Winter 2015

De Novo Protein Structure Modeling and Energy Function Design

Lin Chen
Old Dominion University

DE NOVO PROTEIN STRUCTURE MODELING AND ENERGY FUNCTION DESIGN

by

Lin Chen
B.S. May 2001, Lanzhou University, P.R. China
M.S. December 2009, New Mexico State University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
December 2015

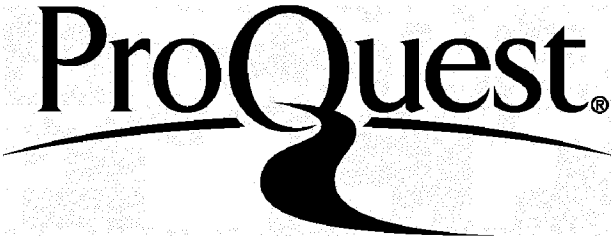
Approved by:

Jing He (Director)

Desh Ranjan (Member)

Lesley H. Greene (Member)

Yaohang Li (Member)



ABSTRACT

DE NOVO PROTEIN STRUCTURE MODELING AND ENERGY FUNCTION DESIGN

Lin Chen
Old Dominion University, 2015
Director: Dr. Jing He

The two major challenges in protein structure prediction problems are (1) the lack of an accurate energy function and (2) the lack of an efficient search algorithm. A protein energy function accurately describing the interaction between residues is able to supervise the optimization of a protein conformation, as well as select native or native-like structures from numerous possible conformations. An efficient search algorithm must be able to reduce a conformational space to a reasonable size without missing the native conformation. My PhD research studies focused on these two directions.

A protein energy function—the distance and orientation dependent energy function of amino acid key blocks (DOKB), containing a distance term, an orientation term, and a highly packed term—was proposed to evaluate the stability of proteins. In this energy function, key blocks of each amino acids were used to represent each residue; a novel reference state was used to normalize block distributions. The dependent relationship between the orientation term and the distance term was revealed, representing the preference of different orientations at different distances between key blocks. Compared with four widely used energy functions using six general benchmark decoy sets, the DOKB appeared to perform very well in recognizing native conformations. Additionally, the highly packed term in the DOKB played its important

role in stabilizing protein structures containing highly packed residues. The cluster potential adjusted the reference state of highly packed areas and significantly improved the recognition of the native conformations in the ig_structal data set. The DOKB is not only an alternative protein energy function for protein structure prediction, but it also provides a different view of the interaction between residues.

The top-k search algorithm was optimized to be used for proteins containing both α-helices and β-sheets. Secondary structure elements (SSEs) are visible in cryo-electron microscopy (cryo-EM) density maps. Combined with the SSEs predicted in a protein sequence, it is feasible to determine the topologies referring to the order and direction of the SSEs in the cryo-EM density map with respect to the SSEs in the protein sequence. Our group member Dr. Al Nasr proposed the top-k search algorithm, searching the top-k possible topologies for a target protein. It was the most effective algorithm so far. However, this algorithm only works well for pure α-helix proteins due to the complexity of the topologies of β-sheets. Based on the known protein structures in the Protein Data Bank (PDB), we noticed that some topologies in β-sheets had a high preference; on the contrary, some topologies never appeared. The preference of different topologies of β-sheets was introduced into the optimized top-k search algorithm to adjust the edge weight between nodes. Compared with the previous results, this optimization significantly improved the performance of the top-k algorithm in the proteins containing both α-helices and β-sheets.

ACKNOWLEDGMENTS

I would like to express my gratitude to several persons. Without them, I would not be able to complete this dissertation.

Firstly, I would like to express my deepest appreciation to my advisor, Dr. Jing He whose patient guidance and endless encouragement helped me in all the time of research and writing of this dissertation.

I also wish to extend many, many thanks to my committee members, Dr. DeshRanjan, Dr. Lesley H. Greene, Dr. Mohammad Zubair and Dr. Yaohang Lifor their valuable time and their precious advice to this dissertation.

I want to thank the Computer Science Department of Old Dominion University for the financial support.

At last, I would like to give my special thanks to my mother, my wife and my daughter for their support and encouragement in the past five years.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

Chapter

1. INTRODUCTION 1

2. PROTEIN ENERGY FUNCTION DESIGN 29
   2.1 METHOD 34
   2.2 RESULTS AND DISCUSSION 52
   2.3 CONCLUSIONS 73

3. PROTEIN TOP-K TOPOLOGY PROBLEM 78
   3.1 METHOD 88
   3.2 RESULTS AND DISCUSSION 106

REFERENCES 109

VITA 129

LIST OF TABLES

Table		Page
1.	Definition of key blocks	37
2.	The percentage of high dense residues in decoy sets	49
3.	The performance of four potentials on five decoy sets	62
4.	The number of proteins with better/same/worse rank for the native conformations	63
5.	The performance of seven potentials for CASP8_30_r decoys	64
6.	The rank of the native conformation of CASP8_30 decoys	66
7.	Improved recognition of the native conformation among the decoys of 1dbb	71
8.	The rank of the native conformation for ig_structal decoys	75
9.	The average RMSD of the top-ranked conformations for CASP8 decoys	76
10.	The rank of the native conformation in DecoysRus set	77
11.	The rank of the native topology in α-β proteins	108

LIST OF FIGURES

Figure Page

1. The 20 amino acids that make up proteins....3
2. Peptide bond formation....5
3. Four distinct protein structures for cyclophilin A....6
4. Dihedral angles of the backbone of the proteins....7
5. Parallel strands and anti-parallel strands....8
6. The number of protein structures solved by multiple methods....11
7. Flow for solving the atomic structure of proteins with X-rays....13
8. Procedure that builds a 3D electron density map using Cryo-EM....16
9. Resolution distribution of a density map in the EMDB....19
10. The number of released protein sequences and structures....20
11. The flow chart for de novo modeling....26
12. The definition of 19 rigid-body blocks in OPUS-PSP....35
13. The distance and orientation representation of a pair of key blocks....37
14. Web-based energy function....43
15. The density distributions for all 30 block pairs of ASP-ARG....44
16. The pair correlation functions for all 30 block pairs of ASP-ARG....45
17. The distance energy functions for all 30 block pairs....47
18. Examples of the distance energy....55
19. The distribution of orientation energy for block pair (16,16) of (PHE,PHE) and (18,14) of (TRP,ASN)....56
20. Probability difference for low-energy residue pairs in highly packed clusters....70

21. The plot of the energy for all decoys of 1dbb and 1nsn....72

22. The native structure of 1acy and a decoy in decoy set ig-structal....74

23. Helix sticks and the topologies....79

24. Application of interpretation tree in finding a match of model features to image features....82

25. The graph of the pure α-helix proteins in the top-k topologies search algorithm....86

26. A 4-stranded β-sheet....87

27. The flow chart of the top-k topology search program....88

28. The input information of 2KUM for the top-K topology search algorithm....89

29. The protein 2KUM and the corresponding skeleton....91

30. The graph of 2KUM, built with SSE-Ss and SSE-Ds....94

31. Comparison of the lengths of SSE-S and SSE-D for a node....95

32. The graph of 2KUM with the unary constraints....96

33. Popular topologies and β-sheet....98

34. The graph of 2KUM with some of the edge weights....99

35. The reverse pseudo tree for the first four shortest paths....103

CHAPTER 1

INTRODUCTION

Proteins are involved in almost all functional processes within living organisms, including metabolic reaction catalyzing, molecule transportation, DNA replication, molecule storage, immune protection, etc. [13] [14]. Those functions depend on the interactions among proteins in which the differences are due to the composition of the proteins in the sequence and their three-dimensional (3D) structures in space [15; 16; 17; 18; 19]. For a given protein sequence, the corresponding structure has been determined by a common principle [20; 21] between the amino acids. Revealing how protein sequences fold is an essential requirement in understanding classical biological phenomena and in providing useful information for drug design and other biotechnological applications.

A protein is a polymer consisting of a sequence of amino acids. As shown in Figure 1, there are 20 types of amino acids. Each amino acid has a chemical formula $H_2NC\alpha HRCOOH$ and a specific side-chain, denoted as R [22]. These side-chains cause each amino acid to have a specific property [23]. Based on the charge, the hydrophobicity, the size or the chemical characteristics of their side-chain R, amino acids are classified as positive, neutral, negative, hydrophilic, hydrophobic, aliphatic, aromatic, or acidic [24]. The classification reveals the role that each amino acid plays in protein folding and it provides a hint for predicting the protein's structure.

Neighboring amino acids in the sequence combine to generate the protein chain. The carboxyl group (COOH) in one amino acid reacts with the amine group (NH_2) in the next amino acid in the sequence [22]. One water molecule is dehydrated and a peptide bond

between the two amino acids occurs in this condensation reaction (Figure 2). The two ends of the polypeptide chain are known as the amino terminus (N-terminus) and the carboxyl terminus (C-terminus). The monomer that an amino acid loss –OH on one side and –H on another side is called the residue. It has the following chemical formula: $-HNC\alpha HRCO-$. This dehydration procedure occurs between all amino acid pairs in the sequence. The backbone of a protein consists of the $C\alpha$, the CO group, and the NH group [25, 26]. The arrangement of these backbone atoms represents the topology of the proteins.

The composition of the amino acid residues in the protein sequence decides the 3D structure of a protein. A specific residue sequence will fold into a particular 3D protein structure. During this procedure, the protein structure has four distinct levels: a primary structure, a secondary structure, a tertiary structure, and a quaternary structure [25], as discussed below.

1. Primary structure refers to the linear residue sequence of the protein chain from the N-terminus to the C-terminus. This is an unbranched chain of residues that can contain from tens to hundreds of amino acids [27; 28]. Figure 3A shows a part of the amino acid chain of the protein cyclophilin A, indexed as 2X2A in the protein data bank (PDB) [29].

   Secondary structure refers to the local energetically favorite segments of the protein structure. Major types of the secondary structure include: the α-helix [30], the β-sheet [31; 32; 33], and the turn/loop [34] (Figure 3 B). The α-helix and the β-sheet have regular geometries with specific dihedral angle values [26], which are due to the hydrogen bonds among the residues in the peptide backbone. The turns/loops connect these regular sub-structures to form the tertiary structure.

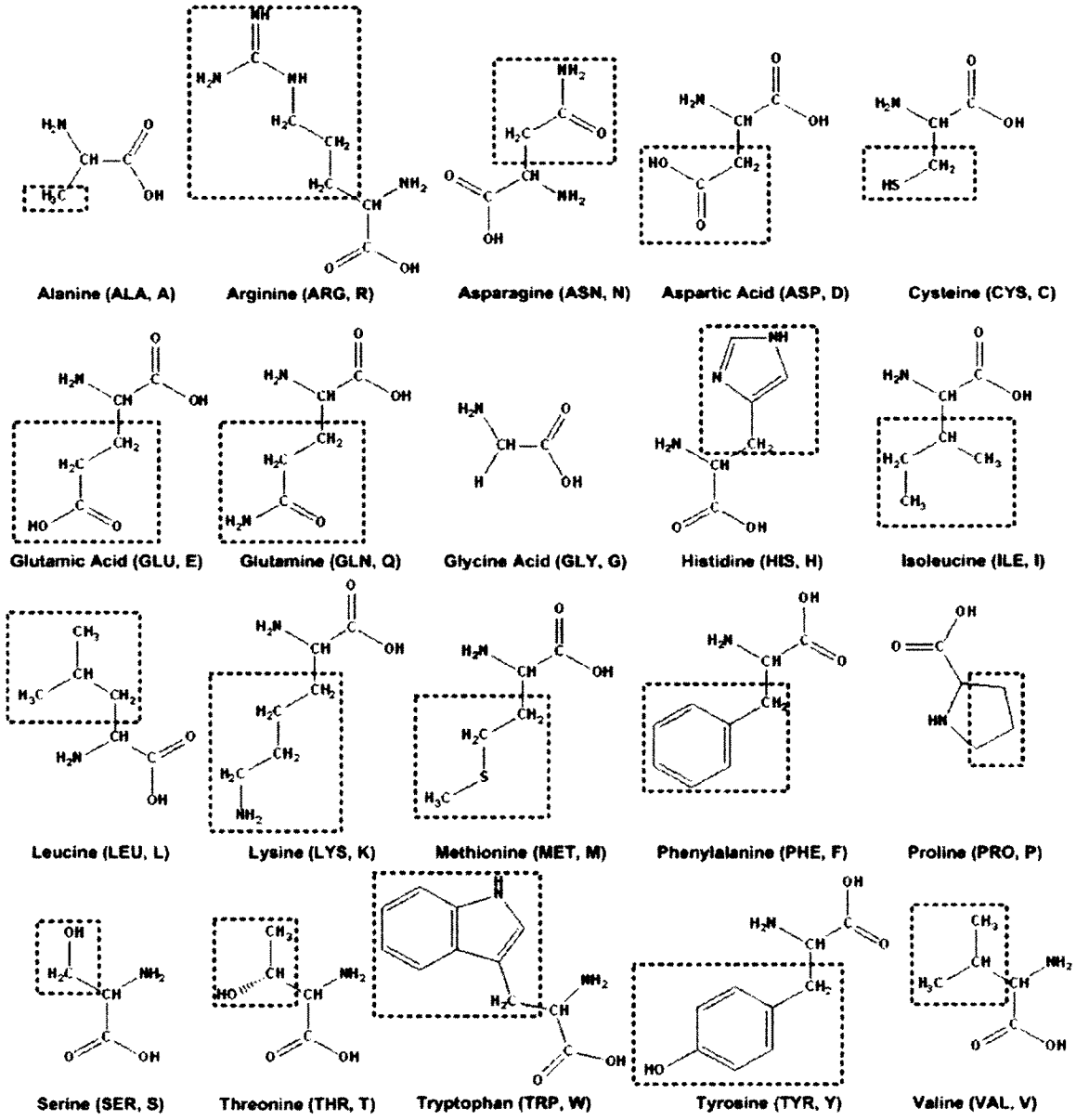


Figure 1. The 20 amino acids that make up proteins. The amine group (NH_2), carbon alpha ($C\alpha$), and the carboxyl group (COOH) are fully shown without removing H and OH. The side-chains (R) are highlighted with rectangular-shaped frame composed of dash lines (GLY has no side chain).

a. α-helix: In this secondary structure, each NH group at residue i connects to the CO group at residue i+4 along the backbone with the hydrogen bond. This is a right-handed coiled or spiral conformation where each helix turn has 3.6 residues and translation along α-helix axis is 1.5 Å. The height of the α-helix turn is 5.4 Å (3.6*1.5) [35]. The backbone dihedral angles (φ, ψ) for the residues are shown in Figure 4; these are used to describe the protein conformation and the value of the angles (φ, ψ) is within -60° and -45°, respectively {Dickerson, 1969

b. #2713} for the α-helix. Side-chains for each residue are attached to the external surface of this helical structure. However, the 20 residues have unequal propensities for forming an α-helix. Alanine, aspartic acid, glutamic acid, isoleucine, leucine, and methionine all have high helix-forming propensities, whereas glycine and proline have poor helix-forming propensities [36, 37, 38].

c. β-sheet: In this type of secondary structure, two or more different segments along the primary structure, called β-strands, form a twisted, pleated sheet [31]. This structure is stabilized with at least two or three backbone hydrogen bonds. Neighboring β-strands can be either parallel or anti-parallel (Figure 5). For the parallel sheet, two strands have the same direction, which is defined as the direction from the N-terminus to the C-terminus for a protein chain, and which have backbone dihedral angles (φ, ψ) of -120° and 115°, respectively. For the anti-parallel sheet, the two strands have opposite directions and the dihedral angles are -140° and 135° [31]. Both the parallel strand pairs and the anti-parallel strand pairs adopt the hydrogen bonds between the hydrogen of the amine (NH) group and the oxygen of the carboxyl (C=O) group. The diversity is that the residue i forms hydrogen bonds to

the residues j-1 and j+1 if two atoms, C_α^i and C_α^j , are adjacent Cα in two hydrogen-bonded β-strands in the parallel β-sheets; whereas, the residue i forms hydrogen bonds to residue j [39].

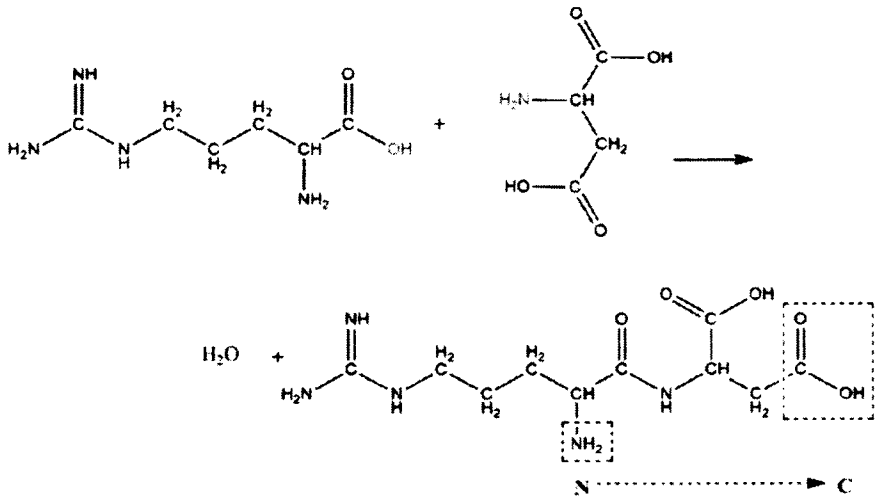


Figure 2. Peptide bond formation. The COOH group in ARG (first amino acid) reacts with the NH_2 group in ASP (second amino acid), and one water molecule is dehydrated. The polypeptide chain is displayed from the N-terminus to the C-terminus (highlighted as the rectangle-shaped dash lines).

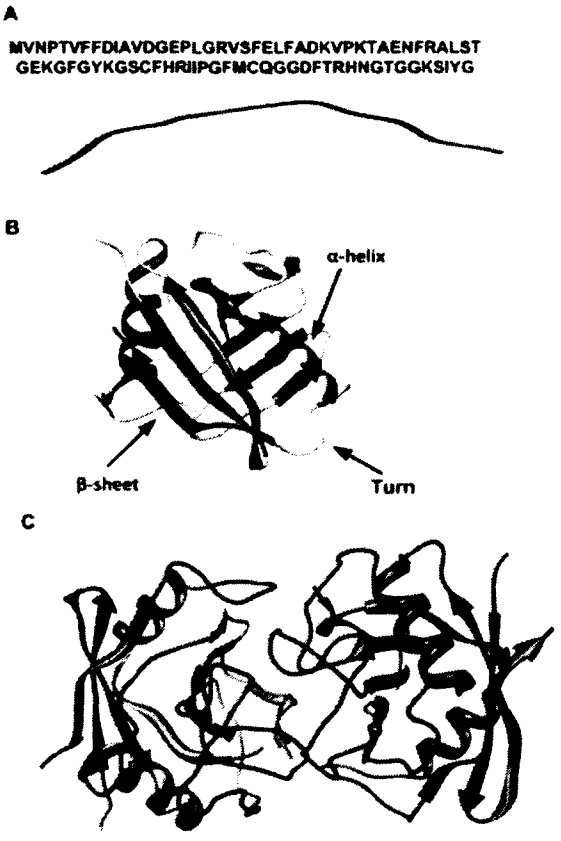


Figure 3. Four distinct protein structures for cyclophilin A (2X2A). (A) Primary structure: a linear residue sequence; (B) Tertiary structure: the turns/loops connect the secondary structure: α-helices (red) and β-sheets (blue); (C) Quaternary structure: a set of organized tertiary structures.

