenz abhaengig sind ist dieses Vorgehen bei Sequenzen ohne bekannte Strukturen von verwandten Proteinen (sogennannte Free Modeling Targets) nicht praktikabel. Ab-initio Vorhersagemethoden wurden speziell fuer diese Free Modeling Targets entwickelt. Wie alle ab-initio Methoden besteht unser Ansatz aus zwei Teilen: 1.) einer Bewertungsfunktion, und 2.) einer Suchmethode. Obwohl Monte Carlo haeufig als Suchfunktion in der Strukturvorhersage verwendet wird hat unser Ansatz die Besonderheit eine rein probabilistische Bewertungsfunktion zu verwenden die nicht auf der Boltzmann-Statistik aufbaut.

Im ersten Teil dieser Arbeit (Kapitel 2) wurde zuerst eine probabilistische Bewertungsfunktion basierend auf Sequenz-Strukturkompatibilitaet (Schenk et al. 2008)(aus unserer Protein Threading Methode(Torda et al. 2004)) fuer ab-initio Vorhersagen verwendet. Die Bewertungsfunktion besteht aus Sequenz- und Strukturtermen die als Bernoulli- bzw. Gaussverteilungen modelliert wurden. Eine Bayes'sche Klassifizierung von Proteinfragmenten von bekannten Strukturen aus der Protein-Datenbank (PDB) wurde erstellt um die wahrscheinlichsten Klassen in dem Datensatz zu finden. Die probabilistische Beschreibung dieser Klasse erlaubt es uns Wahrscheinlichkeiten von vorgeschlagenen Konformationen fuer eine gegebene Sequenz zu berechnen. Im Gegensatz zu Metropolis Monte Carlo verwendet unser Akzeptanzkriterium das Verhaeltnis von Konformationswahrscheinlichkeiten direkt. Die Bewertungsfunktion haengt ausserdem von dem Wechselspiel zwischen kartesischen und internen Koordinaten der Proteinkonformationen ab. Weil wir keine Boltzmannverteilung der Konformationen zur Verfuegung haben, wird die Glaette der Verteilung durch eine kuenstliche Funktion im "Simulated Annealing Monte Carlo" gesteuert. Eine Temperaturvariable zwingt das System bei niedrigen Werten in wahrscheinlichere Zustaende, und deckt bei hoeheren Werten eine grosse Anzahl von Zustaenden ab. Waehrend das System abgekuehlt wird macht die Suchmethode entweder voreingenommene Schritte indem ein Fragment aus einer Bibliothek ausgewaehlt wird, oder unvoreingenommene Schritte indem die beiden Torsionswinkel (phi und psi) zufaellig aus dem Intervall (-pi, pi) gewaehlt werden. Die Ergebnisse des ersten Teils zeigen dass 1.) die ungewoehnliche Verbindung von Monte Carlo mit einer ausschliesslich probabilistischen Bewertungsfunktion funktion-

iert und proteinaehnliche Konformationen generiert, 2.) die Sekundaerstruktur haeufig korrekt vorhergesagt wird, 3. die erzeugten Modelle nicht kompakt genug sind da ein Loesungsmittelterm und ein Term fuer langereichweitige Interaktionen in der Bewertungsfunktion fehlen, und dass voreingenommene Suchschritte bei den meisten Sequenzen immer zu geraden helikalen Strukturen fuehren. In der Summe hat die Methode die Erwartungen erfuellt und weitere Verbesserungen aufgezeigt: 1.) hinzufuegen eines Loesungsmittelterms und Beruecksichtigung von langen Interaktionen in der Bewertungsfunktion und 2.) Erweiterung des Schrittrepertoires der Suchfunktion um den Konformationsraum besser abzudecken.

Im zweiten Teil der Arbeit (Kapitel 3) wurde die Bewertungsfunktion erweitert. Fuer den Loesungsmittelterm wurde eine Kugel definiert um Kontakte mit Loesungsmittelmolekuelen zu messen. Fuer jede Aminosaeure wurde eine Kugel mit festem Radius (z.B. 10A) definiert und die C-alpha Atome innerhalb dieses Radius berechnet. Die Proteinfragmente wurden dann erneut Klassifiziert wobei die Loesungsmittelzugaenglichkeit mit einer Normalverteilung modelliert wurde. Ausserdem wurde vorruebergehend ein Wasserstoffbrueckenterm durch ein elektrostatisches Modell eingefuehrt. Der Einfluss der Wasserstoffbruecken wurde durch einen Gewichtungsfaktor gesteuert. Um mit der erhoehten Auzahl an Freiheitsgraden der Bewertungsfunktion zurechtzukommen wurde das Schrittrepertoire der Suchfunktion verbessert indem die voreingenommenen Schritte korrigiert wurden, und "kontrollierte" Schritte eingefuehrt wurden. Ausserdem wurden alternative Methoden zur Berechnung von Konformationswahrscheinlichkeiten implementiert um die Abhaengigkeiten zwischen den Fragmenten zu reduzieren: "average", "center average" und "simple". Die Ergebnisse von CASP7, CASP8 und non-CASP Targets zeigen eine signifikante Verbesserung gegenueber der Bewertungsfunktion im ersten Teil. Die Modelle die fuer einfache non-CASP Sequenzen erzeugt werden sind den experimentellen Strukturen sehr aehnlich (RMSD von Modell zu 1FSV: 3,5A). Modelle von schwierigen und groesseren CASP-sequenzen sind recht kompakt und manchmal Zeigen beeindruckend exakte Vorhersagen der Sekundaerstruktur. In Zukunft sollte ein (probabilistischer) Wasserstoffbrueckenterm entwickelt werden der mit der Bewertungsfunktion konsistent ist. Ein solcher Term koennte die Kompaktheit verbessern und die Modelle verfeinern weil lange Interaktionen beruecksichtigt werden.