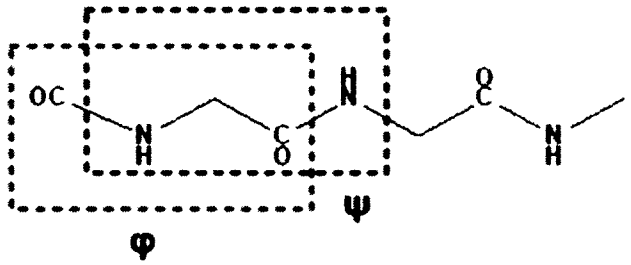


Figure 4. Dihedral angles of the backbone of the proteins. φ (phi) involves C-N-Cα-C atoms (red), ψ (psi) involves N-Cα-C-N atoms (blue).

d. Turn/loop refers to a structural motif in which the Cα atoms of two residues in the structural motif are less than 7Å [34]. In the backbone, the dihedral angles of the turns are not constant. This is in contrast to the α-helices and the β-sheets.

2. Based on the separation between the two end residues of the turns in the sequence, there can be several types of turns: α-turn (i, i$\pm$4), β-turn (i, i$\pm$3), γ-turn (i, i$\pm$2), δ-turn (i, i$\pm$1), and π-turn (i, i$\pm$5) [40].
3. Tertiary Structure refers to the 3D structure of a single protein chain, as shown in Figure 3 B. Thus, tertiary structures are the arrangement of different secondary structures of the same protein chain in 3D space [25]. This structure is stabilized by intra-protein interactions, such as hydrogen bonds, disulfide bonds, electrostatic interactions and van der Waals forces [41; 42; 43]. The tertiary structure is likely determined by the primary structure. Predicting the tertiary structure from the

primary structure is known generally as protein structure prediction [44].

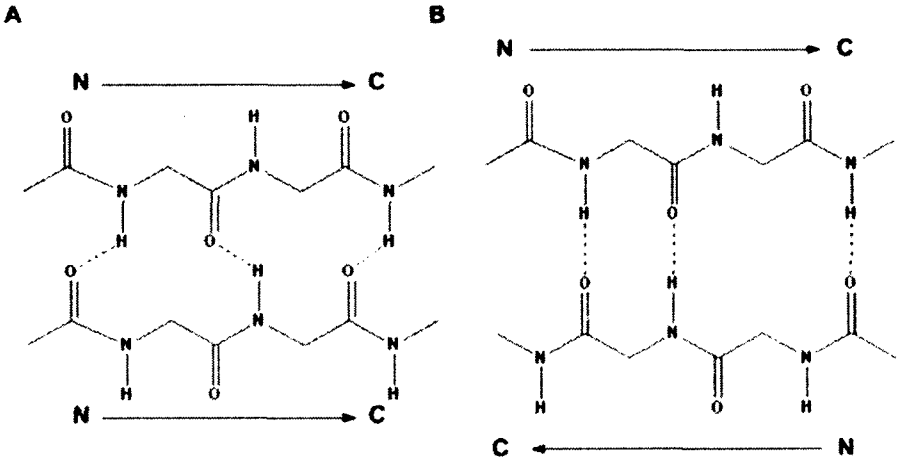


Figure 5. Parallel strands (A) and anti-parallel strands (B).

4. Quaternary Structure refers to the 3D structure of a multi-chain protein. Proteins with two or more polypeptides are called multimers [25]. These subunits are combined with non-covalent interactions and disulfide bonds as the tertiary structure. This level structure is not included in this current study.

Protein structures have been studied for more than fifty years since the first protein structure was reported by John Kendrew in 1958 [45, 46] who was awarded the 1962 Nobel Prize in Chemistry [47]. In the 1960s, due to the rapid development of high-

resolution structure determination techniques, molecular biology became a well-known field. In 1962, Aaron Klug developed crystallographic electron microscopy and applied it to solve nucleic acid-protein complexes [48]. He received the 1982 Nobel Prize in Chemistry for his contribution to protein structure determination. Michael Rossmann proposed a replacement technique to predict unknown protein structures from existing structures [49; 50]. In 1971, the Protein Data Bank (PDB) was established at Brookhaven National Laboratory to deposit 3D structures of proteins and nucleic acids [51]. Initially, PDB only contained seven structures, but now it has over 100, 000 structures. In 1976, Robert Langridge developed the first visualization program to visualize the protein structures, and he established a computer graphics lab at the University of California, San Francisco [52; 53]. In 1978, Kurt Wüthrich introduced Nuclear Magnetic Resonance (NMR) into the study of protein structure [54]; Wüthrich received the 2002 Nobel Prize in Chemistry for his contribution to studying the structure of biological macromolecules [47]. In 1982, Jane Richardson developed ribbon diagrams to represent protein structure, and this has become the standard way of visualizing proteins [55]. In 1983, Jacque Dubochet succeeded in producing biological specimens by freezing them in vitreous ice [56]. This technique is the key to developing the cryoelectron microscopy (Cryo-EM) technique, which can determine protein structures in an aqueous environment [57]. In 2000, the National Institute of General Medical Sciences (NIGMS) funded the Protein Structure Initiative (PSI) to support protein structure determination [58]. Many automated tools have been developed to support high-throughput pipelines to solve complex structures, build computational models to predict 3D structure, and explore the function and potential medical impact of different protein structures [59]. In 1976, Johann Deisenhofer and Robert Huber, who

received the 1985 Nobel Prize in Chemistry, reported the first structure of membrane proteins [60]. In 1969, Martin Karplus developed a protein prediction program, known as Chemistry at HARvard Macromolecular Mechanics (CHARMM). He was awarded the 2013 Nobel Prize in Chemistry for "the development of multiscale models for complex chemical systems" [47]. In recent decades, due to the rapid increase in the gap between the number of known sequences (45 m) [61] and the number of determined structures (101,000) [51], highly effective computational structure prediction methods will play a key complementary role for protein structure determination [62]. According to the principles of thermodynamic theory [63], solved protein structures can provide information that can be used to develop knowledge-based protein determination techniques [62]. More protein structures determined from X-ray crystallography and nuclear magnetic resonance (NMR) leads to better templates and more accurate knowledge-based energy function for computational prediction methods.

The techniques for determing a protein 3D model are roughly classified into two types: experimental structure determination and protein structure prediction [62]. The former uses experimental methods to collect the structural information for a specific protein and generate a 3D model from that infromation. The current experimental techniques contains X-ray crystallography, NMR, and Cryo-EM. The prediction techniques extract the structural information from the solved structures and build the 3D models according to the principles obtained from the information. The prediction techniques can be categorized into template-based modeling and free modeling. Template-based modeling searches templates related to the target sequence and then aligns the target sequence to the template structure to generate structures for the target.

Depending on whether or not there are highly similar sequence templates in the PDB, template-based modeling is further classified into homology modeling and threading, [64].

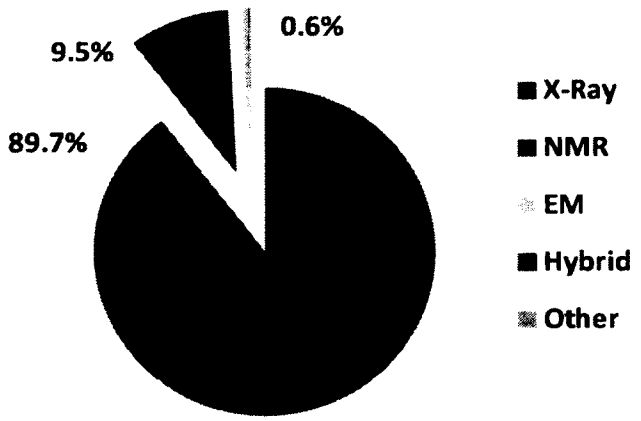


Figure 6. The number of protein structures solved by multiple methods [6].

Free modeling contains either 'ab initio' or 'de novo' modeling, which generates 3D models without templates. An ab initio prediction, such as a molecular dynamic (MD) simulation, uses the sequence information and the fundamental physical principles to search for a structure with minimum global energy [65; 66; 67; 68]. In addition to sequence information, de novo introduces Cryo-EM density maps to facilitate identification of the protein structure [69]. In the current PDB [6], 89.7% of the known structures were determined by X-ray crystallography, 9.5% were determined by NMR, 0.6% were determined by Cryo-EM, and the rest of the known structures were generated by either prediction

methods or hybrid methods (Figure 6).

Experimental Techniques

1. X-Ray Crystallography

An X-ray is a type of electromagnetic radiation that has a wavelength between 0.01 to 10 nanometers [70]. It was classified as an unknown type of radiation (X-ray) after Wilhelm Röntgen, who received the 1901 Nobel Prize in Physics, first discovered it in 1895 [71]. X-ray crystallography is used to determine crystal structures since the wavelength of an X-ray is similar to the size of atoms [72]. A crystal structure is composed of repeated unit cells along three principal directions that may not be perpendicular [73]. Using X-rays to strike the crystals, atoms through the electrons scatter the X-ray wave and generate the diffraction pattern of regular spots called reflections, which are two-dimensional (2D) images [70]. William Lawrence Bragg, and his father William Henry Bragg, proposed Bragg's law [74] in 1912, which provides the tool to convert those reflections into a 3D model of the density of electrons within the crystals. They shared the 1915 Nobel Prize in Physics for their contribution to crystallography [47]. This technique has been widely used to determine the structure of molecules and minerals.

Crystal structure determination has been studied and applied to inorganic crystals and organic crystals. The first X-ray structure of a protein, myoglobin, was reported by John Kendrew and Max Perutz in 1958 [46], who shared the Nobel Prize in Chemistry in 1962. To date, the PDB contains over 80,000 protein structures that are determined using this X-ray technique [6].

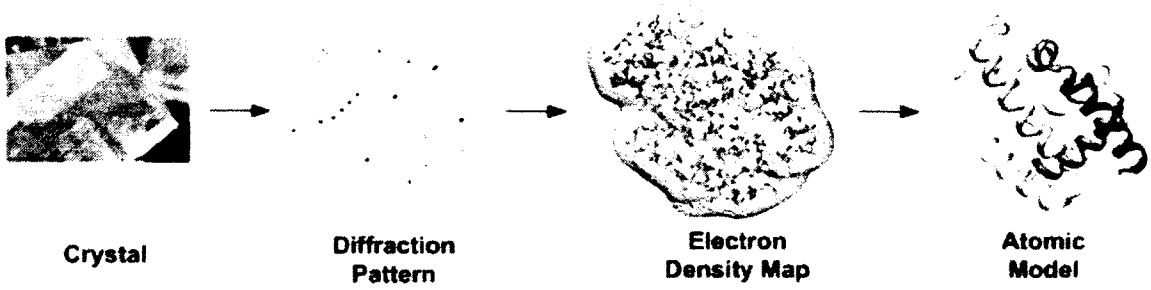


Figure 7. Flow for solving the atomic structure of proteins with X-rays.

Four basic steps are involved in resolving the molecule structure with X-ray diffraction [74]. The first step is to generate a single-crystal of the molecules. This step is the most difficult step in this technique because it is almost impossible to predict and time-consuming to obtain the appropriate crystallization condition for a specific molecule. Many proteins, such as membrane-bound proteins, appear to be stubbornly resistant to crystallization due to their special characteristics and structures [75]. In the second step, the single-crystal is subjected to an X-ray with a particular wavelength to obtain the regular pattern of reflections from various kinds of orientations [76]. This step usually generates thousands of 2D reflections. In the third step, these 2D reflections are converted into a 3D electron density map using Bragg's law [77]. This step is completed with the help of computational programs. Finally, a refined model of the atomic arrangement, the crystal structure, is generated with information about the chemical structure obtained from the other techniques [78].

Protein structures from X-ray diffraction still need complementary techniques to

overcome the drawback of the X-ray method. It is very difficult to obtain a single-crystal and sometimes it is impossible to generate this type of crystal for specific proteins [75]. Moreover, the crystal protein structure is a structure with a perfect atomic arrangement so it cannot represent the dynamic structures of a protein in a solvent [79]. NMR has the advantage of working in the solution.

2. NMR

Nuclear magnetic resonance (NMR) spectroscopy was proposed by Isidor Rabi in 1938 [14; 80], who was awarded the Nobel Prize in Physics in 1944 [81]. The development of NMR provided a powerful tool for understanding molecular structures. NMR not only generates the structural data but also provides more information on dynamics, conformational equilibrium, folding, and intra- and intermolecular interaction [80]. It has been widely used in determining molecule structures, drug screening and design, chemical analysis, and material science [82]. Kurt Wüthrich was awarded the Nobel Prize in Chemistry in 2002 for his study of applying NMR to biomolecules in solution, in particular for the determination of protein structures [54]. Since NMR tries to identify the relationship between target atoms, the protein structures generated from NMR contains many target structures instead of a single structure, which suggests that the possible structures fluctuate around the global energy minimum. More than 10,000 protein structures measured with NMR have been deposited into the PDB [6].

The NMR technique is based on a magnetic field that absorbs and emits electromagnetic radiation [83]. The orbits of atoms are further represented by angular momentum and magnetic moment. The magnetic moment with the same angular momentum has a $-\frac{1}{2}$ spin and a $\frac{1}{2}$ spin. These spins degenerate, which means that the

spins switch between two identical energy states. Thus, all the nuclides with even numbers of protons and/or neutrons have a total spin of zero, while all the nuclides with odd numbers of protons have a non-zero spin. The isotopes with a non- zero spin, such as ^{1}H, ^{13}C and ^{15}N, can be used in NMR spectroscopy [84]. The degenerate spin state can be split into two states with a different energy within an appropriate magnetic field. When the spin stays at the –½ state it is called the ground state. This spin will absorb energy and jump to the excited state (½ state) when electromagnetic radiation of the correct frequency is applied to this spin state. This frequency satisfies $\Delta E = h\nu_0$, in which h is Planck's constant and ν_0 is the radiation frequency. After a while, the spin then relaxes to the ground state by emitting magnetic radiation. All the same nuclei resonate at the same frequency if no other factors are involved. However, this frequency will be perturbed with the surrounding shells of electrons and cause a chemical shift [85]. Furthermore, the electrons on the neighboring bonded atom also inference this frequency and splits it into several peaks, which is called spin-spin coupling or J coupling. In organic synthesis, these chemical shifts and J couplings (correlation spectroscopy) are used to determine the molecular structure from the formula [83].

NMR cannot directly generate protein coordinates. J coupling represents the spin-spin coupling through the bond [83], which only represents the relationship between the connected atoms. The Nuclear Overhauser Effect (NOE) permits distance measurements between hydrogen nuclei through space [86]. It is possible to observe nuclei interactions when the connecting pairs of hydrogen atoms are separated by less than 5 Å [87]. All this information is used as the constraints to calculate the 3D protein structures with computer programs, such as the combined assignment and dynamics algorithm (CYANA) [88] or

XPLOR-NIH [89]. Inter-atomic distances and torsion angles are used to find an ensemble of structures consistent with the NMR constraints [90]. NMR structures are usually represented by a bundle of structures because the NMR constraints describe a range of possible values and many distances that have no exact value.

NMR is unable to deal with large proteins that are over 30 kDa [91], although it has the advantage in working with protein solution. The constraint information obtained with NMR represents the dynamics structures of proteins in an aqueous solution. These structures are natural, native structures of proteins instead of the crystal structure obtained under non-physiological conditions from X-rays. However, a large protein results in fast relaxation and broader lines in the NMR spectrum [92]. The corresponding spectrum has poor resolution and low sensitivity. Moreover, in a large protein, the more resonance lines from more NMR-active nuclei increase the spectral overlap [93].

3. Cryoelectron Microscopy

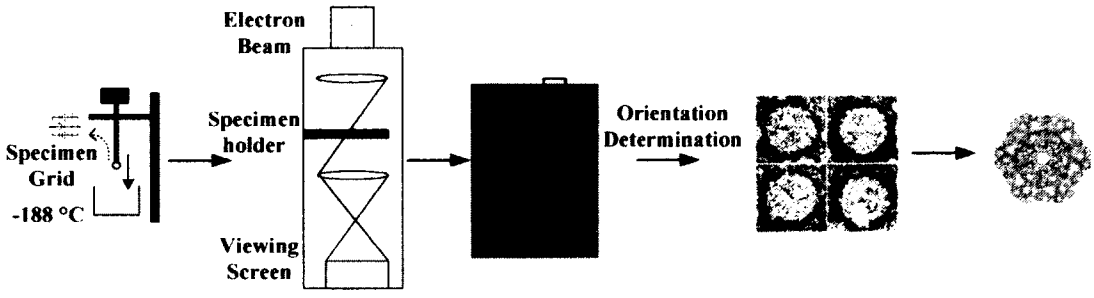


Figure 8. Procedure that builds a 3D electron density map using Cryo-EM [3].

Cryoelectron microscopy (Cryo-EM) provides an alternative technique for determining protein structures, especially with relatively large proteins (mass greater than 200 kDa) [94]. Cryo-EM uses transmission electron microscopy (TEM) to detect the molecular structure. TEM was built by Max Knoll and Ernst Ruska in 1931 in order to obtain significantly higher resolution than is possible with light microscopes [95], and the later was awarded the 1986 Nobel Prize in Physics [47]. This high resolution is due to the small de Broglie wavelength of electrons. The TEM image arises from the interactions of the electrons transmitted through the specimen. Direct images of biological specimens with impressive contrast have been obtained from specimens in the frozen state [96]. This technique is particularly suitable for the study and characterization of polymers, metals, and ceramic materials. However, the vacuum environment of TEM, which is used to avoid the scattering of electrons by gas molecules, is quite harsh to biomolecules. This environment can dehydrate proteins and destroy their structure. Cryo-EM has been used to make the sample tolerate the vacuum environment since Jacques Dubochet discovered that an aqueous solution of biological specimen can form a vitrified layer [97]. Rapid cooling of an aqueous protein solution in liquid ethane or liquid nitrogen temperature generates vitreous ice; in this condition, water would be immobilized before water molecules have time to crystallize and destroy cells. This technique preserves the natural state of proteins in the solution [98].