Safety and Risks

All the research work presented in this manuscript has been purely computational. No experiments with direct use of chemical or biology material were involved at all. Therefore, the standard safety and risk measures (often taken care of by chemists or pharmacists) are irrelevant.

Bibliography

Abagyan, R., Frishman, D. and Argos, P.: 1994, Recognition of distantly related proteins through energy calculations, *Proteins* **19**, 132–140.

Abkevich, V., Gutin, A. and Shakhnovich, E.: 1994, Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry* **33**, 10026–10036.

Aitchison, J. and Brown, J.: 1957, *The lognormal distribution with special reference to its uses in economics*, Cambridge University, London.

Allen, M. and Tildesley, D.: 1989, *Computer simulation of liquids*, Oxford University Press, New York.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: 1990, Basic local alignment search tool, *J Mol Biol* **215**, 403–410.

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

Andricioaei, I. and Straub, J. E.: 1996, Generalized simulated annealing algorithms using tsallis statistics: application to conformational optimization of a tetrapeptide, *Phys Rev E* **53**, R3055–R3058.

Anfinsen, C.: 1973, Principles that govern the folding of protein chains, *Science* **181**, 223–230.

Anfinsen, C. B.: 1972, The formation and stabilization of protein structure, *Biochem J* **128**, 737–749.

Anfinsen, C., Haber, E., Sela, M. and White, F.: 1961, The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, *Proc Natl Acad Sci U S A* **47**, 1309–1314.

Anfinsen, C., Redfield, R., Choate, W., Page, J. and Carroll, W.: 1954, Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease, *J Biol Chem* **207**, 201–210.

Avbelj, F. and Moult, J.: 1995, Determination of the conformation of folding initiation sites in proteins by computer simulation, *Proteins* **23**, 129–141.

Baker, D.: 2004, A surprising simplicity to protein folding, *Nature* **405**, 39–42.

Baker, D. and Sali, A.: 2001, Protein structure prediction and structural genomics, *Science* **294**, 93–96.

Baldwin, R. and Rose, G.: 1999, Is protein folding hierarchic? I. Local structure and peptide folding, *Trends Biochem Sci* **24**, 26–33.

Bayes Rev., T.: 1763, An essay toward solving a problem in the doctrine of chances, *Philos Trans R Soc London* **53**, 370–418.

Beiner, M.: 2007, Proteins: is the folding process dynamically encoded?, *Soft Matter* **3**, 391–393.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E.: 2000, The Protein Data Bank, *Nucleic Acids Res* **28**, 235–242.

Bernal, J. and Crowfoot, D.: 1934, X-ray photographs of crystalline pepsin, *Nature* **133**, 794–795.

Binder, K. and Baumgartner, A.: 1997, Applications of Monte Carlo methods to statistical physics, *Rep Prog Phys* **60**, 487–560.

Blanco, F., Rivas, G. and Serrano, L.: 1994, A short linear peptide that folds into a native stable bold beta-hairpin in aqueous solution, *Nat Struct Biol* **1**, 584–590.

Bloch, F., Hansen, W. and Packard, M.: 1946, Nuclear induction, *Phys Rev* **70**, 460–474.

Blundell, T., Sibanda, B., Sternberg, M. and Thornton, J.: 1987, Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326**, 347–352.

Bogan, A. and Thorn, K.: 1998, Anatomy of hot spots in protein interfaces, *J Mol Biol* **280**, 1–9.

Bonneau, R. and Baker, D.: 2001, Ab initio protein structure prediction: progress and prospects, *Ann Rev Biophys Biomol Struct* **30**, 173–189.

Bork, P. and Gibson, T.: 1996, Applying motif and profile searches, *Methods Enzymol* **266**, 162–184.

Bowie, J. and Eisenberg, D.: 1994, An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function, *Proc Natl Acad Sci U S A* **91**, 4436–4440.

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

Brasseur, R.: 1990, *Molecular description of biological membranes by computer aided conformational analysis*, CRC Press, Boca Raton, FL.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M.: 1983, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations Supported in part by grants from the National Science Foundation and the National Institutes of Health., *J Comp Chem* **4**, 187–217.

Bruyninckx, H.: 2002, Bayesian probability, Website. `http://www.mech.kuleuven.ac.be/~bruyninc/pubs/urks.pdf`.

Bryant, S. and Altschul, S.: 1995, Statistics of sequence-structure threading, *Curr Opin Struct Biol* **5**, 236–244.

Bryant, S. and Lawrence, C.: 1993, An empirical energy function for threading protein sequence through the folding motif, *Proteins* **16**, 92–112.

Burley, S.: 2000, An overview of structural genomics, *Nat Struct Biol* **7**, 932–934.

Burley, S., Almo, S., Bonanno, J., Capel, M., Chance, M., Gaasterland, T., Lin, D., Sali, A., Studier, F. and Swaminathan, S.: 1999, Structural genomics: beyond the human genome project, *Nature Genetics* **23**, 151–158.

Callihan, D. and Logan, T.: 1999, Conformations of peptide fragments from the FK506 binding protein: comparison with the native and urea-unfolded states, *J Mol Biol* **285**, 2161–2175.

Cappe, O., Godsill, S. and Moulines, E.: 2007, An overview of existing methods and recent advances in sequential Monte Carlo, *Proc IEEE* **95**, 899–924.

Center, P. S. P.: 2007, Seventh community wide experiment on the critical assessment of techniques for protein structure prediction (CASP7), Web site. `http://www.predictioncenter.org/casp7/Casp7.html`.

Chandonia, J. and Brenner, S.: 2006, The impact of structural genomics: expectations and outcomes, *Science* **311**, 347–351.

Cheatham III, T. and Kollman, P.: 2000, Molecular Dynamics Simulation of Nucleic Acids, *Ann Rev Phys Chem* **51**, 435–471.

Cheeseman, P., Self, M., Kelly, J., Taylor, W., Freeman, D. and Stutz, J.: 1988, Bayesian classification, *Proc 7th Natl Conf. Artif. Intell.*

Cheung, M., Garcia, A. and Onuchic, J.: 2002, Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse, *Proc Natl Acad Sci U S A* **99**, 685–690.

Chivian, D., Robertson, T., Bonneau, R. and Baker, D.: 2003, Ab initio methods, *Methods Biochem Anal* **44**, 547–558.

Choi, V.: 2005, On Updating torsion angles of molecular conformations, *J Chem Inf Model* **46**, 438–444.

Chothia, C.: 1992, One thousand families for the molecular biologist, *Nature* **357**, 543–544.

Chou, K. and Carlacci, L.: 1991, Simulated annealing approach to the study of protein structures, *Protein Eng Des Sel* **4**, 661–667.

Colonna-Cesari, F. and Sander, C.: 1990, Excluded volume approximation to protein-solvent interaction. The solvent contact model, *Biophys J* **57**, 1103–1107.

Contact, N. and Hunter, P.: 2006, Structure prediction methods, *EMBO Rep* **7**, 249–252.

Covalt, J., Roy, M. and Jennings, P.: 2001, Core and surface mutations affect folding kinetics, stability and cooperativity in IL-1β: does alteration in buried water play a role?, *J Mole Biol* **307**, 657–669.

Cox, R.: 1946, Probability, frequency and reasonable expectation, *Am J Phys* **14**, 1–13.

Creamer, T., Srinivasan, R. and Rose, G.: 1997, Modeling unfolded states of proteins and peptides. II. Backbone solvent accessibility, *Biochemistry* **36**, 2832–2835.

Das, R. and Baker, D.: 2008, Macromolecular modeling with rosetta, *Ann Rev Biochem* **77**, 363–382.

Deane, C. and Blundell, T.: 2003, Protein comparative modelling and drug discovery, *Prac Med Chem* **27**, 445–458.

DeBolt, S. and Skolnick, J.: 1996, Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: Atomic burial position and pairwise non-bonded interactions, *Protein Eng* **9**, 637–655.

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the EM algorithm, *J R Stat Soc Series B* **39**, 1–38.

DePristo, M., De Bakker, P., Lovell, S. and Blundell, T.: 2003, Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles, *Protein Struct Funct Genet* **51**, 41–55.

Derreumaux, P.: 2000, Ab initio polypeptide structure prediction, *Theoretica chimica acta* **104**, 1–6.

Dill, K.: 1990, Dominant forces in protein folding, *Biochemistry* **29**, 7133–7155.

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

Dodson, E.: 2007, Computational biologyProtein predictions, *Nature* **450**, 176–177.

Doll, J. and Freeman, D.: 1994, Monte Carlo methods in chemistry, *IEEE Comp Sci Eng* **1**, 22–32.

Eddy, S.: 1998, Profile hidden Markov models, *Bioinformatics* **14**, 755–763.

Edelsbrunner, H. and Koehl, P.: 2005, The geometry of biomolecular solvation, *Comb Comp Geom* **52**, 243–275.

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

Eisenberg, D., Weiss, R. and Terwilliger, T.: 1984, The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc Natl Acad Sci U S A* **81**, 140–144.

Eramian, D., Eswar, N., Shen, M. and Sali, A.: 2008, How well can the accuracy of comparative protein structure models be predicted?, *Protein Sci* **17**, 1881–1893.

Everitt, B. and Hand, D.: 1981, *Finite mixture distributions*, Chapman and Hall, New York.

Fasman, G.: 1989, *Prediction of protein structure and the principles of protein conformation*, Springer, New York.

Finkelstein, A.: 1997, Protein structure: what is it possible to predict now?, *Curr Opin Struct Biol* **7**, 60–71.

Finkelstein, A., Galzitskaya, O. and Badretdinov, A.: 1996, A folding pathway solving Levinthal's paradox, *Prog Biophys Mol Biol* **65**, 53–53.

Finney, J.: 1996, Overview lecture. Hydration processes in biological and macromolecular systems, *Faraday discussions* **103**, 1–18.

Fiser, A., Do, R. and ŠALI, A.: 2000, Modeling of loops in protein structures, *Protein Sci* **9**, 1753–1773.

Fiser, A., Feig, M., Brooks, C. L. and Sali, A.: 2002, Evolution and physics in comparative protein structure modeling, *Acc Chem Res* **35**, 413–421.

Fiser, A. and Sali, A.: 2003, Modeller: generation and refinement of homology models, *Methods Enzymol* **374**, 461–491.

Frank, H. and Evans, M.: 1945, Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes, *J Chem Phys* **13**, 507–532.

Fujitsuka, Y., Chikenji, G. and Takada, S.: 2006, SimFold energy function for de novo protein structure prediction: consensus with Rosetta, *Proteins* **62**, 381–398.

Gibbs, N., Clarke, A. and Sessions, R.: 2001, Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model, *Proteins* **43**, 186–202.