Three basic steps are needed to generate an electron density map using the Cryo-EM technique (Figure 8). The first step is to prepare the vitrified specimen [98]. The aqueous solution of proteins is dropped into the holes of a supporting grid. The self-supported water film spans the holes. This file is thin enough to transmit electrons, which

is, typically, less than 100 nm [99; 100]. This specimen grid is quickly frozen at -180° in liquid nitrogen [98]. In the second step, the structural information is generated using two approaches: single particle analysis (SPA) [101] and cryo-electron tomography (CryoET) [102]. SPA aligns the two-dimensional images from the same orientations to reduce the random noise. Resolution in the range of 7-10 Å needs about 300-100,000 images [103]. The number of protein molecules exposed under the electron beam limits the resolution. In contrast, rather than looking at a large number of projections, CryoET fixes one protein particle and collects the images by the controlled angles [102]. The resolution depends on how much electron exposure the protein molecule is able to tolerate before its structure is degraded by the electron beam. The information extracted from a single particle, such as the conformational changes that occur during protein binding, is beneficial in dynamic study. This approach has a resolution in the range of 20-40 Å [99]. Finally, the 2D images that are obtained are merged into a 3D image, which is referred as the molecular electron density map.

Although the resolution of Cryo-EM is increasing steadily, there is still a long way to go before it is able to reach the atomic level for general specimens. In the past 30 years, the resolution of Cryo-EM has improved from 35 Å for the Semliki forest virus [104] to atomic-level resolution [105; 106; 107; 108]. However, the resolution of most density maps is still greater than 6 Å [109]. The Electron Microscopy Data Bank (EMDB) [110] contains a total of 1897 map entries; of those, 34% have a resolution less than 20 Å, 35% have a resolution less than 10 Å, and 22% have a resolution ranging from 5 Å to 10 Å, only 6% have atomic-level resolution (Figure 9). For protein structure or quaternary structure that can be observed in the 10-30 Å range, for which the rigid-body fitting of known

structures is the primary method for modeling. The secondary structure can be extracted in the 8-10 Å range; α-helices are resolvable in the 8-10 Å range and β-sheets are resolvable in the 6-8 Å range [100]. In the 3-6 Å range, the full atomic model can be built directly with the existing X-ray modeling techniques [58; 106; 111; 112].

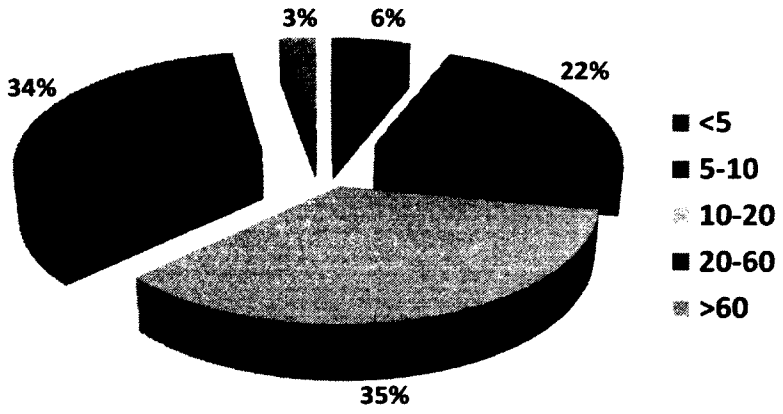


Figure 9. Resolution distribution of a density map in the Electron Microscopy Data Bank (EMDB) [7].

CryoEM is becoming a complementary tool of X-ray crystallography and NMR for analyzing large, uncrystallized structures. Although CryoEM currently lacks atomic-level resolution, it offers an opportunity to determine protein structures in their natural state. Rapid freezing prevents the rearrangement of water molecules into ice crystals and the rearrangement of the target protein. Combining computational prediction techniques,

it is possible to obtain the atomic structure from an intermediate-resolution density map.

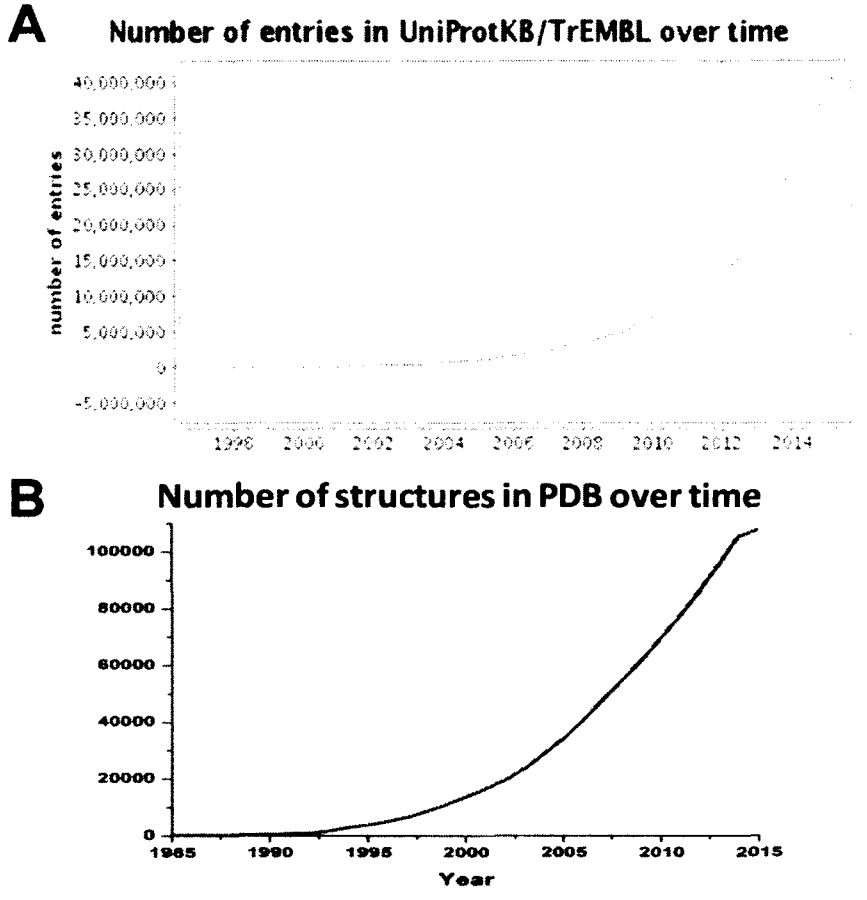


Figure 10. (A) The number of released protein sequences in UniProt [5] over time; (B) The number of released protein structures in PDB [6] over time.

Computational Prediction Techniques

Due to the rapidly increasing gap between the number of released sequences in UniProt [5] and the number of released structures in PDB [6], current molecular biology research is in urgent need of computational prediction tools that can help scientists identify protein structures from amino acid sequences. DNA is transcribed to RNA in the cell nucleus, which contains protein-coding region ("exons") and non-coding regions ("introns"). The splicing process cuts the introns and only retains the exons. The spliced RNA containing only exons leaves the cytoplasm to produce proteins [113]. The protein sequences were generated by the sliced RNA and they fold into a 3D model. The number of sequences in UniProt currently exceeds 45M [5] and increases by thousands of sequences each day. To understand the functions of proteins, we need to know the 3D structure for each protein sequence. However, the experimental techniques used to determine a protein structure are tedious and expensive. It might take months or years to successfully determine the structure for a specific protein. There are about 110 K protein structures in PDB [6]. Bridging this immense gap is almost an impossible task with experimental techniques. Thus, there is an urgency to develop highly effective computational techniques to predict 3D models from a 1D sequence.

In Critical Assessment of protein Structure Prediction (CASP), the protein structure prediction techniques were classified into template-based modeling and free modeling [64]. Although the template-based method has gained popularity over the free modeling methods because it has had greater success in achieving high resolution models [114], template-based modeling presents two major challenges: the selection of appropriate templates and the alignment. It is still rare to achieve accuracy above 80% for the target

proteins with a template that has less than 50% sequence identity by comparative modeling. These prediction errors indicate that the current known structures do not cover the complete information that is necessary for modeling. We checked the Cryo-EM density maps ranging between 6 Å and 10 Å [109] (~20% of all of the density maps). About 90% of the solved density maps are generated from the template with over 95% sequence similarity. Because it is difficult to further increase the model quality when using the template-based modeling method, the free modeling method could be a complementary strategy to obtain high quality models for the target protein.

1. Comparative Modeling

Comparative modeling, also known as homology modeling, constructs the atomic resolution model of the target protein from the protein templates of the known structures. These template proteins have a relatively high sequence similarity (> 30%) on the alignment that maps the target protein sequence to the template protein sequence [115; 116]. Based on biological observation, the proteins with similar amino acid sequences are usually evolutionarily related and have similar 3D structures [117]. Given a protein sequence, homologous protein structures could be used as the templates. Depending on the degree of similarity between the target and template sequences [115; 116], in recent years the predicted structure has been found to reach a 3.5 Å resolution, sometimes even 1Å [118].

The homology modeling procedure includes four steps [119; 120]: template selection, target-template alignment, model construction, and model assessment. The protein sequences in which the proteins have over 30% similarity display a high similarity in the 3D structure. Current multiple sequence alignment (PSI-BLAST) and Hidden Markov Models (HMMs) provide 80% accuracy for the sequence alignment. Usually several

candidate template structures are identified in this step. Then, the target sequence is aligned to the template structure. In the modeling step, the coordinates of the target protein can be generated using several methods: 1) single template refinement [121]; 2) fragment assembling and segment matching [122; 123]; and 3) spatial restraints with NMR spectroscopy or the electron density map of Cryo-EM [124; 125] in which Cryo-EM fitting contains rigid-body fitting and flexible fitting, which searches for the best fit between the template and electron density map with a cross-correlation coefficient [126]. Finally, the homology models are assessed with energy functions, which include knowledge-based potentials and physics-based potentials.

2. Threading

The threading modeling is a more sophisticated method that is used when the level of sequence identity is less than 30%. When no high sequence similarity exits and the templates are found, the protein model is still able to be built from the supper secondary structures (folds). First, based on the protein classification databases, such as the Structural Classification of Proteins (SCOP) [127] or the CATH Protein Structure Classification [128], a structure template database can be constructed that will remove all the protein structures with high sequence similarity. Second, a scoring function can be designed to measure the relationship between the sequence and the structure. Third, the target sequence can be aligned to the structure templates with a good score. Fourth, the most probable model with the best score will be selected as the predicted model. Since fewer and fewer new folds have been found in recent years [129], Zhang proposed that the target proteins in the current PDB that are less than 2.5 Å with over 82% alignment coverage always have similar folds[130; 131].

3. Ab initio Modeling

Ab initio modeling refers to an process that can predict the protein tertiary structure from its amino acid sequence based on the force field governing protein folding [21]. It is distinguished from template modeling, which uses known structures during the predication procedure. Ab initio modeling assumes that all the structure information is contained in the amino acid sequence; in other words, given a specific protein sequence, only one protein tertiary structure corresponds to it. This assumption was demonstrated in the 1950s by Christian Anfinsen [14]. Denatured ribonuclease A spontaneously refolded to its native tertiary structure and regained its function. Changing the psi and phi angles for each residue of the protein sequence can generate numerous protein models. The native structure for this protein sequence must be contained in the search space.

However, enumerating all the models is an impossible task. For example, each residue has 10 different conformations; a sequence with 100 residues has 10^{100} models [132]. It would take years and years of computational time to traverse all of them. The current ab initio approaches consider hybrid approaches guided by knowledge-based and physics-based potentials. Pure physics-based protein folding with MD simulation is able to generate a native-like conformation for a 100-residue long protein sequence with approximately 1000 CPU years [64]. ROSETTA built over 92 residues using 9-mers from other PDB proteins [133]. Despite the fact that 1.8 Å RMSD to the native structure was obtained, the computational cost was over 150 CPU days. I-TASSER, the current best modeling approach in CASP developed by Zhang, built protein models with various fragment sizes, improving the model size up to 155-residues long based on knowledge-based potential [64].

Despite the expensive computational cost, the ab inito approach has attracted the attention of many researchers because it is an eventual solution to protein structure prediction. The purely physics-based ab initio simulation identified the pathway of protein folding. However, due to the expensive cost and lack of accurate potential function, the best current results come from the combination of knowledge-based and physics-based approaches.

4. De novo Modeling

De novo modeling refers to building protein models from an electron density map and a protein sequence. This method is used for protein modeling when there are no appropriate templates that have more than 30% similarity with the target. The density maps with resolution in the 10-30 Å range are so-called low resolution density maps. No useful detailed structure information can be extracted from these density maps. The resolution in the 5-10 Å range is the intermediate resolution. These intermediate resolution density maps are unable to provide atomic structural information. However, the secondary structure (SSE) can be extracted from the intermediate density maps, which provides complementary structural information that can be used to model the protein structures. The dense mass in SSE causes higher electron density in the density map, in which α-helices are detected as rods and in β-sheets are detected as the plate areas [134; 135; 136; 137; 138].

Two major approaches can be used to generate the models for proteins using Cryo-EM density maps [132]. One method is the fitting and refinement method [125; 139; 140; 141; 142; 143; 144; 145; 146; 147; 148; 149], which is a template-based modeling approach, such as ROSETTA [150], MODELLER [124], S-FLEXFIT [151]. The other method is the de novo

approach, which uses the secondary structure elements and the skeleton to build the models. Density maps are used to reveal α-helices at about the 8-10 Å range and β-sheets at about the 6-8 Å range. Several programs provide the tools to identify these SSEs. HELIXHUNTER [152], EMATCH [138; 153], and HELIXTRACER [137] have been successfully applied to identify the α-helices. SHEETMINER [136], SHEETTRACER [97], and SSEHUNTER [154] were developed to detect the β-sheets. The skeletonization algorithm in SSEHUNTER [154] has been used to trace the backbone [155]. Several programs, including PHDpsi [156], Jufo [157; 158], PSIPRED [159], and PORTER [160], have predicted the SSEs in the sequence with up to 80% accuracy. Combining the SSE information in the sequence and in the structure, we expect that more medium resolution structures could be modeled.

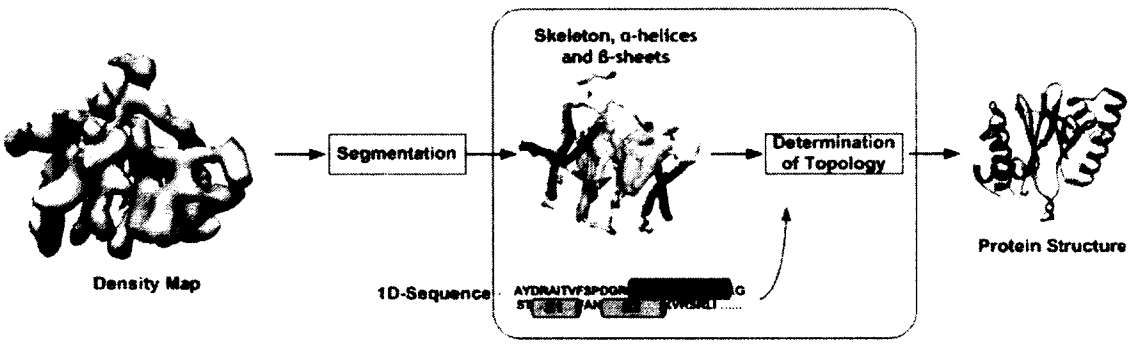


Figure 11. The flow chart for de novo modeling.

There are three basic steps in de novo modeling (Figure 11). First, the SSEs are identified from the Cryo-EM density map. From the discussion about Cryo-EM, previously presented in the this paper, the α-helices and β-sheets can both be recognized at the intermediate resolution. Second, the sequences that pass through the SSEs identified in the density map, known as the topologies, are traced or the SSEs in the sequence and in the structure corresponded. Due to the extremely large searching space, several algorithms have been developed to speed up the searching [69; 132; 161; 162; 163; 164]. Finally, based on the topologies, the backbone is placed into the density map, the side-chain is added and optimized, guided by the folding principles [165].

Three major factors limit the application of the de novo method: 1) it lacks an accurate energy function locates the global minimum for the native protein structure; 2) it lacks an efficient search algorithm covers the conformational space without missing the native conformation; and 3) it is unable to select a native-like structure from the decoy structures. For a long protein sequence, enumerating the conformations is extremely expensive. The efficient search approaches must be able to reduce this huge search space to a reasonable size. The reduced search space must contain the native structure. Exploring the entire conformation space would generate plenty of decoys for the target protein. An accurate energy function is needed to distinguish the native structure from the modeled conformations and guide the conformation optimization.

This work focused on generating an accurate protein energy function and reducing the topology searching space. The physical interactions between residues in a tertiary structure are described by energy functions. In other words, the protein energy

function is a score function to evaluate the stability of the protein conformations. The native structure for a specific sequence has the highest stability and the lowest energy. An accurate energy function is the major obstruction for protein structure prediction. Chapter 2 introduces how to generate an energy function using the statistic thermodynamics theories and the datasets from PDB in our work. By evaluating several widely used benchmarks, our energy function is able to surpass the most popular energy functions currently being used for protein structure prediction. Chapter 3 addresses how to improve the pruning algorithm developed by Dr. Kamal Al Nasr. To reduce the search space when searching the correspondence between the SSEs in the sequence and in the structure, Dr. Kamal Al Nasr designed an efficient algorithm to identify the top-K topologies for pure α-helices proteins. This present work extends his algorithm to search the top-K topologies for proteins that have both α-helices and β-sheets by considering the features of β-sheets that occur in nature.

CHAPTER 2

PROTEIN ENERGY FUNCTION DESIGN

One of the most challenging tasks in protein tertiary structure prediction is to distinguish the native conformation of a protein among the decoys that have similar conformations. In thermodynamics, Gibbs free energy (G) is used to evaluate stability of the protein structure, which is taken as an isolated thermodynamic system.

As shown in equation 1, U is the internal energy, which is a scalar of temperature [166]. Higher internal energy causes atoms in the isolated system to move faster and to increase system temperature. P represents pressure, V stands for volume, and T is for temperature. In our research, since the residue number is constant for a specific protein sequence, we assume that P, V, and T are constant for all the protein conformations generated from this sequence. S represents the entropy [63], which is a measure of disorder. A system consisting of well-arranged atoms has a low entropy. In contrast, a chaos system has a high entropy. The third law of thermodynamics states that the entropy of a perfect crystal equals zero [166]. However, the disorder of the system cannot be measured directly, and the entropy difference between two protein conformations cannot not be calculated directly.

$$G = U + PV - TS \qquad (1)$$

The corresponding partial derivative equation is:

$$dG = dU + VdP + PdV - SdT - TdS \qquad (2)$$

In an isothermal, isobaric, and isochoric environment, the above equation equates to:

$$dG = dU - TdS = dF \qquad (3)$$

In equation 3, F represents Helmholtz free energy and is related to the partition function Q:

$$F = -k_B T ln Q = -k_B T ln \sum_i \exp(-\frac{E_i}{k_B T}) \qquad (4)$$

Where E_i is the energy at state i. Thus, the relative Gibbs free energy has the relationship [63, 167]:

$$dG = -k_B T ln g(r) = -k_B T ln \frac{\rho_r}{\bar{\rho}} \qquad (5)$$

In equation 5, *g(r)* is the paired distribution function, ρ_r is the density at distance *r*, and $\bar{\rho}$ is the density for the bulk system. With equation 5, we are able to design a known-based nergy function from the native protein structures in the Protein Data Bank (PDB) [51].