Gibson, K. and Scheraga, H.: 1967, Minimization of Polypeptide energy, II. Preliminary Structures of Oxytocin, Vasopressin, and an Octapeptide from Ribonuclease, *Proc Natl Acad Sci U S A* **58**, 1317–1323.

Gilis, D. and Rooman, M.: 1996, Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials, *J Mol Biol* **257**, 1112–1126.

Ginalski, K., Grishin, N., Godzik, A. and Rychlewski, L.: 2005, Practical lessons from protein structure prediction, *Nucleic Acids Res* **33**, 1874–1891.

Ginalski, K., Pas, J., Wyrwicz, L., Grotthuss, M., Bujnicki, J. and Rychlewski, L.: 2003, ORFeus: detection of distant homology using sequence profiles and predicted secondary structure, *Nucleic Acids Res* **31**, 3804–3807.

Greer, J.: 1981, Comparative model-building of the mammalian serine proteases, *J Mol Biol* **153**, 1027–1042.

Hansmann, U. H. E.: 2003, Protein folding in silico: an overview, *Comput Sci Eng* **5**, 64–69.

Hanson, R., Stutz, J., Cheeseman, P., Branch, A. I. R. and Center, A. R.: 1991, *Bayesian classification theory*, NASA Ames Research Center, Artificial Intelligence Research Branch, National Technical Information Service, distributor.

Hao, M. and Scheraga, H.: 1994, Monte Carlo simulation of a first-order transition for protein folding, *J Phys Chem* **98**, 4940–4948.

Harano, Y. and Kinoshita, M.: 2004, Large gain in translational entropy of water is a major driving force in protein folding, *Chem Phys Lett* **399**, 342–348.

Hardin, C., Pogorelov, T. and Luthey-Schulten, Z.: 2002, Ab initio protein structure prediction, *Curr Opin Struct Biol* **12**, 176–181.

Hartree, D. and Blatt, F.: 1958, The calculation of atomic structure, *Am J Phys* **26**, 135–136.

Heckerman, D.: 1990, Probabilistic interpretations for mycin's certainty factors, *Readings in uncertain reasoning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Heinemann, U., Illing, G. and Oschkinat, H.: 2001, High-throughput three-dimensional protein structure determination, *Curr Opin Biotechnol* **12**, 348–354.

Henikoff, S. and Henikoff, J.: 1992, Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A* **89**, 10915–10919.

Herbrich, R.: 2002, *Learning kernel classifiers*, Mit Press, Cambridge, MA.

Hohenberg, P. and Kohn, W.: 1964, Inhomogeneous electron gas, *Phys Rev* **136**, 864–871.

Horton, R., Moran, L., Scrimgeour, G., Perry, M. and Rawn, D.: 2006, *Principles of biochemistry*, Prentice Hall, London.

Howson, C. and Urbach, P.: 1991, Bayesian reasoning in science, *Nature* **350**, 371–374.

Huang, E., Samudrala, R. and Park, B.: 2000, Scoring functions for ab initio protein structure prediction, **143**, 223–245.

Hubbard, T.: 1997, New horizons in sequence analysis, *Curr Opin Struct Biol* **7**, 190–193.

Huggins, L. M.: 1971, 50 years of hydrogen bonding theory, *Angew Chem Int Ed* **10**, 147–208.

Hummer, G., Garde, S., Garcıá, A. and Pratt, L.: 2000, New perspectives on hydrophobic effects, *Chem Phys* **258**, 349–370.

Hunter, P.: 2006, Into the fold, *EMBO Reports* **7**, 249–252.

Ishida, T., Nishimura, T., Nozaki, M., Inoue, T., Terada, T., Nakamura, S. and Shimizu, K.: 2003, Development of an ab initio protein structure prediction system ABLE, *Genome Inform* **14**, 228–237.

Jaramillo, A. and Wodak, S.: 2005, Computational protein design is a challenge for implicit solvation models, *Biophys j* **88**, 156–171.

Jaroszewski, L., Rychlewski, L., Zhang, B. and Godzik, A.: 1998, Fold prediction by a hierarchy of sequence, threading, and modeling methods, *Protein Sci* **7**, 1431–1440.

Johnson, M. S., Srinivasan, N., Sowdhamini, R. and Blundell, T. L.: 1994, Knowledge-based protein modelling, *CRC Crit Rev Biochem Mol Biol* **29**, 1–68.

Jones, D.: 1997, Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs, *Proteins* **29**, 185–191.

Jones, D.: 2001, Predicting novel protein folds by using FRAGFOLD, *Proteins* **45**, 127–132.

Jones, D. and McGuffin, L.: 2003, Assembling novel protein folds from super-secondary structural fragments, *Proteins* **53**, 480–485.

Jones, D., Taylort, W. and Thornton, J.: 1992, A new approach to protein fold recognition, *Nature* **358**, 86–89.

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

Kalos, M.: 2007, Monte Carlo methods in the physical sciences, *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, IEEE Press Piscataway, NJ, USA.

Kang, Y., Gibson, K., Nemethy, G. and Scheraga, H.: 1988, Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model, *J Phys Chem* **92**, 4739–4742.

Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. and Hughey, R.: 1999, Predicting protein structure using only sequence information, *Proteins* **37**, 121–125.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. and Hughey, R.: 2003, Combining local-structure, fold-recognition, and new fold methods for protein structure prediction, *Proteins* **53**, 491–496.

Karplus, M. and McCammon, J.: 2002, Molecular dynamics simulations of biomolecules, *Nat Struct Biol* **9**, 646–652.

Karplus, M. and Petsko, G.: 1990, Molecular dynamics simulations in biology, *Nature* **347**, 631–639.

Kauzmann, W.: 1959, Some factors in the interpretation of protein denaturation, *Adv Protein Chem* **14**, 1–63.