There are two types of energy functions, in general. The physical-based functions, such as CHARMM [168] and AMBER [169], are built upon the principles of physics. Those energy functions usually ignore the energy contribution from the entropy. In this case, *dG* = *dU*, as the internal energy was used to evaluate the stability of the protein system. The internal energy contains both the bonded energy and the non-bonded energy. The bonded energy contains both rotational and vibrational energy, but no transfer energy. The non-bonded energy contains the contributions from electrostatic force (Coulomb force) and non-electrostatic force (van der Waals force, dispersion force) [170]. At room temperature (298K), *TdS* in (3) could be significant. Due to the missing of entropy term in equation (3), the performance of the physical-based functions is very far from what is expected under room temperature. In contrast, the knowledge-based energy functions represent the statistics extracted from large number of known structures [171, 172] with equation (5).

Knowledge-based energy functions, or so-called "statistical potentials" have been used in numerous applications, such as structure prediction [173], protein design [174] and docking [175]. As shown in equation (5), knowledge-based energy is Gibbs free energy, which contains the contributions from both internal energy (U) and entropy (S). For this reason, the knowledge-based energy functions have much better performance than the physics-based energy functions. The only limits of accuracy for the knowledge-based energy functions are the number of protein structures in the PDB and an appropriate reference state (bulk density), which represents the environment for an interaction between residues.

In spite of the successful cases demonstrated by the statistical potentials, it is challenging to develop an energy function that approximates well in various physical environments. Some statistical energy functions use all-atom interactions, such as DFIRE [176] and DOPE [177], while others use reduced representations for amino acids. Although the all-atom functions characterize the fine details of a conformation, it is challenging to represent the dependencies among the atoms that are connected by one or more consecutive covalent bonds. Various reduced representations, or "coarse-grained models," have been proposed. Some of them use the Cα atom [172, 178] or the side chain center to represent each amino acid [179; 180; 181]. OPUS-PSP breaks an amino acid into multiple blocks and uses nineteen blocks to represent twenty amino acids [182]. Random-Walk function uses twenty vector-pairs on the side chain to represent an amino acid [183].

In addition to the above mentioned pair-wise functions, three-body and four-body potentials have been investigated by various groups. It has been suggested that the pair-wise potentials are not sufficient to characterize the three-dimensional interactions due to the simple decomposition of such interactions to two-dimensional problems [184; 185].

Krishnamoorthy and Tropsha use Delaunay Tessellation to derive a four-body function [186]. Feng, *et al.* extended a two-body potential to a four-body potential [187]. In spite of the theoretical advantages, multi-body potentials have yet to outperform pairwise functions in distinguishing the native from the decoys in large datasets.

One of the advantages for pairwise coarse-grained potentials is the simplicity in describing both the distance and relative orientation of a pair. Earlier pairwise functions are primarily based on the distance between the pair [171; 172; 176; 188; 189; 190]. Recently, the relative orientation of the pair has been incorporated [182; 183; 191; 192; 193]. The block representation of OPUS-PSP groups the rigid portion of the chemical structure into a block, but still provides the flexibility in representing the side chain. In spite of the innovative block representation, OPUS-PSP is an orientation-dependent, but not a distance-dependent function. In principle, the joint distance and orientation function should be more sensitive in distinguishing the fine conformational differences. In practice, this is not feasible until sufficient representative data are available.

The CABS model incorporates both distance and orientation in the potential function, although the number of orientations is limited [194]. In spite of the recent attempts [195; 196] it is still challenging to derive an effective function that is both distance- and orientation-dependent. In our study, we present a both distance- and orientation-dependent function, DOKB, that is based on the block representation. We illustrate the importance of using both distance and orientation in characterizing the pairwise potential.

Side chain packing is one of the most important factors used to distinguish one conformation from another. In the block representation of OPUS-PSP, most side chains consist of multiple blocks. In principle, all blocks of an amino acid should be used in

calculating the energy. However, backbone-backbone interaction is not specific in distinguishing the conformations. Although the interaction between backbone and side chains is more specific, it is challenging to accurately represent the dependency among multiple blocks. Previous studies have shown the dependency between the backbone and the side chain conformations [197].

A large number of known structures is required to derive statistically meaningful dependency among multiple blocks when relative distances and orientations are encoded. We hypothesize that some blocks in a side chain are more influential than others. In fact, representations that bias the functional group of the side chain were proposed for 9 of the 20 amino acids [198]. In this paper, we present the results of a simple and effective approach that uses a key block to represent each side chain, except TYR and ILE, for which two are used. The minimum representation using key blocks can highlight the most characteristic portion of the side chain during packing.

Protein structures are known to present as a scale-free interaction network [199; 200; 201; 202] in which clusters or "hot-spots" play critical roles in stability. The densely packed clusters presumably have the most constraints in packing the side chains, and they are perhaps the regions to identify the difference between a native structure and a decoy.

One of the drawbacks of a pairwise potential is that it does not distinguish the local environment of the pair. For instance, a pair with the relatively same geometry has the same potential, regardless of where it is located. Multi-body potentials aim to fix this drawback, although it is not clear if replacing the pairwise potential with multi-body potential for all regions is an effective approach. We characterized the residue pairs in the low-energy cluster within the highly packed clusters and translated the knowledge to an

energy term in an attempt to incorporate the energy difference between a highly packed environment and a loosely packed environment. The added cluster energy term appears to improve the performance in the ig-structal dataset.

2.1 Method

2.1.1 Definition of the Relative Geometry of a Pair of Key Blocks

Based on our understanding of protein energy, blocks in OPUS-PSP were used to represent the interaction between residues. The residue-level interaction uses Cα/Cβ or side-chain center to represent the position of residues. It is a concise description of the residue and simplifies the calculation. However, it not only ignores the conformation of the side-chain, but also takes a very rough approximation of the distance between functional atoms. The atom-level model keeps all the interaction information of the non-bonded atoms in the residues, but it ignores all the connection information within the residue. We used the block-level model in an attempt to obtain a balance between the residue-level and the atom-level. It contains the necessary connection information of the bonded atoms without increasing calculation burden.

DOKB borrowed the definition of blocks from OPUS-PSP [182]. As shown in Figure 12, 20 residues are decomposed into 19 rigid-body blocks. The interaction between two residues is converted into the summation of blocks interaction. This definition assumes that all heavy atoms are in the same plane, and blocks consist of the bonded heavy atoms. Each block only appears once in each residue. Based on the block shapes, all blocks are categorized into three classes: point blocks, linear blocks, and plane blocks.

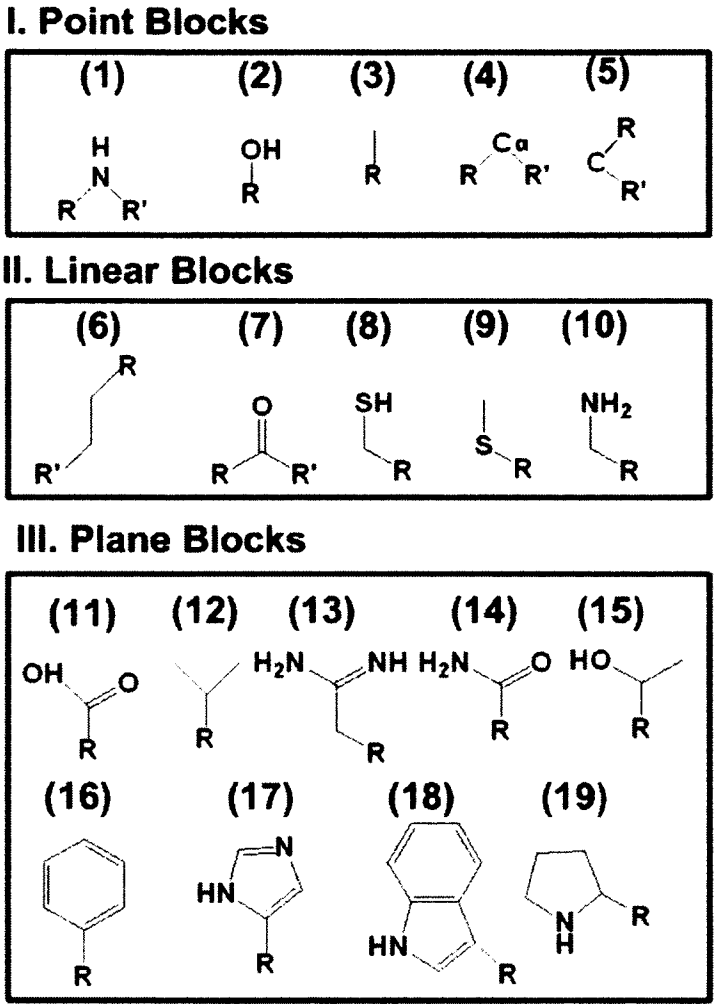


Figure 12. The definition of 19 rigid-body blocks in OPUS-PSP [2]. R and R' are not parts of blocks.

The block representation in DOKB is based on OPUS-PSP [182] with some modifications (Table 1). Instead of using multiple blocks to represent an amino acid, a key block at the distal end of the side chain was used, except for ILE and Tyr (Table 1).

The backbone blocks are not included in DOKB, and the energy solely calibrates the packing of the key blocks in the side chains. Unlike in Lu [182], the 22 key blocks are amino acid specific.

The local frame of each block was defined according to Lu [182]. If the central block i is plane block (class III), the relative geometrical relationship between block i and j is represented by (r, θ, ϕ), in which r is the center-to-center distance between i and j. θ and ϕ characterize the orientation of j block in i's frame (Figure 13). In particular, θ is the angle formed by AB and z-axis, and ϕ is the angle formed by x-axis and the projection of AB on the xy plane (Figure 13). A pair of key blocks is considered in energy calculation if $r \leq 7$Å, slightly larger than the popular contact cutoff of 6.5Å between C_α atoms in order to consider more pairs. Since the number of protein structures is limited, we partitioned the geometrical space of a pair of blocks into bins of $(\Delta r, \Delta\theta, \Delta\phi)$, in which $\Delta r = 0.5$Å, and $\Delta\theta = \Delta\emptyset = 30°$. If the central block i is linear block (class II), the geometrical relationship between block i and j is represented by (r, θ). There are 6 bins from 0° to 180°. If the central block i is point block (class I), there is no orientation energy and only distance energy is calculated.

2.1.2 The Energy Function

The statistical energy function was developed using a data set of 4,180 known protein structures that were extracted using PISCES [203]. The data set contains those structures that were solved by X-ray crystallography and have (1) no more than 40% of sequence similarity; (2) at least 1.8 Å resolution; (3) an R-factor of 0.25 or better and (4) at least 40 amino acids in the sequence.

Table 1. Definition of key blocks [2]

Residue	Key block[a]	Block[b]	Residue	Key block	Block
GLY	None	None	THR	15	15
SER	2	2	PHE	16	16
ALA	3	3	HIS	17	17
CYS	8	8	TRP	18	18
MET	9	9	PRO	19	19
LYS	10	10	ILE	20, 21	12,3
ASP	11	11	LEU	22	12
VAL	12	12	TYR	23, 26	2, 16
ARG	13	13	GLU	24	11
ASN	14	14	GLN	25	14

[a] Key block index in DOKB.
[b] Block index in OPUS-PSP.

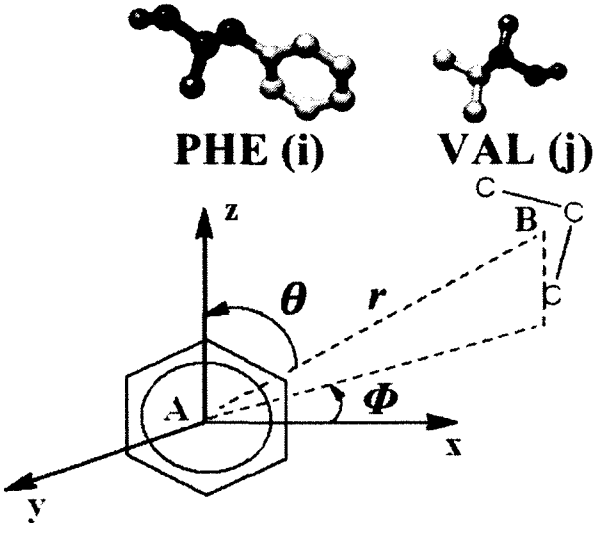


Figure 13. The distance and orientation representation of a pair of key blocks [2]. The key blocks (cyan) of PHE and VAL are represented by the local frame centering at A and B respectively. The distance between A and B is r. θ and ϕ are the angles of B in A's local frame.

The energy of the entire protein E_{Total} contains the energy from all key block pairs using eq. (6). The energy between a pair of key blocks contains the distance energy $E_{dist}^{i,j}(r)$ and the orientation energy $E_{ort}^{i,j}(r,\theta,\varphi)$, where i and j are the block indexes in Table 1. Each residue is represented with its most distal block of the side chain (Table 1) except for ILE and TYR that are represented by two blocks. Note that $E^{i,j}$ may not be the same as $E^{j,i}$ since the orientation energy $E_{ort}^{i,j}(r,\theta,\phi)$ depends on the orientation of block j in the local frame of block i.

$$E_{Total} = E_{dist} + E_{ort} = \sum_i \sum_j E^{i,j}(r,\theta,\phi) = \sum_i \sum_j (E_{dist}^{i,j}(r) + E_{ort}^{i,j}(r,\theta,\phi)) \quad (6)$$

The observed density $\rho^{i,j}(r)$ (7) at distance r was calculated by $N^{i,j}(r)/4\pi r^2 \Delta r$, in which $N^{i,j}(r)$ is the number of the observed block pairs in bin $floor(r/0.5)$. $\rho^{i,j}(r)$ was derived for $0 < r \leq 25$Å. The reference state, $\bar{\rho}^{i,j}$, uses the ideal gas state that is supposed to be the density at infinite distance. We observed that the density is approximately constant when $15\text{Å} \leq r \leq 20\text{Å}$. This character was similarly reported in [176]. Due to the limit of the protein size in the training data, density may not be realistic when $r > 20$Å. Therefore, we used an average density calculated from $15\text{Å} \leq r \leq 20\text{Å}$ as the expected density.

$$E_{dist}^{i,j}(r) = -k_B T ln g(r) = -k_B T ln \frac{\rho^{i,j}(r)}{\bar{\rho}^{i,j}} \quad (7)$$

The orientation energy $E_{ort}^{i,j}(r,\theta,\phi)$ (8) was designed to adjust the distance energy $E_{dist}^{i,j}(r)$ that represents the average energy at distance r. $N^{i,j}(r,\theta,\phi)$ is the number of block pairs observed at distance r with orientation (θ,ϕ), and $\bar{N}^{i,j}(r)$ is the average number of the block pairs (i,j) of all the orientations with distance r. In

particular, $\bar{N}^{i,j}(r) = N^{i,j}_{total}(r)/\#bin$, where $N^{i,j}_{total}(r)$ is the total number of block pairs (i,j) at distance r.

$$E^{i,j}_{ort}(r,\theta,\phi) = -k_B T ln \frac{N^{i,j}(r,\theta,\phi)}{\bar{N}^{i,j}(r)} \qquad (8)$$

The cluster energy is calculated at the residue level instead of the block level to ensure sufficient low-energy, highly packed cases are available for all pairs. A pair of residues (m,n) is in the highly packed cluster if m has at least 15 neighbors. A pair of residues (m,n) is in a low-energy region if the energy at m is no more than -$15k_BT$. The energy at m is simply the summation of the pairwise energy for all residues that are neighbors of m. A residue n is considered a neighbor of residue m if n has a block within 7 Å to any block of residue m. Although the highly packed clusters often have low energy for the cluster center residue, it is not absolutely necessary. $P_{cluster}(m,n)$ is the probability for (m,n) in the low-energy region to appear in a highly packed cluster (Figure 20). $P_{all}(m,n)$ is the probability for (m,n) in the low energy region to appear in the entire structure regardless of highly packed or loosely packed regions (Figure 21). In particular, $N_{Low,highly\ packed}(m,n)$ is the number of (m,n) pairs in which m has no more than$-15k_BT$ kcal/mol and (m,n) is in a highly packed cluster. $N_{highly\ packed}(m,n)$ is the number of (m,n) pairs in which (m,n) is in a highly packed cluster. $N_{Low}(m,n)$ is the number of (m,n) pairs in the low-energy regions, and $N_{all}(m,n)$ is the number of (m,n) pairs in the entire structure regardless of where it is located. Alternatively, the energy of a protein can be calculated using (12) if the cluster energy is a concern.

$$P_{cluster}(m,n) = \frac{N_{Low,highly\ packed}(m,n)}{N_{highly\ packed}(m,n)} \qquad (9)$$

$$P_{all}(m,n) = \frac{N_{Low}(m,n)}{N_{all}(m.n)} \tag{10}$$

$$E_{cluster}^{m,n} = \begin{cases} -k_B T ln \frac{P_{cluster}(m,n)}{P_{all}(m,n)} & m \text{ is highly packed} \\ 0, & m \text{ is not highly packed} \end{cases} \tag{11}$$

$$E_{Total} = \sum_i \sum_j (E_{dist}^{i,j}(r) + E_{ort}^{i,j}(r,\theta,\varphi)) + \sum_m \sum_n bE_{cluster}^{m,n} \tag{12}$$

2.1.3 Energy Function Generation

For convenience, we generated a web-based energy function table and posed it at: http://www.cs.odu.edu/~jhe/software/DOKB/Block.htm. The webpage frame is shown in Figure 14.

Each cell in the 20*20 array contains a link to the distance energy (eq. 7) and the orientation energy (eq. 8) for the specific residue pair. The 20*20 array contains all 400 residue pairs. For cell (i, j), i represents the residue in the row and j represents the residue in the column. Since the orientation defined in Figure 13 is not symmetric, mostly $E_{ort}^{i,j} \neq E_{ort}^{j,i}$, the energy table is not symmetric either. We are using (ASP, ARG) pair as the sample to show the procedure of generating the distance energy and the orientation energy.

Click cell (ASP, ARG) on the energy table to display the "Information between ASP and ARG" page (Figure 14). This page consists of the density distribution table (top) and the corresponding energy table (bottom).

In the density distribution table, the first column contains the links of the distance density distribution pages for the block pairs between ASP and ARG (marked with the red rectangle in Figure 14). Click any of the links in the first column. There is a two-column distance distribution table (not shown in Figure 14). In this table, the numbers in the first column are the distance value between blocks from 0Å to 24.5Å with interval 0.5

Å, while the numbers in the second column are the density value at each distance. If block i in pair (i, j) is point block, the pair has only distance energy and no orientation distribution. If block i in block pair (i, j) is a linear block or plane block, the links for the orientation distributions at each distance are attached after the distance density distribution (circled with red in Figure 14). The orientation distribution page contains a table (each bin has 30° range, 1*6 array for linear block, 6*12 array for plane block). Each cell in the table contains the ratio value ($\frac{N^{i,j}(r,\theta,\phi)}{\bar{N}^{i,j}(r)}$ in eq. 8) and the pair number in the whole dataset for this orientation and distance (r, θ, φ).

Similar to the density distribution table, the first column in the energy table contains the links of the distance energy pages for the block pairs between ASP and ARG (highlighted with the red rectangle in Figure 14). Each distance energy page contains a two-column table. The first column contains the distance value from 0Å to 24.5Å with interval 0.5Å. The second column contains the energy value at each distance. Since there is no energy when the distance equals to zero, "nan" is used to fill the space. When two blocks are placed very close, the energy is increasing rapidly to infinite. A large number "21474" is used here to represent the infinite value. When calculating the energy of a protein, value 9 is used to represent the infinite value for easy plot purpose.

Step 1. Generate the density distribution function

According to eq. 7, the density for all 30 block pairs of (ASP, ARG) is plotted in Figure 15. ASP consists of block 1, 4, 5, 7, 11, and ASN consists of block 1, 4, 5, 6, 7, 13. There are total 5*6 block pairs. Each pair density distribution (i, j) represents the packing conformation of block j around block i. When the distance between two blocks are long enough, the interaction between these two blocks could be ignored. In other

words, the energy between these two blocks is zero and the density value is constant after 15Å [176]. However, since the size of the proteins is not infinite, the density value is not real constant and decreases rapidly after 20Å. Based on the protein size in our dataset, we assume the density within 15Å-20Å is constant and pick this range as the constant range.

To convert the density distribution to the pair correlation function, we pick the reference density value from the constant range. $\bar{\rho}^{i,j}$ Within the constant range, each density plot in Figure 15 has one average density value. 30 pairs have 30 different average densities. Even when all the block pair distributions are collected from a same residue pair, the average densities are slightly different. We picked the median value from these 30 average density values to be the reference density value. This reference value was used to generate the pair correlation function for each block pair next step. This value (17.1) was also taken as the density distribution of (ASP, ARG) and written in the main webpage (left in Figure 14).

Step 2. Generate the pair correlation function

The pair correlation function was generated from the density distribution function by the reference value obtained in last step. We divided all 30 density distribution plots of (ASP, ARG) by the reference value. The resulting plots are called "pair correlation function," which describes the packing conformation between rigid balls (Figure 16). The red line marks the value 1, which is the value for the reference state. For the plots, there most likely is more than one peak for each block pair. These show us that, for the central block i, the neighbor block j has two or three preferred distances. This observation is quite similar to the ideal gas packing of single atoms.

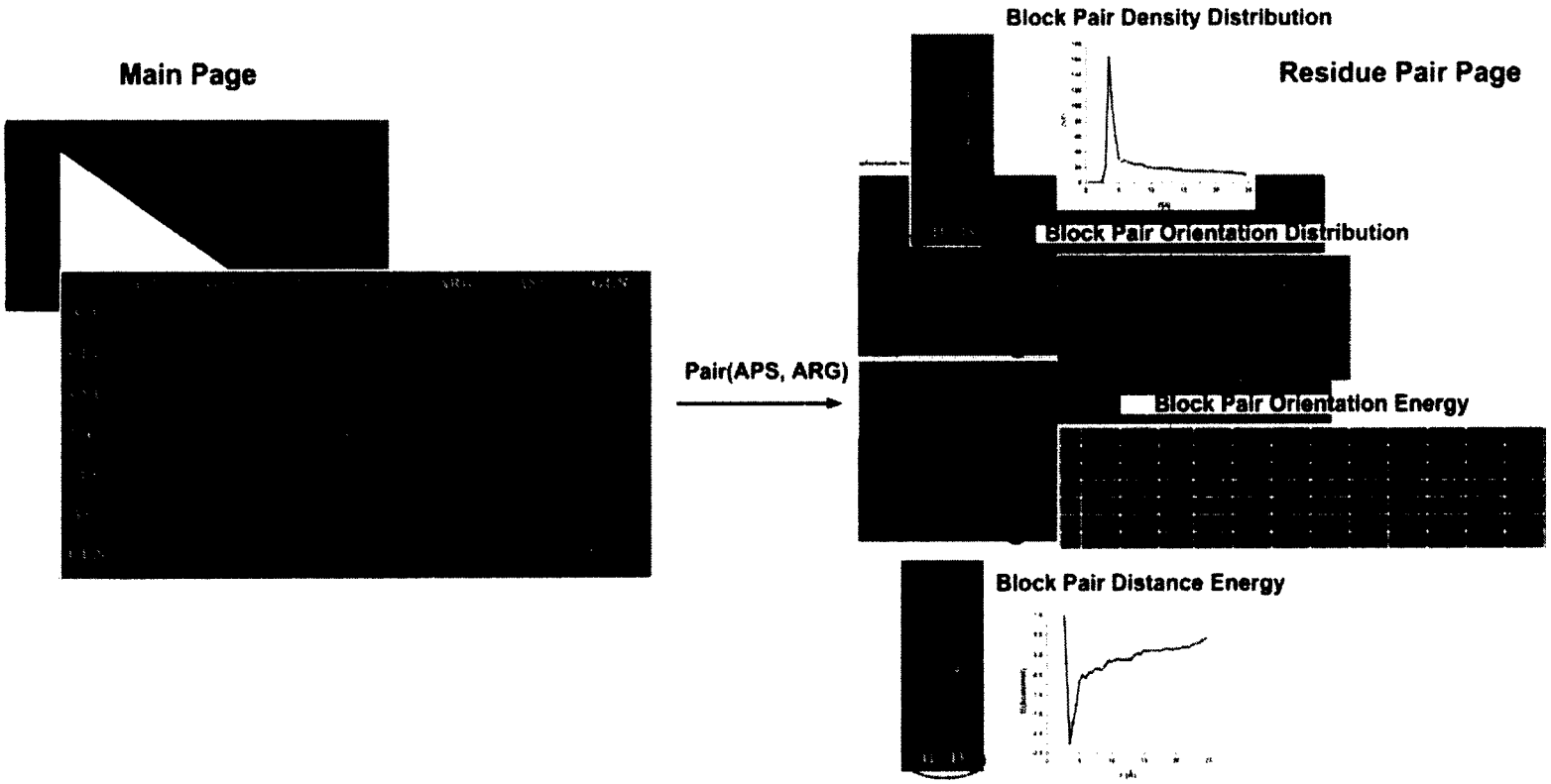


Figure 14. Web-based energy function. Main page contains 20*20 cells, in which each cell represents the interaction of one residue pair. For each cell link, there is a residue pair page, which contains the density distribution table (up) and the energy table (bottom).

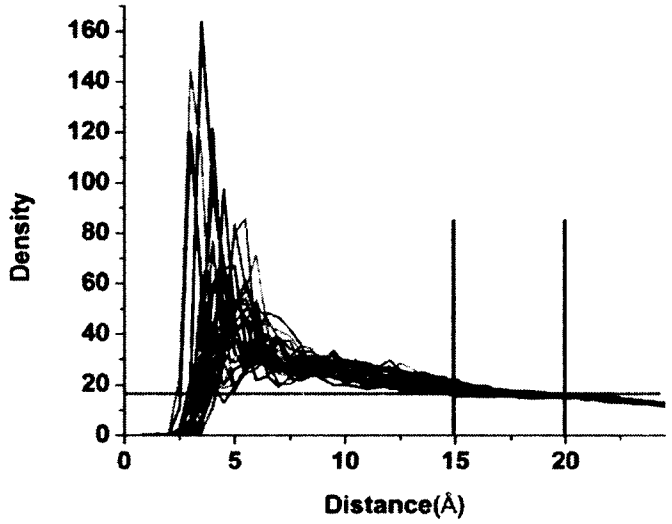


Figure 15. The density distributions for all 30 block pairs of ASP-ARG. The range between 15Å and 20Å are constant value range. The red line marks the bulk density value for 30 block pairs.

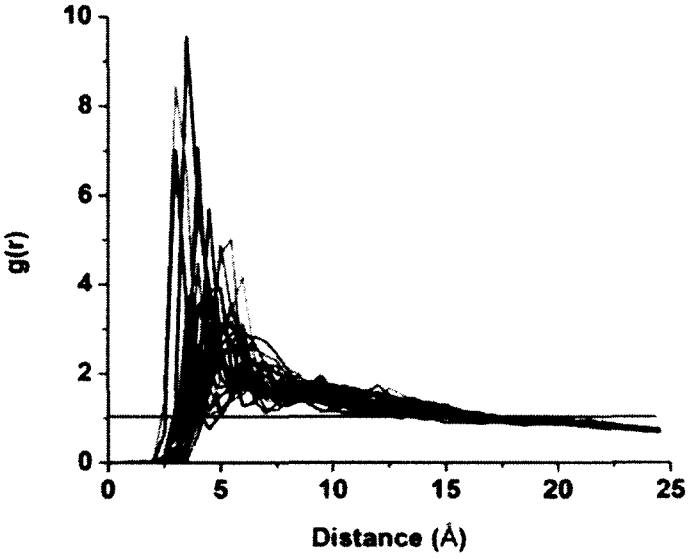


Figure 16. The pair correlation functions for all 30 block pairs of ASP-ARG. The red line represents the reference state, whose g(r) value is 1.

Step 3. Generate the distance energy function

With equation 5, the distance energy function was converted to the pair correlation function between blocks (Figure 17). The red line marks the zero energy. When two blocks are far enough away, there is no interaction, and the energy between them is zero. When two blocks are very close, they repulse or overlap each other, and the energy between them increases rapidly to infinity. Some block pairs have unfavorable positions, even though the distance is not close. In Figure 17, the plot of block pair (11, 7) has a valley at 3.0Å, which is the most favorable packing distance. However, at 4.5Å, there is an energy peak of 0.259, which represents a very unfavorable distance between these two blocks. After that, another valley appears around 6.0Å. From the ideal gas packing model for block pair (i, j), we know that the first valley represents the first packing shell around block i and the second valley represents the second packing shell around block i. The unfavorable distance at 4.5Å is a position between two shells. Block j could not be placed at this distance without increasing the system energy.

Step 4. Generate the orientation function

The orientation energy function was generated with equation 8. In step 3, we generated the distance energy function. From a statistical standpoint, the energy at each specific distance is the average energy of all possible orientations. Some orientations are favorable (the interaction energy should be less than the distance energy) while others are unfavorable (the interaction energy should be larger than the distance energy). For block pair (i, j), if i is the linear or the planar block, the orientation energy was calculated.

At each distance, the orientation was represented with (θ, φ) (Figure 13). θ ranges from 0° to 180°. φ ranges from 0° to 360°. We divided the range into several bins, each

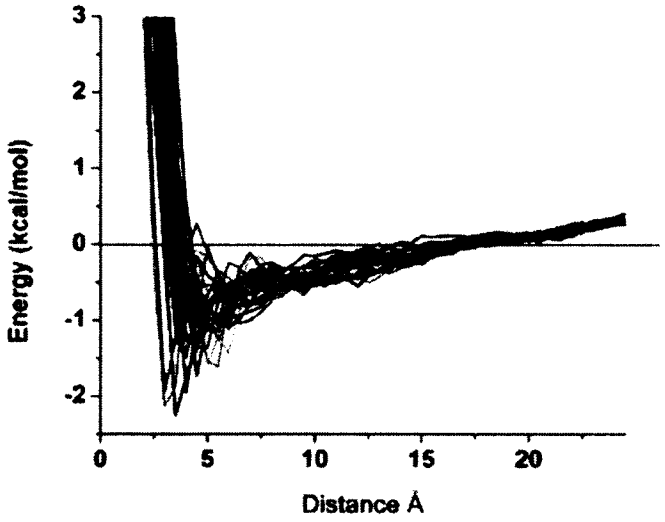


Figure 17. The distance energy functions for all 30 block pairs. The red line marks the zero energy; the energy value below zero means preferred

with a range of 30°. Then, all possible orientations were simplified to a 6*12 table. The average pair number was obtained by dividing the total pair number at this distance with 6*12 for planar block. Each cell contains the pair number in this specific distance and orientation and the ratio between the pair number and the average pair number. Then we used equation 8 to calculate the orientation with the ratio value. When the pair number is greater than the average number, the orientation is a favorite orientation. The orientation energy is a negative value. Otherwise, the orientation energy is positive value (Figure 14).

Step 5. Cluster Energy

The cluster energy was generated to fix the bias from the dense area of the proteins. The distance energy and the orientation energy in step 1-4 were generated for all block pairs without distinguishing the dense area or the loose area. The function is an average value for all areas. However, the dense area obviously has a different entropy value. The messy area has higher entropy, while the highly organized area has lower entropy. To describe the energy contribution of the dense area, we introduced the cluster energy.

The cluster energy is necessary to distinguish the native structure from the highly similar decoys. We were using five decoy sets to evaluate the energy function, DecoyRus, MOULDER, hg, ig, and ITASSER. We used the definition for the high dense residue in 2.1.2. The percentage of highly dense residues for each decoy was calculated. Since each protein has many decoys, the mean value for each protein was posted in Table 2. The dense percentage in Decoy'R'us is from 4.74% to 26.34%, and only 4 of them are over 20%. The high dense percentage in hg is from 9.17% to 18.52%, and in ig is from 25% to 30.72%, in MOULDER is from 8.57% to 36.62%, and in ITASSER is from 2% to 19.5%. The energy with only the distance energy and the orientation energy performed very poor in decoy set ig. Without the cluster energy, the energy was not sensitive enough to distinguish the native from decoys. This is because the decoys in ig have very high similarities. They have almost the same backbone conformations and slightly different side-chain conformations. The distance energy and the orientation energy are very close. The entropy for the dense area cannot be ignored anymore.

Table 2. The percentage of high dense residues in decoy sets

DecoyRus	Mean (%)	DecoyRus	Mean (%)	hg	Mean (%)	hg	Mean (%)
1ctf	18.19	1dkt-A	26.34	1ash	15.9	1mba	13.95
1r69	13.88	1fca	14.36	1bab-B	13.35	1mbs	18.52
1sn3	20.57	1nkl		1col-A	9.17	1myg-A	17.04
2cro		1pgb	9.26	1cpc-A	10.39	1myj-A	16.51
3icb	18.53	1trl-A	20.12	1ecd	10.17	1myt	15.77
4pti	15.81	4icb	26.72	1emy	17.32	2dhb-A	14.28
4rxn		1b0n-B	0.04	1flp	13.19	2dhb-B	14.7
1fc2	2.19	1bba	0.07	1gdm	16.56	2lhb	15.57
1hdd-C	0.35	1ctf	10.79	1hbg	16.35	2pgh-A	14.47
2cro	6.21	1dtk	5.8	1hbh-A	14.48	2pgh-B	13.42
4icb	0.53	1fc2	4.74	1hbh-B	15.23	4sdh-A	13.4
1bg8-A	6.41	1igd	11.59	1hda-A	13.9		
1bl0	15.15	1shf-A	10.89	1hda-B	12.85		
1eh2	14.14	2cro	2.31	1hlb	14.1		
1jwe	13.96	2ovo	9.47	1hlm	15.9		
smd3	9.52	4pti	6.04	1hsy	18.08		
1beo				1ith-A	13.57		
1ctf	22.36			1lht	17.9		

Table 2. Continued

ig	Mean (%)	ig	Mean (%)	ig	Mean (%)	ig	Mean (%)
1acy	28.5	1gaf	26.5	1mfa	26.1	2fbj	27.69
1baf	27.63	1ggi	27.52	1mlb	28.43	2gfb	28.51
1bbd	28.12	1gig	26.83	1mrd	27.22	3hfl	26.43
1bbj	27.1	1hil	29.27	1nbv	29.55	3hfm	25.98
1dbb	27.99	1hkl	27.11	1ncb	27.99	6fab	30.72
1dfb	29.32	1iai	28.64	1ngq	26.77	7fab	25.08
1dvf	28.3	1ibg	28.79	1nmb	29.44	8fab	27.16
1eap	27.26	1igc	28.46	1nsn	25		
1fai	29.76	1igf	27.74	1opg	29.056		
1fbi	29.37	1igi	28.68	1plg	27.57		
1fgv	27.78	1igm	29.04	1rmf	28.51		
1fig	27.93	1ikf	28.34	1tet	27.26		
1flr	27.07	1ind	25.71	1ucb	27.8		
1for	29.61	1jel	28.61	1vfa	27.68		
1fpt	28.61	1jhl	29.35	1vge	30.16		
1frg	29.09	1kem	27.48	1yuh	26.51		
1fvc	28.76	1mam	27.83	2cgr	28.74		
1fvd	28.66	1mcp	28.99	2fb4	29.35		

Table .2 Continued

MOULDER	Mean (%)	YangZhang	Mean (%)	YangZhang	Mean (%)	YangZhang	Mean (%)
1bbh	16.97	1abv	20	1g1cA	9.3	1orgA	17.2
1c2r	18.57	1af7	12.618	1gjxA	10	1pgx	8.1
1cau	14.47	1ah9	5.4	1gnuA	11.8	1r69	10.8
1cew	17.16	1aoy	8.9	1gpt	13.7	1sfp	12.5
1cid	16	1b4bA	8.9	1gyvA	10.1	1shfA	8.4
1dxt	18.53	1b72A	7.4	1hbkA	12.5	1sro	5.3
1eaf	20.34	1bm8	10.1	1itpA	10.3	1ten	11.9
1gky	17.79	1bq9A	8.4	1jnuA	12.6	1tfi	2
1lga	32.16	1cewI	13.9	1kjs	10.3	1thx	21.2
1mdc	11.63	1cqkA	11.1	1kviA	8.6	1tif	9.9
1mup	20.92	1csp	6	1mkyA3	9.6	1tig	9.9
1onc	15.26	1cy5A	19.5	1mla_2	13.9	1vcc	14.6
2afn	24.47	1dcjA	7.8	1mn8A	8.5	256hA	11.3
2cmd	29.66	1di2A	9.1	1nOuA4	16.2	2a0b	18.1
2fbj	19.54	1dtjA	8	1ne3A	3.8	2cr7A	14.2
2mta	8.57	1egxA	12.5	1no5A	13.4	2f3nA	8.4
2pna	16.96	1fadA	16.4	1npsA	17	2pcy	13.7
2sim	36.62	1fo5A	8.2	1o2fB	12.3	2reb_2	9
4sbv	27.51			1of9A	8.4		
8i1b	32.33			1ogwA	10.3		

The cluster energy was calculated for residue pairs as described in 2.1.2. and posted in Figure 20. The ideal cluster energy should use the block pairs in equation 11 for each specific distance and each specific orientation. However, since the dataset is not large enough, the dense block pair number is too small at the specific distance and the specific orientation to represent the dense area. Thus, we use the residue pairs to replace the block pairs within a range.