Kavraki, L. E.: 2007, Representing proteins in silico and protein forward kinematics, Connexions Web site. `http://cnx.org/content/m11621/1.15/`.

Kendrew, J.: 1959, Three-dimensional structure of globular proteins, *Rev Mod Phys* **31**, 94–99.

Kihara, D., Zhang, Y., Lu, H., Kolinski, A. and Skolnick, J.: 2002, Ab initio protein structure prediction on a genomic scale: Application to the Mycoplasma genitalium genome, *Proc Natl Acad Sci U S A* **99**, 5993–5998.

Kirkpatric, S., Gelatt, C. and Vecchi, M.: 1983, Optimization by simulated annealing, *Science* **220**, 671–680.

Kocher, J., Rooman, M. and Wodak, S.: 1994, Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches, *J Mol Biol* **235**, 1598–1613.

Kolinski, A.: 2004, Protein modeling and structure prediction with a reduced representation, *Acta Biochim Pol* **51**, 349–372.

Kolinski, A. and Gront, D.: 2007, Comparative modeling without implicit sequence alignments, *Bioinformatics* **23**, 2522–2527.

Krasley, E., Cooper, K., Mallory, M., Dunbrack, R. and Strich, R.: 2006, Regulation of the oxidative stress response through Slt2p-dependent destruction of cyclin C in Saccharomyces cerevisiae, *Genetics* **172**, 1477–1486.

Landau, D. and Binder, K.: 2005, *A guide to Monte Carlo simulations in statistical physics*, Cambridge University Press, London.

Lathrop, R.: 1994, The protein threading problem with sequence amino acid interaction preferences is NP-complete, *Protein Eng Des Sel* **7**, 1059–1068.

Lathrop, R. and Smith, T.: 1996, Global optimum protein threading with gapped alignment and empirical pair score functions, *J Mol Biol* **255**, 641–665.

Lattman, E.: 2004, The state of the protein structure initiative, *Proteins* **54**, 611–615.

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

Lee, J., Kim, S., Joo, K., Kim, I. and Lee, J.: 2004, Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing, *Proteins* **56**, 704–714.

Lemer, C., Rooman, M. and Wodak, S.: 1995, Protein structure prediction by threading methods: evaluation of current techniques, *Proteins* **23**, 337–355.

Lesk, A.: 2004, *Introduction to protein science*, Oxford University Press, New York.

Levinthal, C.: 1969, Mossbauer spectroscopy in biological systems, *Proceedings of a meeting held at Allerton House. P. Debrunner, JCM Tsibris, and E. Munck, editors. University of Illinois Press, Urbana, IL*.

Levitt, M., Hirshberg, M., Sharon, R. and Daggett, V.: 1995, Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution, *Comp Phys Commun* **91**, 215–231.

Levitt, M. and Sharon, R.: 1988, Accurate simulation of protein dynamics in solution, *Proc Natl Acad Sci U S A* **85**, 7557–7561.

Levy, Y. and Onuchic, J.: 2004, Water and proteins: A love-hate relationship, *Proc Natl Acad Sci U S A* **101**, 3325–3326.

Levy, Y. and Onuchic, J.: 2006, Water mediation in protein folding and molecular recognition, *Annu Rev Biophys Biomol Struct* **35**, 389–415.

Li, L., Mirny, L. and Shakhnovich, E.: 2000, Kinetics, thermodynamics and evolution of nonnative interactions in a protein folding nucleus, *Nat Struct Biol* **7**, 336–342.

Li, W. Z., Jaroszewski, L. and Godzik, A.: 2001, Clustering of highly homologous sequences to reduce the size of large protein databases, *Bioinformatics* **17**, 282–283.

Li, Z., Laidig, K. E. and Daggett, V.: 1998, Conformational search using a molecular dynamics-minimization procedure: applications to clusters of coulombic charges, Lennard-Jones particles, and waters, *J Comput Chem* **19**, 60–70.

Lifson, S., Hagler, A. T. and Dauber, P.: 1979, Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the C:O.cntdot..cntdot..cntdot.H- hydrogen bonds, *J Am Chem Soc* **101**, 5111–5121.

Liu, J.: 2001, *Monte Carlo strategies in scientific computing*, Springer, New York.

Lushington, G. H.: 2008, Comparative modeling of proteins, *Mol Model Proteins* **443**, 199–212.

Madej, T., Gibrat, J. and Bryant, S.: 1995, Threading a database of protein cores, *Proteins* **23**, 356–369.

Marchler-Bauer, A., Anderson, J., DeWeese-Scott, C., Fedorova, N., Geer, L., He, S., Hurwitz, D., Jackson, J., Jacobs, A., Lanczycki, C. et al.: 2003, CDD: a curated Entrez database of conserved domain alignments, *Nucleic Acids Res* **31**, 383–387.

Mardia, K., Kent, J., Bibby, J. et al.: 1979, *Multivariate analysis*, Academic press, New York.

Margulis, C. J., Stern, H. A. and Berne, B. J.: 2002, Helix unfolding and intramolecular hydrogen bond dynamics in small α-helices in explicit solvent, *J Phys Chem B* **106**, 10748–10752.

Marqusee, S., Robbins, V. and Baldwin, R.: 1989, Unusually stable helix formation in short alanine-based peptides, *Proc Natl Acad Sci U S A* **86**, 5286–5290.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, R. and Sali, A.: 2000, Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct* **29**, 291–325.

McDonald, I. K. and Thornton, J. M.: 1994, Satisfying hydrogen bonding potential in proteins, *J Mol Biol* **238**, 777–793.

Melo, F., Sanchez, R. and Sali, A.: 2002, Statistical potentials for fold assessment, *Protein Sci* **11**, 430–438.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A. and Teller, E.: 1953, Equations of state calculations by fast computing machines, *J Chem Phys* **21**, 1087–1092.

Montelione, G., Zheng, D., Huang, Y., Gunsalus, K. and Szyperski, T.: 2000, Protein NMR spectroscopy in structural genomics, *Nat Struct Biol* **7**, 982–985.

Morozov, A., Kortemme, T., Tsemekhman, K. and Baker, D.: 2004, Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations, *Proc Natl Acad Sci U S A* **101**, 6946–6951.

Moult, J.: 2005, A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr opin struct biol* **15**, 285–289.

Moult, J.: 2008, Comparative modeling in structural genomics, *Structure* **16**, 14–16.

Murray, D. and Honig, B.: 2002, Electrostatic control of the membrane targeting of C2 domains, *Mol Cell* **9**, 145–154.

Nassal, M., Leifer, I., Wingert, I., Dallmeier, K., Prinz, S. and Vorreiter, J.: 2007, A structural model for duck hepatitis B virus core protein derived by extensive mutagenesis, *J Virol* **81**, 13218–13229.

Ngan, S., Inouye, M. and Samudrala, R.: 2006, A knowledge-based scoring function based on residue triplets for protein structure prediction, *Protein Eng Des Sel* **19**, 187–193.

Nölting, B.: 1999, *Protein folding kinetics: biophysical methods*, Springer, Berlin.

Ooi, T., Oobatake, M., Nemethy, G. and Scheraga, H.: 1987, Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, *Proc Natl Acad Sci U S A* **84**, 3086–3090.

Orengo, C., Jones, D. and Thornton, J.: 1994, Protein superfamilies and domain superfolds, *Nature* **372**, 631–634.

Oschkinat, H., Griesinger, C., Kraulis, P., Sørensen, O., Ernst, R., Gronenborn, A. and Clore, G.: 1988, Three-dimensional NMR spectroscopy of a protein in solution, *Nature* **332**, 374–376.

Panchenko, A., Marchler-Bauer, A. and Bryant, S.: 2000, Combination of threading potentials and sequence profiles improves fold recognition, *J Mol Biol* **296**, 1319–1331.

Papoian, G., Ulander, J., Eastwood, M., Luthey-Schulten, Z. and Wolynes, P.: 2004, Water in protein structure prediction, *Proc Natl AcaSci* **101**, 3352–3357.

Park, B. and Levitt, M.: 1995, The complexity and accuracy of discrete state models of protein structure, *J Mol Biol* **249**, 493–507.

PDB: 2009, Protein Data Bank (PDB) - an information portal to biological macromolecular structures, PDB statistics. `http://www.rcsb.org/pdb`.

Pearson, W. R. and Lipman, D. J.: 1988, Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A* **85(8)**, 2444–2448.

Pedersen, J. and Moult, J.: 1995, Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms, *Proteins* **23**, 454–460.

Pedersen, J. and Moult, J.: 1996, Genetic algorithms for protein structure prediction, *Curr Opin Struct Biol* **6**, 227–231.

Petsko, G. and Ringe, D.: 2004, *Protein structure and function*, New Science Press, London.

Pratt, L. and Pohorille, A.: 2002, Hydrophobic effects and modeling of biophysical aqueous solution interfaces, *Chem Rev* **102**, 2671–2692.

Ptitsyn, O.: 1987, Protein folding: hypotheses and experiments, *J Protein Chem* **6**, 273–293.

Ptitsyn, O., Lim, V. and Finkelstein, A.: 1972, Analysis and simulation of biochemical systems, *Proc 8th FEBS Meet.*

Ptitsyn, O. and Rashin, A.: 1973, Stagewise mechanism of protein folding, *Doklady Akademii Nauk SSSR* **213**, 473–475.

Ptitsyn, O. and Rashin, A.: 1975, A model of myoglobin self-organization, *Biophys Chem* **3**, 1–20.

Purcell, E., Torrey, H. and Pound, R.: 1946, Resonance absorption by nuclear magnetic moments in a solid, *Phys Rev* **69**, 37–38.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A., Read, R. and Baker, D.: 2007, High-resolution structure prediction and the crystallographic phase problem, *Nature* **450**, 259–264.

Radford, S.: 2000, Protein folding: progress made and promises ahead, *Trends Biochem Sci* **25**, 611–618.

Ramachandran, G., Ramakrishnan, C. and Sasisekharan, V.: 1963, Stereochemistry of polypeptide chain configurations., *J Mol Biol* **7**, 95–99.

Rey, A. and Skolnick, J.: 1991, Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins, *Chem Phys* **158**, 199–219.

Rohl, C., Strauss, C., Misura, K. and Baker, D.: 2004, Protein structure prediction using Rosetta, *Methods Enzymol* **383**, 66–93.

Rost, B.: 1999, Twilight zone of protein sequence alignments.

Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A.: 2000, Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Sci* **9**, 232–241.

Sagui, C. and Darden, T.: 1999, Molecular dynamics simulations of biomolecules: long-range electrostatic effects, *Ann Rev Biophys Biomol Struct* **28**, 155–179.

Sali, A.: 1995, Modeling mutations and homologous proteins, *Curr Opin Biotechnol* **6**, 437–451.

Sali, A. and Blundell, T. L.: 1993, Comparative protein modelling by satisfaction of spatial restraints, *Curr Opin Biotechnol* **234**, 779–815.