2. 2 Results and Discussions

2.2.1 The Distance Energy Adjusted by the Orientation Energy

We have developed a statistical energy function that is based on the characterization of the distance and orientation for each pair of key blocks. The main terms in the energy function include a distance term $E_{dist}^{i,j}(r)$ and an orientation term $E_{ort}^{i,j}(r,\theta,\phi)$ for each pair of key blocks i and j. The orientation term reflects the energy fluctuation of those pairs with different orientations but at the same distance bin. We observed in this study, as many previous studies [180], that the distance energy is critical in distinguishing block conformations. Using the ideal gas as reference, we derived the distance energy (see Methods). Note that the distance energy is about zero at the range of 15Å-20 Å, since the average density at this range was used as the reference (Figure 18 A and C). As an example, the lowest energy for block pair (16,16), of residue pair (PHE,PHE), is at the distance bin of 5.0-5.5Å with distance energy of $E_{dist}^{16,16}(5.0) = -2.02K_BT$ kcal/mol. This lowest energy distance of about 5Å agrees well with that derived from the Multiwell function, in which the geometrical center of the side chain atoms was used (Figure 18 A and B). The lowest distance energy for block pair (18,14), $E_{dist}^{18,14}(5.5) = -0.53K_BT$ kcal/mol, is at a slightly longer distance of 5.5-6Å, due to the

larger block 18 of TRP. The distance energy suggests that block 16, the distal end of the PHE, is more likely to interact with block 16 at the distance bin of 5.0-5.5Å (Figure 18 A) than for block 18 of TRP to interact with block 14 of ASN (Figure 18 C). This reflects the popular hydrophobic interaction between PHE and PHE in native proteins.

The orientation energy is an effective term to recognize the short-distance feasible geometry for a pair. To illustrate the nature of the preferred orientations at each distance bin, we show the orientation energy $E_{ort}^{i,j}(r,\theta,\phi)$ for two pairs of blocks (16,16) (Figure 19 A, B, C) and (18,14) (Figure 19 D, E, F) at three consecutive distance bins. As expected, most of the orientation bins have positive energy for block pair (16,16) to be at the short distance of 3.5-4.0Å (Figure 19 A). In our energy function, there is no need to introduce the repulsion term as in OPUS-PSP [11], since the orientations causing collision have zero or extremely low occurrences (Figure 19 A). The orientation energy was assigned to an extremely high value 9 if there were no observed cases in the orientation bin. Note that the distance energy of (16,16) is high and suggests it is unfavorable to have the pair in such short distance range compared to other distances (red bar in Figure 18 A and Figure 19 A). However, the orientation energy shows that if the pair is in such a distance bin, the orientations must be mostly restricted to three distinct peaks, roughly at 90°, -30° (red peak value), 90°, 0° (blue peak value), and 30°, -180° (green peak value) (Figure 19 A). The resulting energy of the pair at the red peak value (Figure 19 A) is $E_{dist}^{16,16}(3.5) + E_{ort}^{16,16}(3.5,90,-30) = 0.47 - 2.99 = -2.52$, an overall favorable energy. The orientation energy makes it possible to recognize the feasible geometry that would have been missed if a distance-only function was used.

As the distance increases, before reaching the most favorite distance at about 5Å, the distance energy becomes lower with $E_{dist}^{16,16}(4.5) = -1.63$ (green bar in Figure 18 A and Figure 19 C). There are more peaks in the orientation energy, but the height of the peaks decreases. For example, there are ten orientation peaks with orientation energy lower than -1.0 in the distance bin of 4.5Å-5Å (Figure 19 C), but there are only six in the bin of 3.5-4Å (Figure 19 A). The highest orientation peak in Figure 14 C has an overall energy $E_{dist}^{16,16}(4.5) + E_{ort}^{16,16}(4.5,90,-30) = -1.63 - 1.77 = -3.4$. The overall energy suggests that it is more popular for the pair (16,16) to adopt a relative geometry of $(4.5,90,-30)$ than $(3.5,90,-30)$. This is reasonable since there are more observed cases in the bin of 4.5Å-5Å than in the bin of 3.5-4 Å (Figure 18 A).

Note that there is no need to introduce weight parameters for the two terms in our energy function, because the orientation energy was characterized for those pairs at the same distance. but with different orientations. The two terms are not derived independently. With both the distance energy term and the orientation energy term, there is no need to use a repulsion term as in OPUS-PSP [182], since the statistically derived distance term naturally shows the repulsion at the short distance. This simplifies the calculation of the energy since there is no need to scan all the atoms for repulsions.

2.2.2 Transition of the Most Preferred Orientations at Different Distances for a Pair of Blocks

Numerous existing studies have suggested that each pair of amino acid side chains have preferred geometrical positions [191; 204; 205]. Lu, *et al.* characterized the preferred orientations for each pair of blocks on the side chains [182]. The nature of the multiple

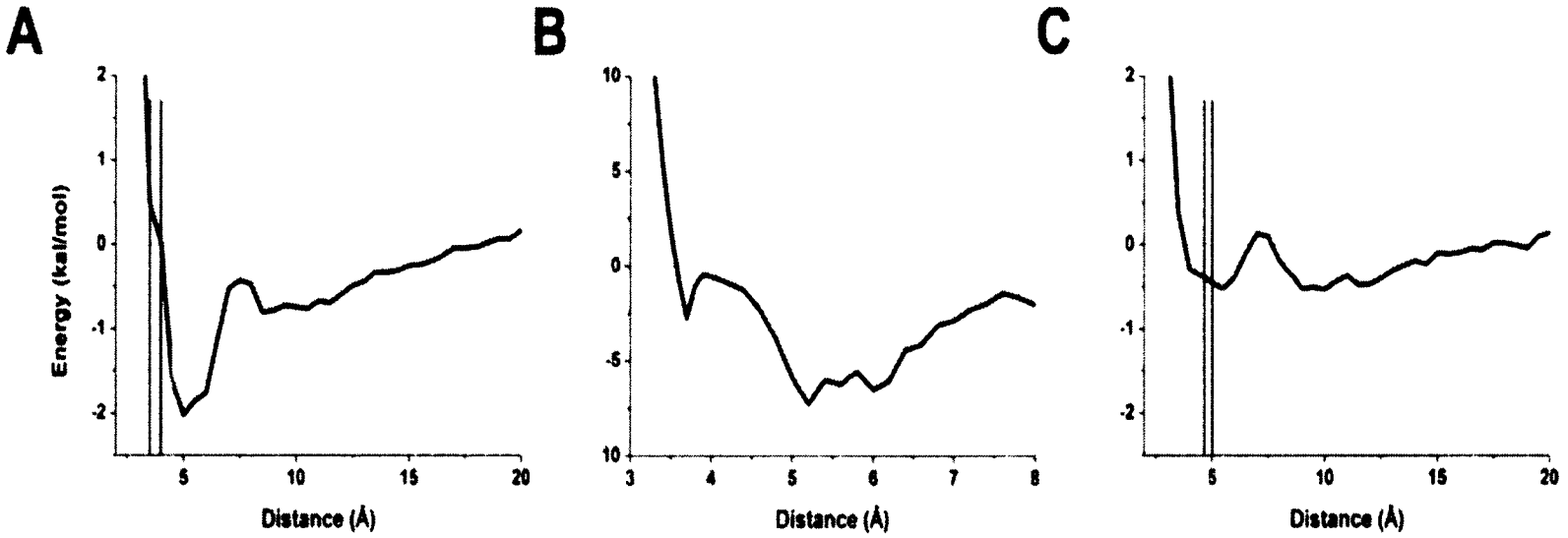


Figure 18. Examples of the distance energy [2]. The distribution of the distance energy is shown for block pair (16,16) of residue (PHE,PHE) in (A), and (18,14) of (TRP, ASN) in (C). (B): the distribution of the distance energy of residue pair (PHE,PHE) using the side chain geometry center representation of Multiwell function in. The red, blue and green lines are the sample distances corresponding to the colored frames in Figure 14.

(16,16)

A $E_{dist}(3.5) = 0.47$

ϕ/θ	0	30	60	90	120	150
-180			-2.03	-1.29		
-150						
-120						
-90						
-60						
-30			-1.93	-2.99		
0			-0.93	-2.69		
30						
60						
90						
120						
150			-0.28	-0.59		

B $E_{dist}(4.0) = 0.01$

ϕ/θ	0	30	60	90	120	150
-180			-0.87	-1.05		
-150		-0.05	-0.90	-0.46		
-120						
-90						
-60			-0.46	-0.72		
-30			-1.15	-2.55	-0.30	
0		-0.38	-1.32	-1.96		
30			-0.30	-0.66		
60						
90		-0.34				
120			-0.34	-0.49		
150		-0.59	-0.94	-1.32		

C $E_{dist}(4.5) = -1.63$

ϕ/θ	0	30	60	90	120	150
-180			-1.14	-0.62		
-150			-1.04			
-120		-0.78		-0.16		
-90		-0.46	-0.83	-0.94		
-60						
-30			-1.43	-1.32		
0	-0.05			-1.69		
30		-1.10				
60						
90		-1.34				
120						
150			-1.06	-1.77		

(18,14)

D $E_{dist}(4.5) = -0.37$

ϕ/θ	0	30	60	90	120	150
-180		-0.09		-0.74		
-150		-0.24			-2.52	-0.55
-120	-0.43	-0.09			-1.63	
-90						
-60						
-30						
0			-0.78		-0.90	
30		-0.09				
60	-0.09					
90		-0.08			-0.69	-0.24
120			-0.09		-0.74	
150		-0.43		-0.24	-0.65	

E $E_{dist}(5.0) = -0.45$

ϕ/θ	0	30	60	90	120	150
-180		-0.71	-1.64	-1.34		
-150		-0.71	-0.98	-1.60	-0.26	
-120		-0.53			-0.57	-0.71
-90						
-60						
-30						
0		-0.02	-1.32	-0.87		
30			-0.61			
60						
90						
120		-0.02	-1.47	-0.71	-0.26	
150			-0.53	-0.84	-0.26	

F $E_{dist}(5.5) = -0.53$

ϕ/θ	0	30	60	90	120	150
-180		-0.59	-0.37	-1.21		
-150			-1.54	-1.72		
-120			-0.17	1.79		
-90						
-60						
-30						
0			-1.10	-0.62	-0.95	
30			-1.12	-0.80		
60			-0.17	-0.57		
90			-0.65	-0.87	-0.17	
120				-0.03		
150	-0.26	-0.30	-0.44	-0.85		

Figure 19. The distribution of orientation energy for block pair (16,16) of (PHE,PHE) and (18,14) of (TRP,ASN) [2]. The orientation energy $E_{ort}^{(16,16)}(r,\theta,\phi)$ is shown for distance bin $r = 3.5$ (A), 4.0(B), 4.5(C) respectively. The distance energy of the bin is given on top of each frame. The lowest three energy orientations in each bin are highlighted in red, blue, and green respectively. Similar information is shown for $E_{ort}^{(18,14)}(r,\theta,\phi)$ in distance bin $r = 4.5$ (D), 5.0(E), and 5.5(F) respectively.

preferred orientations is also shown in our data. For example, block pair (16,16) has roughly three distinct preferred orientations at $(\theta, \phi) = (90°, -30°)$, (30°, -180°), and (90°, 150°) (Figure 19 A, B, C). These three preferred orientations appear to agree with the previous finding in Lu, *et al.* [182], in spite of different definitions of the orientation. However, OPUS-PSP energy function in Lu, *et al.* is not a distance dependent function. With the fine distance bins increments of 0.5Å, we are able to see the same preferred orientations are preferred at different levels when they are at different distance bins. In particular, two of the three preferred orientations, 90°, -30° (red peak in Figure 19 A) and 30°, -180° (green peak in Figure 19 A) are the two most distinct peaks in the distance bin of 3.5 Å. However, in the distance bin of 4.5 Å, the third preferred peak (90°, 150°) becomes the most favored (red peak in Figure 19 C) in this bin. Our energy function will distinguish two pairs with the same orientation, but at 1 Å difference in distance. For example, the same orientation has different energy of $E^{16,16}(3.5,90,-30) = -2.52$ verses $E^{16,16}(4.5,90,-30) = -2.95$ depending on the distance. We observed such transitions of the preferred orientation peaks in different distance bins for many other pairs of blocks. Figure 19 D, E, F shows another such example for block 18 of TRP and Block 14 of ASN. Our energy function is both distance- and orientation-dependent, and can distinguish the level of preferences for the orientations at different distances.

2.2.3 Performance in Five Decoy Sets

We downloaded five decoy sets: DecoysRus [206], MOULDER [207], hg_structal, ig_structal (http://dd.compbio.washington.edu/) and ITASSER (http://zhanglab.ccmb.med.umich.edu/decoys/). Each decoy set consists of a number of proteins. For each protein, a number of decoys and the true structure were provided in the set. The

decoys were ranked based on the energy. Four energy functions DOKB, Multiwell [180], OPUS-PSP [182], and DFIRE-2.0 [176], were compared. OPUS-PSP and DFIRE-2.0 were downloaded from their websites respectively. The Multiwell function was previously developed in our group [180].

An energy function's capability in recognizing native conformations was evaluated using three criteria:

1) The number of the native conformations that are ranked as the top 1 on the list (Table 3 column 2)

2) The mean rank of the native conformations in a decoy set (Table 3 column 3)

3) The number of proteins for which the native conformation is ranked closer to the top when the energy function is compared with DOKB (Table 4).

As an example, both DOKB and OPUS-PSP were able to rank the native structure as the top 1 by the potential energy for 31 of 34 proteins in DecoysRus (Table 3). Both functions failed to rank the native structure as the 1st for three proteins. However, the mean rank of the native structure among the 34 proteins is 26.3 for DOKB and 37 for OPUS-PSP. Table 10 lists the detailed ranking information for DecoysRus. DOKB ranks 70th for 1fc2 in fisa, which is much lower than 312th from OPUS-PSP. In lmds, the ranks of 1bba and 1fc2 of DOKB are also slightly lower than the corresponding results from OPUS-PSP. The mean rank in Table 3 reflects the ranks of the native conformations when they are not ranked as the top 1. DOKB ranked the native conformations closer to the top than OPUS-PSP for 3 proteins in the DecoyRus set (Table 4, row 2, column 2). It appears that DOKB performs slightly better than OPUS-PSP for two decoy sets (hg_structal, and I-TASSER), the same for two sets (DecoysRus, and MOULDER), and

significantly better in ig_structal set, in terms of ranking the natives as the top 1 (Table 3). When criterion (2) and (3) are used, DOKB outperforms OPUS-PSP in all the five decoy sets (Table 3 column 3, Table 4 column 2). Both DOKB and OPUS-PSP use blocks to represent side chains, and both are orientation dependent [182]. The results in the five decoy sets suggest that DOKB is more sensitive distinguishing the fine packing details than OPUS-PSP. It is possible that having both distance and orientation dependency contributed to the improved sensitivity even when less number of blocks were included in the calculation of the energy.

Both DOKB and DFIRE-2.0 ranked the same number of native conformations as the top 1 in three of the five decoy sets (Table 3). DOKB performed better in the other two decoy sets, particularly in the ig_structal set, in which DFIRE-2.0 failed to rank any native conformations as the top 1. When all the three criteria are considered (Table 3, Table 3 column 3), DOKB ranks the native conformations slightly better than DFIRE-2.0 in four of the five decoy sets, and significantly better in the ig_structal set. DFIRE-2.0 is an all-atom potential and DOKB is a coarse-grained potential, in which each amino acid is represented as a point except for TYR and ILE. The test using the five decoys suggests that it is possible for a coarse-grained potential to be comparable or even to outperform an all-atom potential in terms of recognizing native conformations. The Multiwell potential function is a pair-specific, distance-dependent function. A side chain is represented by the geometrical center of the side chain atoms in Multiwell [180]. The comparison between DOKB and Multiwell shows that DOKB has an overall better performance of recognizing native structures, particularly in DecoysRus and I-TASSER decoy sets. This is not surprising since DOKB is both distance and orientation dependent

and appears to distinguish fine conformation details. However, Multiwell performs the best among the four functions in hg_structal decoy set. It is possible that the distance-only energy functions may perform just as well or even better in some cases, since distance is the most important character to represent a pair of blocks or a pair of residues.

2.2.4 Performance in CASP8 Decoys

DOKB was tested using a dataset containing thirty CASP8 targets. The targets were downloaded from http://www.predictioncenter.org/download_area/CASP8/predictions_trimmed_to_domains/. Only those target proteins whose majority decoys have sequence length similar to that of the native structure were included in the dataset for convenience of testing. CASP8_30 dataset contains all the decoys of the 30 targets, and CASP8_30_r contains those decoys with less than 10Å backbone RMSD from the native. Seven energy functions were evaluated using the CASP8 targets: DOKB, OPUS-PSP [182], DFIRE-2.0 [176], Multiwell [180], Four-body [187], General-four-body [208] and Short-range [179]. The Four-body, General-four-body and Short-range potentials are available at http://gor.bb.iastate.edu/potential/. The energy functions were primarily evaluated on two metrics. One is the capability to recognize the native conformation among the decoys. The other is the backbone RMSD of the top-ranked conformation sorted by the energy.

Table 5 summarizes the results of the seven energy functions using thirty CASP8 targets. DFIRE-2.0, DOKB, and Multiwell appear to perform the best and have comparable capability of recognizing the native structures. They were able to rank the native conformation as the top 1 for 22, 21, and 22 proteins with a mean rank of 9.4, 9.13, and 10.3 respectively. OPUS-PSP recognized 19 natives, slightly less than the previous

three methods. However, OPUS-PSP failed to rank the native among the top 100 for 8 proteins, and therefore has a large mean rank value. The test using CASP8 decoys suggest that DOKB is more sensitive in distinguishing the native conformation than OPUS-PSP. Since CASP_30_r contains those decoys with less than 10Å RMSD from the native, and our results suggest that DOKB is fairly sensitive in recognizing the native among the conformations that are not quite wrong. When those decoys with large RMSD from the native are incorporated in the test, all the four functions perform slightly worse, except DFIRE-2.0 (Table 6). This suggests that DFIRE-2.0 is more robust in handling very wrong conformations and those near native conformations.

In addition to the capability of recognizing the native conformations, we evaluated the capability to distinguish near-native conformations using the backbone RMSD of the top-ranked conformation when the native is not included in the decoy set. DFIRE-2.0 appears to have overall the smallest RMSD (4.0Å), followed by the Four-body potential (4.19Å), DOKB (4.22Å) and OPUS-PSP (4.23Å) (Table 4). Short-range potential appears to perform worse than Four-body and General-four-body, similarly reported in a previous paper [196], in spite of the difference in the testing data sets. Although not tested in this paper, the optimized Four-body potential has been shown to perform better than Four-body or General-Four-body functions [196]. It combines the three functions (Four-body, General-four-body, and Short-range) and optimizes the combination. Table 9 summarizes the average RMSD of the top-ranked decoy. RMSD of DOKB is only slightly higher than DEFIRE-2.0 and 4B G POT and lower than other four functions.

Table 3. The performance of four potentials on five decoy sets [2]

Energy Function	Top 1[a]/Total No[b]	Mean[c]
DecoysRus		
DOKB	31/34	26.3
Multiwell	17/34	32.6
OPUS-PSP	31/34	37
DFIRE 2.0	28/34	46.4
MOULDER		
DOKB	19/20	1.4
Multiwell	19/20	2.9
OPUS-PSP	19/20	4
DFIRE 2.0	19/20	6.6
hg_structal		
DOKB	19/29	4.5
Multiwell	24/29	2.4
OPUS-PSP	18/29	6.8
DFIRE 2.0	19/29	7.2
ig_structal		
DOKB	15/61[d]-35/61[e]	21.2[f]-6.3[g]
Multiwell	22/61	8.9
OPUS-PSP	20/61	15.7
DFIRE 2.0	0/61	47.5
I-TASSER		
DOKB	53/56	12.6
Multiwell	16/56	94.4
OPUS-PSP	45/56	30.6
DFIRE 2.0	53/56	2.2

[a] The number of native structures that were ranked 1st by the energy.
[b] The total number of proteins in the decoy set.
[c] The average rank of the native structures in the decoys set.
[d] The number of the native structures that were ranked 1st without $E_{cluster}$.
[e] The number of the native structures that were ranked 1st with $E_{cluster}$.
[f] The average rank of the native structures without $E_{cluster}$.
[g] The average rank of the native structures with $E^{cluster}$.

Table 4. The number of proteins with better/same/worse rank for the native conformations [2]

Decoys	DOKB vs OPUS-PSP	DOKB vs DFIRE 2.0
DecoyRus	3[a]/31[b]/0[c]	6/28/0
MOULDER	1/19/0	1/19/0
hg_structal	9/20/0	15/12/2
ig_structal	34/17/10	59/0/2
I-TASSER	10/45/1	2/52/2
CASP8	10/16/4	5/18/7

[a] The number of proteins for which DOKB ranks the native closer to the top than the other potential.
[b] The number of proteins for which the native was ranked the same between DOKB and the other potential.
[c] The number of proteins for which DOKB ranks the native farther from the top than the other potential.

Table 5. The performance of seven potentials for CASP8_30_r decoys

Target	Length	#Decoys	4B POT[a]	4B G POT[b]	SR[c]	Multi-well		OPUS-PSP		DFIRE 2.0		DOKB	
			Rank	Rank	Rank	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD
T0388	164	213	42	91	73	2	5.29	125	3.44	1	2.9	1	2.9
T0389	134	376	48	5	94	1	3.47	357	3.47	1	3.15	1	3.53
T0392	82	336	275	287	25	93	1.31	7	1.65	7	1.58	1	1.57
T0395	235	16	1	2	6	1	8.25	1	8.25	1	8.25	1	8.25
T0396	102	374	241	37	85	1	1.92	1	2.45	6	2.41	1	1.72
T0397	82	9	1	2	1	1	9.72	1	9.72	1	11.9	1	9.72
T0401	127	361	70	17	1	1	4.4	1	4.28	1	4.29	1	4.64
T0406	147	279	52	110	7	1	7.45	273	3.06	1	3.06	1	3.51
T0407	231	210	2	3	58	1	4.41	185	4.26	1	4.26	5	5.7
T0411	118	393	159	58	3	1	6.21	1	3.2	1	4.11	1	3.7
T0412	165	327	1	244	172	1	3.31	1	3.35	1	3.29	1	5.69
T0414	127	82	22	30	28	1	9.8	1	9.8	1	9.8	1	9.8
T0415	107	289	136	36	14	1	2.81	1	2.78	2	2.61	1	2.61
T0421	221	98	16	26	4	1	4.51	1	8.86	1	4.51	2	4.51
T0425	179	336	48	249	219	5	3.72	1	3.24	1	3.15	1	3.49
T0426	257	295	85	163	116	1	0.55	255	0.8	5	0.94	18	0.9
T0427	218	362	23	25	4	1	3.72	1	3.07	1	3.07	1	3.37
T0428	229	334	1	163	201	1	0.87	324	1.19	13	1.3	4	0.87
T0430	138	53	16	15	3	1	7.42	1	9.05	1	9.05	1	6.96
T0432	130	276	78	131	128	11	3.38	258	7.72	1	1.8	1	3.38
T0433	199	256	1	1	42	1	3.83	1	3.65	1	2.05	1	3.89
T0436	405	226	1	2	71	1	8.7	1	6.39	1	6.39	6	8.7
T0448	207	278	1	2	43	1	5.04	1	3.43	1	4.13	1	4.35
T0449	296	307	1	16	23	1	4.87	1	4.87	1	4.87	1	4.87

Table 5. Continued

Target	Length	#Decoys	4B POT[a]	4B G POT[b]	SR[c]	Multi-well		OPUS-PSP		DFIRE 2.0		DOKB	
			Rank	Rank	Rank	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD
T0451	127	378	42	74	292	7	2.84	1	3.24	1	4.26	1	4.26
T0453	86	325	297	291	143	1	2.16	2	2.07	1	2.07	5	2.38
T0456	87	324	179	305	112	104	2.69	315	2.76	221	2.76	92	3.05
T0457	194	316	7	97	8	1	5.02	1	4.5	1	4.16	108	5.92
T0458	77	345	165	226	46	59	1.77	21	0.89	5	1.12	13	0.79
T0459	91	295	39	190	166	5	1.52	1	1.5	1	2.62	1	1.44
Avg[d]			68.3	96.6	72.9	10.3	4.37	71.4	4.23	9.4	4	9.1	4.22
Total[e]			8/30	1/30	2/30	22/30		19/30		23/30		21/30	

a The Four-body potential of the web server.
b The Four-body general potential of the web server.
c The short-range potential of the web server.
d The average rank of the native structures in the decoys set.
e The number of the native structures that were ranked 1st by the energy.

Table 6. The rank of the native conformation of CASP8_30 decoys [2]

Target	Length	#Decoys	Mutli_well	OPUS-PSP	DFIRE 2.0	DOKB
T0388-D1	164	235	2	126	1	2
T0389-D1	134	440	1	386	1	1
T0392-D1	82	359	96	7	7	1
T0395-D1	235	366	1	1	1	1
T0396-D1	102	436	1	1	6	1
T0397-D1	82	419	13	9	1	1
T0401-D1	127	475	3	1	1	1
T0406-D1	147	321	1	296	1	1
T0407-D1	231	320	1	266	1	6
T0411-D1	118	437	1	1	1	1
T0412-D1	165	357	1	1	1	1
T0414-D1	127	262	1	1	1	1
T0415-D1	107	409	1	1	2	1
T0421-D1	221	350	1	1	1	2
T0425-D1	179	413	6	1	1	2
T0426-D1	257	316	1	255	5	18
T0427-D1	218	415	1	1	1	1
T0428-D1	229	361	1	331	13	4
T0430-D1	138	270	1	1	1	1
T0432-D1	130	313	13	278	1	1
T0433-D1	199	282	1	1	1	3
T0436-D1	405	266	1	1	1	20
T0448-D1	207	292	1	1	1	1
T0449-D1	296	361	1	1	1	1
T0451-D1	127	422	10	1	1	1
T0453-D1	86	347	1	2	1	5
T0456-D1	87	344	104	319	221	92
T0457-D1	194	362	1	1	1	136
T0458-D1	77	369	61	21	5	13
T0459-D1	91	321	6	1	1	1
Avg_rank[a]			11.13333	77.13333	9.4	10.7
Total[b]			20/30	17/30	23/30	18/30

[d] The average rank of the native structures in the decoys set.

[e] The number of the native structures that were ranked the 1st by the energy.

2.2.5 Energy Difference at Highly packed Clusters for Residue Pairs at the Low-energy Region

Our energy function is derived using the block pairs from 4,180 protein structures regardless of where the block pairs are located. The energy difference contributed by different environments can be included, in principle, using higher order terms [134; 209; 210; 211; 212]. However, there are different local environments and it is a challenging problem to determine if and how much the local environments affect the energy. For example, some of the block pairs (14,14) of (ASN,ASN) may reside at the highly packed region, while others are located in the loosely packed region. Some of the pairs are at the regions in which the center residue is at a stable low energy environment, whereas others are in a less stable environment. This is possible because not all pairs are at a comfortable environment, although the protein is overall at a stable low energy environment. Since the highly packed clusters of a protein play significant roles in stabilizing the structure, it is important to represent the energy precisely at such clusters.

We investigated the distance energy for residue pairs at different environments. In particular, we collected the block pairs from the highly packed regions, the low-energy regions, the highly packed and low-energy regions, and all pairs regardless of the environments.

$P_{all}(m,n)$ (Figure 20 A) represents the probability for residue pair (m,n) to be in the low energy region regardless of where the pair is located. It is not surprising that (CYS,CYS) pair has the highest probability to be in a low-energy region, since many (CYS,CYS) form disulfide bond. Figure 20 A suggests that the probability for (ASP,LEU) to be in a low-energy region regardless of highly packed or loosely packed

environment is 0.09. $P_{cluster}(m,n)$ (Figure 20 B) represents the probability for(m,n) of a low-energy region to appear at the highly packed cluster. For example, the probability for (ASP,LEU) of a low-energy cluster when ASP has at least 15 neighbors is 0.045 (Figure 20 B), slightly less than that of $P_{all}(m,n)$. This suggests that it is more likely for (ASP,LEU) to be at the loosely packed environment when the energy at ASP is low. On the other hand, there is not much difference in the two probabilities for many pairs such as (ILE,ILE), (ARG,GLU). The two plots (Figure 20 A and 20 B) have similar colors for most of the pairs except some of those pairs with small polar or charged residues as the center, such as (ASP,LEU), (ASP,LYS), (ASN,GLN), (ASN,LYS) and (GLU,PHE). We derived $E_{cluster}^{m,n}$ (Figure 20 C) to represent the ratio between the two plots and used it to adjust the differences between a pair in a highly packed cluster or a loosely packed environment. Note that the energy function in (1) assigns the same energy for pair (m,n), as long as they have the same relative distance and orientation, regardless of the environment of m. However, Figure 20 suggests that it is less likely for (ASP,LEU) to have low-energy if ASP is in a highly packed cluster, even if ASP and LEU have the same relative geometry (as they do in the loosely packed environment).

2.2.6 Improved Ranking of the Native Structures for ig-structal with the Cluster Energy Term

DOKB performs well for four of the five datasets tested, except the ig-structal set when the cluster energy term was not used. This is a dataset of immunoglobulins, each of which contains a high percentage of β-sheets. We noticed that the proteins in this dataset were more densely packed than the other four data sets. On average, about 28% of the overall residues in the native proteins were highly packed in the ig-structal set, but only

11.2% were highly packed for DecoysRus set. Since the highly packed regions have the most constraints in packing the residues, it is likely to expect differences between the native and the decoy in the highly packed regions that are challenging to fold. We explored the use of the cluster energy term to adjust the slight environmental contribution at the highly packed clusters, as in formula (12). DOKB recognized the native conformation as the top 1 for 35 of the 61 proteins in the ig-structal set, a significant improvement from 15 when no cluster energy term is introduced. The mean rank of the native conformation also improved from 21.2, as in Table 1 to 5.8. Each protein decoy set in the ig-structal set contains a native structure and 60 high similar decoys. The backbones of these decoys are slightly different with the native structure, whereas the side-chains locate differently. Figure 22. shows the native structure of 1acy (red) and one decoy generated according to 1baf (blue). As marked with a rectangle, the side-chains of LYS 158 on two structures are pointing to the opposite directions, although their backbone are very close in space.

Table 7 shows the details of the top 10 decoys of 1dbb, based on the energy. It appears that both the native conformation (row 1 of Table 4) and the other decoys contain over 20% of highly packed clusters (column 2 and 3 of Table 7). In particular, there is a big difference in the number of highly packed clusters between the native (47 clusters) and other decoys (with 58-73 clusters). This might be reasonable since the native conformation is likely to be optimized to reduce the number of unnecessary clusters. As a result, there might be fewer highly packed clusters in the native than in the decoys. Without using the cluster term, native conformation is not

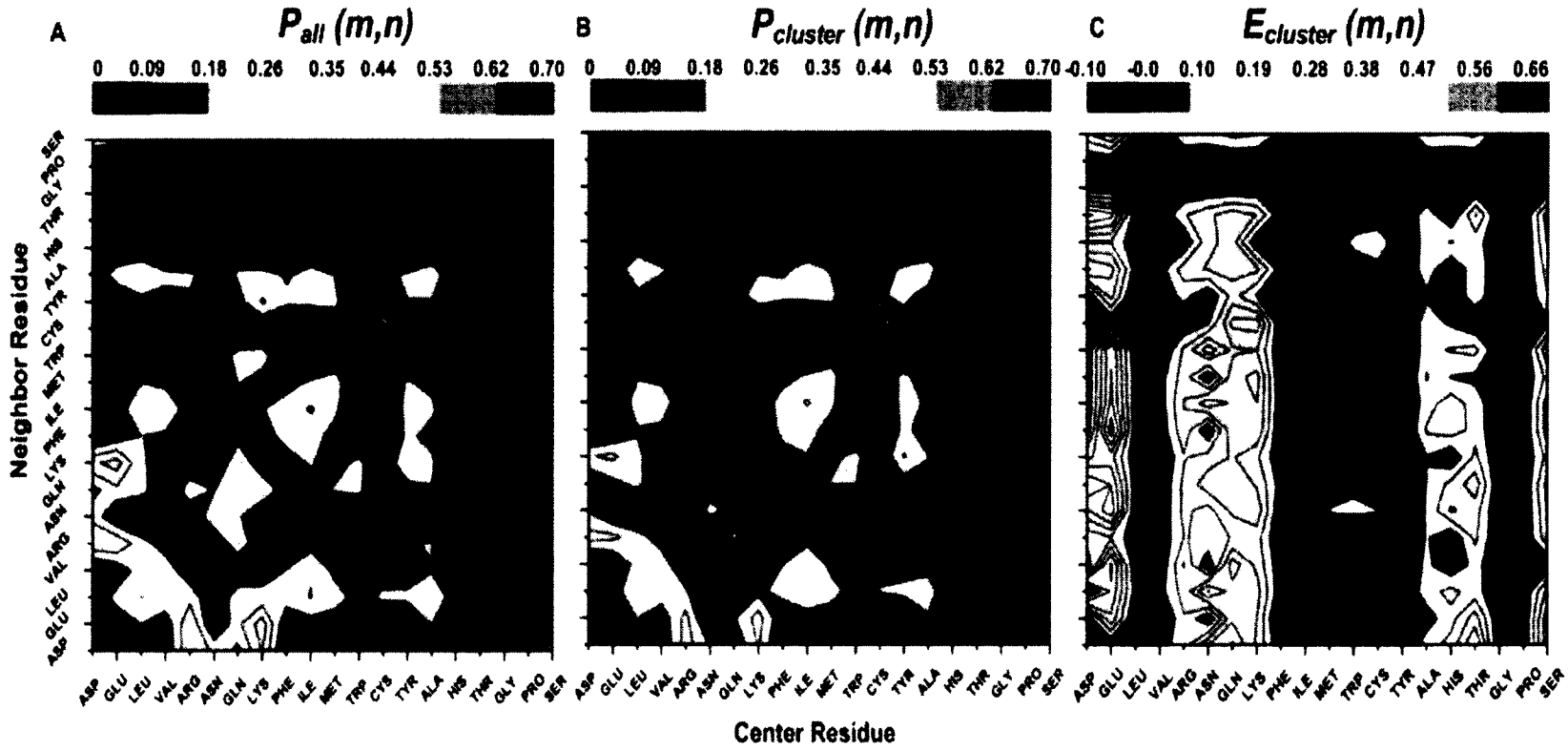


Figure 20. Probability difference for the low-energy residue pairs in highly packed clusters [2]. The probability of the low-energy pairs in the entire structure is shown in (A) and that in the highly packed cluster is shown in (B); an example of the difference in the probability between A and B is indicated by a circle; (C) $E_{cluster}$ plot for all pairs of amino acids.

Table 7. Improved recognition of the native conformation among the decoys of 1dbb in the ig-structal set [2]

Protein Decoy	# Cluster Centers (#Total Residue 231)[a]	Percentage of Highly Packed(%)[b]	RMSD[c]	Energy with $E_{cluster}$[d]	Energy without $E_{cluster}$[e]
0	47	20.3463	0	-5.407	-7.7
1	65	28.1385	1.64411	-2.777	-7.956
2	65	28.1385	2.17777	-3.517	-7.544
3	58	25.1082	1.75195	4.37	0.782
4	64	27.7056	1.94817	-3.394	-7.544
5	66	28.5714	1.89583	-3.082	-7.821
6	65	28.1385	2.40169	-3.761	-8.073
7	63	27.2727	2.58499	-3.085	-7.788
8	73	31.6017	1.9614	-1.941	-8.302
9	72	31.1688	2.13748	-1.294	-7.663
10	60	25.974	1.9794	-4.013	-8.069

[a] The number of cluster centers with more than 15 neighbors.

[b] The percentage of highly packed residues of the total number of residues.

[c] The RMSD between the decoy and the native for all atoms except hydrogen atoms.

[d] The Energy including E_{dist}, E_{ort} and $E_{cluster}$.

[e] The Energy including E_{dist} and E_{ort}.

distinguishable (column 6 Table 7) as top 1. However, the cluster energy term was able to adjust the overall energy in such a way that the native conformation is clearly distinguishable (column 5 Table 7, Figure 17 A). There might be two reasons for the effectiveness of the cluster energy term in the ig-structal set. Firstly, the cluster energy matrix (Figure 16 C) might be effective to down-weight the less likely pairs in the highly packed clusters. Secondly, the native conformation is more likely to be optimized so that unfavorable clusters are minimized. The big difference in the number of highly packed clusters between the native and the decoy might be a major reason for the improved

ranking. Table 7 shows the rank of the native conformation for each protein in the ig-structal. Apparently, the rank of the native structures improved for 41 of the 61 proteins (Table 8), and the native was ranked the 1st either using or not using the cluster term for 9 other proteins. For example, the native was ranked the 29th for 1hkl using energy formula (6), but was ranked 1st using formula (12) that includes the cluster energy term.

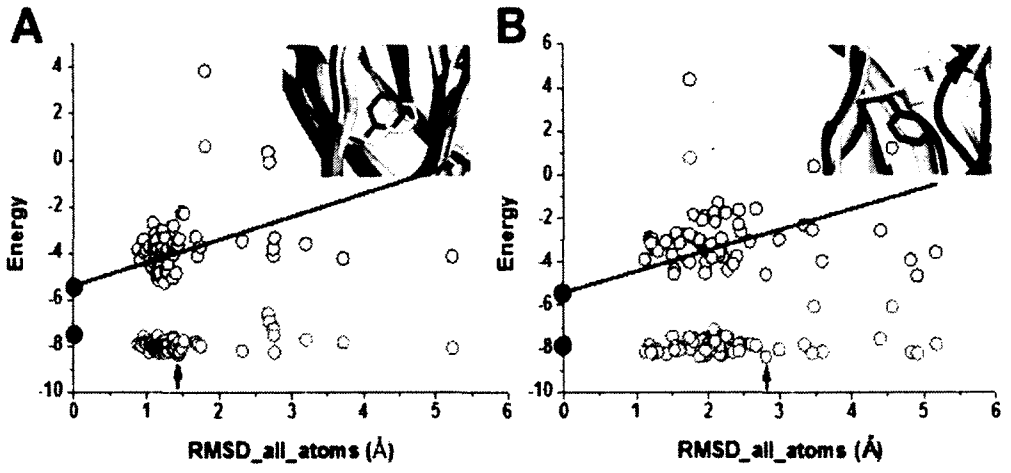


Figure 21. The plot of the energy for all decoys of 1dbb (A) and 1nsn (B) [2]. Black empty circle: the energy of a decoy when the cluster energy term is used; red empty circle: the energy of a decoy when no cluster energy is involved; the energy of the native structure is highlighted with a circle. Snapshots on the right corner represent one clustered part of the native structure and one of the decoys for 1dbb (A) and 1nsn (B) respectively.