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

Samudrala, R., Xia, Y., Huang, E. and Levitt, M.: 1999, Ab initio protein structure prediction using a combined hierarchical approach, *Proteins* **37**, 194–198.

Sánchez, R., Pieper, U., Melo, F., Eswar, N., Martí-Renom, M., Madhusudhan, M., Mirkovic, N. and Sali, A.: 2000, Protein structure modeling for structural genomics, *Nat Struct Biol* **7**, 986–990.

Sanchez, R. and Sali, A.: 1997, Advances in comparative protein-structure modelling, *Curr Opin Struct Biol* **7**, 206–214.

Schellman, J.: 1955, The stability of hydrogen-bonded peptide structures in aqueous solution., *CR Trav Lab Carlsberg [Chim]* **29**, 230–259.

Schenk, G., Margraf, T. and Torda, A.: 2008, Protein sequence and structure alignments within one framework, *Algorithms Mol Biol* **3**, 4–15.

Scheraga, H. A., Khalili, M. and Liwo, A.: 2007, Protein-folding dynamics: overview of molecular simulation techniques, *Ann Rev Phys Chem* **58**, 57–83.

Sela, M., White, F. and Anfinsen, C.: 1957, Reductive cleavage of disulfide bridges in ribonuclease, *Science* **125**, 691–692.

Service, R.: 2005, Structural Biology: Structural Genomics, Round 2.

Shakhnovich, E.: 1997, Theoretical studies of protein-folding thermodynamics and kinetics, *Curr Opin Struct Biol* **7**, 29–40.

Shakhnovich, E., Abkevich, V. and Ptitsyn, O.: 1996, Conserved residues and the mechanism of protein folding, *Nature* **379**, 96–98.

Sharp, K., Nicholls, A., Fine, R. and Honig, B.: 1991, Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects, *Science* **252**, 106–109.

Shi, J., Blundell, T. and Mizuguchi, K.: 2001, FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J Mol Biol* **310**, 243–257.

Siepmann, J. and Frenkel, D.: 1992, Configurational bias Monte Carlo: a new sampling scheme for flexible chains, *Mol Phys* **75**, 59–70.

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

Simons, K., Strauss, C. and Baker, D.: 2001, Prospects for ab initio protein structural genomics, *J Mol Biol* **306**, 1191–1199.

Sippl, M.: 1990, Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins, *J Mol Biol* **213**, 859–883.

Sippl, M.: 1995, Knowledge-based potentials for proteins, *Curr Opin Struct Biol* **5**, 229–235.

Sippl, M., Hendlich, M. and Lackner, P.: 1992, Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin β4, *Protein Sci* **1**, 625–640.

Skolnick, J., Fetrow, J. and Kolinski, A.: 2000, Structural genomics and its importance for gene function analysis, *Nat Biotechnol* **18**, 283–287.

Skolnick, J. and Kolinski, A.: 2001, Computational studies of protein folding, *Comp Sci Eng* **3**, 40–50.

Smyth, M. and Martin, J.: 2000, x Ray crystallography, *J Clin Pathol: Mol Pathol* **53**, 8–14.

Soding, J.: 2005, Protein homology detection by HMM-HMM comparison, *Bioinformatics* **21**, 951–960.

Sohl, J., Jaswal, S. and Agard, D.: 1998, Unfolded conformations of alpha-lytic protease are more stable than its native state, *Nature* **395**, 817–819.

Sun, S.: 1993, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, *Protein Sci* **2**, 762–785.

Tadashi, A., MEGURO, T. and YAMATO, I.: 2002, Development of an atomistic Brownian dynamics algorithm with implicit solvent model for long time simulation, *J Comp Chem Jpn* **1**, 115–122.

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

Taylor, W. and Orengo, C.: 1989, Protein structure alignment, *J Mol Biol* **208**, 1–22.

Thiele, R., Zimmer, R. and Lengauer, T.: 1999, Protein threading by recursive dynamic programming, *J Mol Biol* **290**, 757–779.

Titterington, D., Smith, A. and Makov, U.: 1985, *Statistical analysis of finite mixture models*, Wiley, Chichester, UK.

Torda, A. E.: 2005, *Protein Threading*, The proteomics protocols handbook (Ed. Walker, J.M.), Humana Press, Totowa N.J., pp. 921–938.

Torda, A., Procter, J. and Huber, T.: 2004, Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices, *Nucleic acids research* **32**, W532–W535.

Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Seo, J., Lee, J., Baker, D., Chivian, D., Kim, D. et al.: 2007, Assessment of predictions submitted for the CASP7 domain prediction category, *Protein* **69**, 137–151.

Trigg, G.: 2005, *Mathematical tools for physicists*, Wiley-VCH Verlag GmbH & Co., Weinheim.

Tsallis, C.: 1988, Possible generalization of Boltzmann-gibbs statistics, *J Stat Phys* **52**, 479–487.

Unger, R.: 2004, The genetic algorithm approach to protein structure prediction, *Struct Bond* **110**, 153–176.

Wang, P., Yan, B., Guo, J., Hicks, C. and Xu, Y.: 2005, Structural genomics analysis of alternative splicing and application to isoform structure modeling, *Proc Natl Acad Sci U S A* **102**, 18920–18925.

Weiner, P. and Kollman, P.: 1981, AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions, *J Comp Chem* **2**, 287–303.

Wesson, L. and Eisenberg, D.: 1992, Atomic solvation parameters applied to molecular dynamics of proteins in solution, *Protein Sci* **1**, 227–235.

Wodak, S. and Rooman, M.: 1993, Generating and testing protein folds, *Curr Opin Struct Biol* **3**, 247–259.

Wüthrich, K.: 1986, *NMR of proteins and nucleic acids*, J. Wiley and Sons, Inc., New York.

Wüthrich, K.: 2003, Noble Lecture: NMR studies of structure and function of biological macromolecules, *Biosci Rep* **23**, 119–168.

Xu, Y., Xu, D. and Liang, J.: 2006, *Computational methods for protein structure prediction and modeling: basic characterization*, Springer, New York.

Yadgari, J., Amir, A. and Unger, R.: 1998, Genetic algorithms for protein threading, *Proc 6th Internatl. Conf. Intell. Syst. Mol Biol (ISMB)*.

Yang, J., Chen, W., Skolnick, J. and Shakhnovich, E.: 2007, All-atom ab initio folding of a diverse set of proteins, *Structure* **15**, 53–63.

Ye, Y., Li, Z. and Godzik, A.: 2006, Modeling and analyzing three-dimensional structures of human disease proteins., *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, Pac Symp Biocomput.

Yue, P. and Moult, J.: 2006, Identification and analysis of deleterious human SNPs, *J Mol Biol* **356**, 1263–1274.

Zaki, M. and Bystroff, C.: 2007, *Protein structure prediction*, Humana Press, New York.

Zhang, M. and Kavraki, L. E.: 2002, A new method for fast and accurate derivation of molecular conformations, *J Chem Inf Comput Sci* **42**, 64–70.

Zhang, Y.: 2008, Progress and challenges in protein structure prediction, *Curr Opin Struct Biol* **18**, 342–348.

Zhang, Y. and Skolnick, J.: 2004, Automated structure prediction of weakly homologous proteins on a genomic scale, *Proc Natl Acad Sci U S A* **101**, 7594–7599.

Zhou, H. and Skolnick, J.: 2007, Ab initio protein structure prediction using chunk-TASSER, *Biophys J* **93**, 1510–1518.

Zhou, H. and Zhou, Y.: 2004, Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins* **55**, 1005–1013.

Zwanzig, R., Szabo, A. and Bagchi, B.: 1992, Levinthal's paradox, *Proc Natl Acad Sci U S A* **89**, 20–22.

Résumé

Personal Information

Name:	Nasir Mahmood
Date of Birth:	January 18th, 1978 in Multan, Pakistan
Marital Status:	Married, 1 child
Address:	Present: Reichweindamm 44, 13627 Berlin, Germany Permanent: H/No. 1147/C Mumtazabad, 60600 Multan, Pakistan

Education

2006	**Postgraduate Diploma in Applied Bioinformatics** Cologne University Bioinformatics Center (CUBIC), University of Cologne, Cologne, Germany	Germany
2005	**M.Sc. Computational Visualistics** Faculty of Computer Science, Otto-von-Guericke University of Magdeburg, Magdeburg, Germany	Germany
2003	**MS Computer Science** Punjab College of Information Technology (PCIT), University of Central Punjab, Lahore, Pakistan	Pakistan
2000	**B.Sc. (Hons) Agriculture** University College of Agriculture (UCA), Bahauddin Zakariya University, Multan, Pakistan	Pakistan
1996	**F.Sc. (Pre-Medical)** Govt. College Multan, Pakistan	Pakistan

EXPERIENCES

PERIOD	08.2006 — Present	Berlin,
EMPLOYER	Technical University Berlin	Germany
JOB TITLE	Research Assistant	
PERIOD	04.2006 — 03.2009	Hamburg,
EMPLOYER	Center for Bioinformatics Hamburg (ZBH)	Germany
	University of Hamburg	
JOB TITLE	Research Assistant	
PERIOD	12.2004 — 03.2005	Halle (Saale),
EMPLOYER	Center for Environmental Research (UFZ)	Germany
JOB TITLE	Research Assistant	
PROJECT	Mapping and Visualization of MacMan Project	
PERIOD	08.2003 — 12.2004	Rawalpindi,
EMPLOYER	CERN LAB	Pakistan
JOB TITLE	Research Assistant	
PROJECT	Enabling Handheld Device for Grid Applications	

PUBLICATIONS

2008 Mahmood, N. and Torda, A.: 2008, Protein Structure Prediction Using Coarse Grain Force Fields, *Proc NIC Workshop 2008* 40, 309-312.

POSTERS

2008 Mahmood, N. and Torda, A.: 2008, Monte Carlo Simulation with Herrn Boltzmann *at 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction 2008, Cagliari, Italy*.

Mahmood, N. and Torda, A.: 2008, Protein Structure Prediction: Probabilistic Force Fields *at Computer Simulation and Theory of Macromolecules 2008, Huenfeld, Germany*.

2007 Mahmood, N., Schenk, G. and Torda, A.: 2007, Monte Carlo Simulations with Unusual Probability Sampling *at Methods of Molecular Simulation 2007, Heidelberg, Germany*.

Mahmood, N. and Torda, A.:2007, Protein Structure Prediction using Coarse Grain Force Fields *at German Conference on Bioinformatics 2007, Potsdam, Germany*.

2006 Mahmood, N., Stehr, T., Hoffmann, S., Margraf, T., Mosisch, M., Schenk, G., Reuter, G., Huber, T. and Torda, A.:2006, CASP7 Predictions using the Wurst Server *at 7th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction 2006, USA*.

Versicherung an Eides statt

Nach § 3 Promotionsordung versichere ich an Eides statt, dass ich meine Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als von mir angegebene Hilfsmittel und Quellen nicht benutzt und die den benutzten wörtlich oder inhaltlich entnommenen Stellen als solche kentlich gemacht habe.

Ort, Datum: _____

Unterschrift: _____