2.3 CONCLUSIONS

The joint use of distance and orientation has proven to be an effective way to represent the geometrical relationships in many problems. We have developed a potential function that is both distance- and orientation-dependent, which is based on the coarse-grained model of key blocks. Our results illustrate that both distance and orientation are necessary to represent the fine details in geometrical relationships between the side chains in order to recognize the native conformations. Having only the distance or only the orientation representation may not be accurate enough.

Although both DOKB and OPUS-PSP use the block models, DOKB introduced the distance dependency and the cluster energy term to distinguish the highly packed environment. DOKB appears to be more sensitive in recognizing the native conformations than OPUS-PSP in all the six decoy sets, one of them involving CASP8 data. DOKB also shows comparable with DFIRE-2.0, an all-atom energy function in recognizing the native structures.

The local environment has been known to influence the pairwise energy, and there are various local environments. Highly packed clusters play important roles in stabilizing the structure. The densely packed nature of the highly packed clusters poses potential challenges in side chain packing. Our investigation into the highly packed clusters at the residue level suggests that certain residue pairs in a low-energy cluster have a lower probability to appear in the highly packed clusters than in the entire protein. We translated this finding into a cluster energy term and showed that it improves the native recognition in the ig_structal testing set.

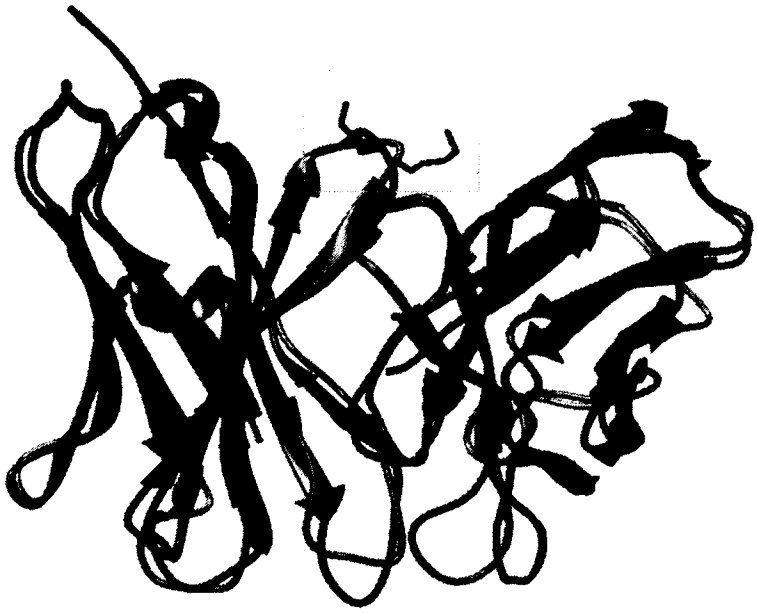


Figure 22. The native structure of 1acy (red) and a decoy (blue) in decoy set ig-structal.

Table 8. The rank of the native conformation with/without $E_{cluster}$ energy for ig_structal decoys [2]

Protein[a]	Rank With/Without $E_{cluster}$[b]	Protein	Rank With/Without $E_{cluster}$	Protein	Rank With/Without E_{cluter}
1acy	8/1	1hil	1/1	1ngq	12/2
1baf	½	1hkl	1/29	1nmb	3/23
1bbd	1/5	1iai	1/35	1nsn	1/57
1bbj	2/55	1ibg	1/4	1opg	9/45
1dbb	1/44	1igc	12/42	1plg	2/16
1dfb	7/20	1igf	1/16	1rmf	1/54
1dvf	½	1igi	6/30	1tet	1/7
1eap	1/1	1igm	21/54	1ucb	1/6
1fai	2/17	1ikf	1/1	1vfa	1/1
1fbi	3/53	1ind	5/3	1vge	1/1
1fgv	1/1	1jel	1/5	1yuh	38/56
1fig	6/58	1jhl	1/25	2cgr	1/15
1flr	15/48	1kem	1/2	2fb4	2/1
1for	1/53	1mam	2/8	2fbj	6/3
1fpt	1/17	1mcp	61/61	2gfb	1/1
1frg	6/1	1mfa	1/1	3hfl	2/35
1fvc	¼	1mlb	1/40	3hfm	61/61
1fvd	1/3	1mrd	3/51	6fab	½
1gaf	1/1	1nbv	12/54	7fab	5/1
1ggi	2/48	1ncb	1/11	8fab	3/1
1gig	3/1				

[a] The protein name in ig_structal decoy set.

[b] The rank of native structure in the decoys with/without $E_{cluster}$.

Table 9. The average RMSD of the top-ranked conformations for CASP8 decoys [2]

	CASP8_30_r Decoys							CASP8 Decoys[b]			
Potential	**4B POT[c]**	**4B G POT[d]**	**SR[e]**	**Multi-well**	**OPUS-PSP**	**DFIRE 2.0**	**DOKB**	**4B POT[f]**	**4B G POT[g]**	**SR[h]**	**4BOPT POT[i]**
Average RMSD[a]	4.48	4.19	5.84	4.37	4.23	4.0	**4.22**	4.6	4.7	6.9	**3.7**

[a] The average RMSD of the top ranked decoy.
[b] The decoy set in [25].
[c] The four-body potential of the web server.
[d] The general-four-body potential of the web server.
[e] The short-range potential of the web server.
[f] The four-body potential results in [25].
[g] The results of general-four-body potential in [25].
[h] The results of the short-range potential in [25].
[i] The four-body optimized potential in [25].

Table 10. The rank of the native conformation in DecoysRus set [2]

Decoys	Multi_well	OPUS-PSP	DFIRE 2.0	4BOPTPOT[a]	DOKB
4state_reduced					
1ctf	1	1	1	2	1
1r69	1	1	1	2	1
1sn3	2	1	1	1	1
2cro	1	1	1	1	1
3icb	2	1	4	1	1
4pti	2	1	1	1	1
4rxn	2	1	1	1	1
fisa					
1fc2	180	312	102	496	70
1hdd-C	12	1	1	3	1
2cro	2	1	1	62	1
4icb	1	1	1	1	1
fisa_casp3					
1bg8-A	1	1	1	1	1
1bl0	1	1	1	3	1
1eh2	2	1	2	-	1
1jwe	1	1	1	1	1
smd3	1	1	1	-	1
lattice_ssfit					
1beo	1	1	1	1	1
1ctf	1	1	1	1	1
1dkt-A	1	1	1	1	1
1fca	37	1	1	1	1
1nkl	1	1	1	1	1
1pgb	1	1	1	1	1
1trl-A	1	1	1	1	1
4icb	1	1	1	1	1
lmds					
1b0n-B	16	1	441	441	1
1bba	497	501	501	470	437
1ctf	1	1	1	501	1
1dtk	44	1	1	70	1
1fc2	395	409	501	99	357
1igd	1	1	1	3	1
1shf-A	20	1	1	1	1
2cro	2	1	1	5	1
2ovo	2	1	1	119	1
4pti	24	1	1	157	1
Avg_rank[b]	32.6	37	46.4	76.6	26.3
Total[c]	17	31	28	15	31

[a] The four-body optimal potential result from [25].

[b] The average rank of the native conformation in the decoy set.

[c] The number of the native conformations that were ranked the 1st by the energy.

CHAPTER 3

PROTEIN TOP-K TOPOLOGY PROBLEM

Cryo-electron microscopy (cryo-EM) is an important technique used to derive the three-dimensional structure of large protein complexes [213; 214; 215; 216; 217]. Using the current advances of the cryo-EM technique, it is possible to produce volumetric images, called density maps, of a protein in the high-resolution range, such as 3–5-Å resolution [105; 106]. At this resolution, the secondary structure is mostly distinguishable, and backbone tracing becomes possible [218]. Due to various experimental difficulties, many proteins have be resolved to the medium-resolution range (5–10 Å) [110], comprising about 22% of the density maps in EMDB. A number of computational methods have been developed to detect α-helices and β-sheets for these medium-resolution density maps [97; 136; 152; 154; 219; 220; 221]. The secondary structure elements identified by the detection tools in the density map (SSE-Ds) refer to the helix sticks and β-sticks detected from the three-dimensional image. Each detected helix is represented by the trace of the central axis of the helix; each detected β-sheet is represented by a curved surface that contains several β-strands. Each detected β-strand is also represented by the trace of the central axis of the β-strand. Although β-strands are often invisible in the medium-resolution image, recent studies have shown significant potential in β-strand detection. In our study of the topology search algorithm, we assumed that some of the β-strands in the β-sheet would be detectable. However, the various X-ray crystallographic modeling building tools, such as O [222] and Coot [223], are unable to directly use these SSE-D anchor points due to the lack of the connection relationship between SSE-Ds. Our research focused on identifying these

connections between SSE-Ds.

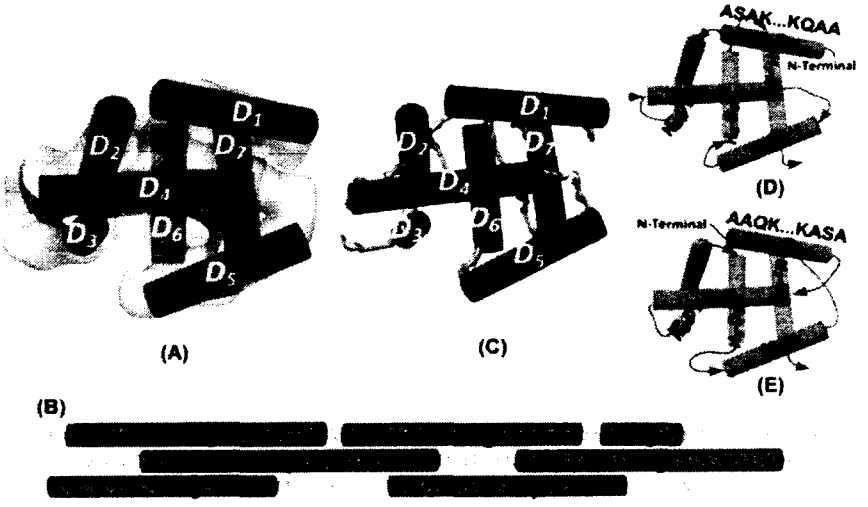


Figure 23. Helix sticks and the topologies [4]. (A) The density map (gray) was simulated to 10-Å resolution, using protein structure 1FLP (PDB ID) and EMAN software [6; 8]. The seven helix sticks were detected from the density map, using *SSETracer* [9] and viewed by Chimera [10]. (B) The helix segments in the protein sequence are marked as H_1 to H_7. (C) The helix sticks (red) were superimposed on the skeleton (green), generated using Gorgon [11; 12]. (D) The correct topology of the SSE-Ds (sticks). (E) A wrong topology.

The 1-d protein sequence is another source used to extract the secondary structures. The secondary structure elements in the sequence (SSE-Ss) refer to the α-

helices and β-strands in the 1-d protein sequence. A number of programs are available to predict SSE-Ss, including SSPro [224], JPred [225], PsiPred [226] and Porter [160]. These programs assign an SSE-S (α-helix, β-strand, loop) to each amino acid in the sequence. Due to prediction errors, a consensus alignment from multiple predictions can obtain better assignments than the results from a single program. The predictions from Porter can have over 80% accuracy.

The topology search problem is defined as determining the correspondence between the SSE-Ss and the SSE-Ds. Figure 23 demonstrates the topologies for the pure α-helix protein 1FLP. In Figure 23 A, 7 helix sticks (D_1, D_2, D_3, D_4, D_5, D_6, D_7) were detected from the simulated 10-Å resolution density map of the protein 1FLP, using SSETracer [9]. The real helix segments in the sequence are marked as H_1 to H_7 from the N-terminal to the C-terminal. The true topology is the correct assignment of SSE-Ss to SSE-Ds, in other words, the order of the SSE-Ds with respect to the SSE-Ss and the direction of each element. For example, the order of the SSE-Ds in the true topology is (D_1, D_2, D_3, D_4, D_5, D_6, D_7) [Figure 23 D]. A wrong topology [Figure 23 E] may contain a wrong order of the sticks, such as $(D_1, D_4, D_3, D_2, D_6, D_5, D_7)$, and a wrong direction for certain sticks, such as S_1 and S_3 in this case. The optimal match should consider factors such as (1) matching the SSE-Ss for α-helices to the SSE-Ds for α-helices and matching the SSE-Ss for β-strands to the SSE-Ds for β-strands, (2) matching the long SSE-Ds in the density map to the long SSE-Ss in the protein sequence, and (3) matching two SSE-Ss connected by a short loop in the protein sequence to two close SSE-Ds in the density map.

The goal of designing an effective topology search algorithm is to reduce the search space. The naïve de novo protein modeling approach builds the protein models for all topologies, then chooses the native model by the geometrical and physical constraints. However, it is an impossible task due to the large search space. For example, in Figure 23, the topology determination for the protein 1FLP is to assign 7 SSE-Ss to 7 SSE-Ds. Based on the fact that there are 7! different orders for the assignment and two directions to assign for each helix, the total topology number is $7!2^7 = 645120$. Building the models for so many topologies is time-consuming and may take years for the larger proteins. Since most of the topologies for a specific protein are invalid due to the geometrical constraints, it is possible to obtain a subset of all topologies with an effective search algorithm using basic geometrical constraints, in which the true topology is included.

Three approaches have been attempted to derive the topology of the SSEs. The naïve approach is to enumerate all possible topologies and to evaluate them one by one [161, 227]. Due to the huge search space, this approach is limited to the proteins fewer than 9 SSEs. Another approach is to use the Monte Carlo simulation to sample the search space [69, 132]. Although this approach can work with a large search space, the stochastic nature of the Monte Carlo approach may miss the native topology. The third approach is to translate the topology problem into a graph problem by exploiting the constraints from a pair of sticks. This approach is performed within Gorgon [11], in which the SSE correspondence results are shown as a ranked list from best to worst. It produces two graphs, one representing the connectivity among the SSE-Ds in the density map, and the other representing the linear relationship of the SSE-Ss [228, 229].

The topology search problem is then translated into an inexact graph-matching problem. The A* search was used in matching the two graphs. The complexity of the A* search depends on the heuristics used. However, this approach requires that the true link between the SSE-Ds be detected correctly. Due to the quality of the skeleton [green in Figure 23 C] generated by Gorgon, the true link may be missed. It is also unclear if the A* search is effective for large proteins with such a complex skeleton.

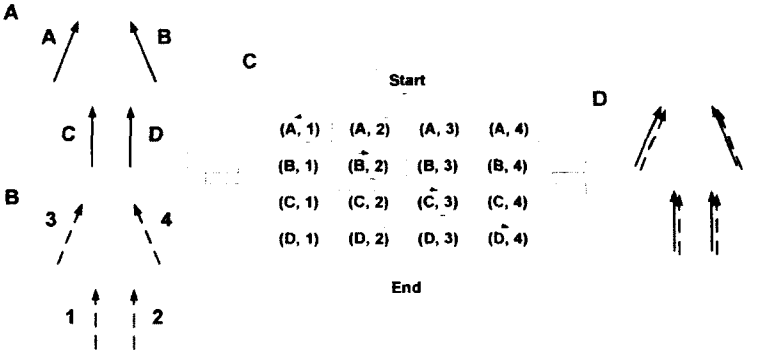


Figure 24. Application of interpretation tree in finding a match of model features to image features. (A) The model features of an object. (B) The features in the image. (C) The interpretation tree graph for the problem. (D) The match based on the best path in the graph.

Al Nasr et al. proposed a dynamic algorithm to search the top-k topologies for the pure α-helix proteins [164]. This algorithm borrowed the idea from the interpretation tree [230] to convert the topology search problem into a single graph. A common application of the interpretation tree is to identify an object from an image by mapping the model features [Figure 24 A] of the object to the image features [Figure 24 B]. The interpretation tree converts the matching problem into a graph [Figure 24 C]. There are two types of constraints. One is the unary constraint, which measures the matching between a model feature (A, B, C, or D) and an image feature (1, 2, 3, or 4). In the graph [Figure 24 C], the unary constraint represents the similarity between the model feature and the image feature for a node. For instance, A fits 1 exactly; any path passing through node (A, 1) will get an extra weight. The other type is the binary constraint, which represents the relationship between two nodes in the graph. In Figure 24 A, both A and B are on top. In Figure 24 B, both 1 and 2 are on top. The weight between node (A, 1) and (B, 2) is high. The best mapping has the maximum path in the graph from the start node to the end node [Figure 24 C]. Several algorithms are available to search the maximum path of the graph. The best paths [(A, 1), (B, 2), (C, 3), (D, 4)] are marked out in Figure 24 C. Figure 24 D represents the best mapping based on the best path. In Al Nasr's algorithm, the SSE-Ss and the SSE-Ds are taken as the model features and the image features in the interpretation tree, respectively. Let (H_1, H_2, ..., H_M) be SSE-Ss in the protein sequence and (S_1, S_2, ..., S_N) be SSE-Ds in the density map, in which $M \geq N$ without losing generality. All topologies can be represented with an M*2N graph. In this graph, each node is an assignment of SSE-S to SSE-D with the direction d, (H_i, S_i, d), in which d is the

direction of SSE-D. Since the protein sequence can enter the SSE-D from two sides, each SSE-S/SSE-D pair is represented by two nodes. When d is +1, the protein sequence enters SSE-D from one side; when d is -1, the protein sequence enters SSE-D from the other side. Regarding the unary constraint, for each node, if the lengths of H_i and S_i are significantly different, no edge is allowed to enter or exit this node. Concerning the binary constraint, for each node pair [(H_i, S_i, d), (H_j, S_j, d')], the edge weight is equal to the difference between the loop lengths in the sequence and the density map, respectively. The loop length in the sequence is the length of the loop between H_i and H_j. The loop length in the density map is the distance between the end point of S_i and the start point of S_j, which will be replaced with the skeleton trace length if a skeleton trace exists between S_i and S_j [Figure 23 C]. Figure 25 shows a graph in the top-k topologies search algorithm for a pure α-helix protein. Each node (Hi, Si, d) represents a valid SSE-S/SSE-D pair by the unary constraint, and each edge represents a valid SSE-S/SSE-D pair by the binary constraint. Two special nodes are added as the start and end nodes. The edge weights between the nodes and these two special nodes are zero. Due to the features of SSE-Ss and SSE-Ds, there are several potential constraints in the graph, as follows: (1) The protein sequence is linear; each edge points downward. If M = N, each edge must link consecutive rows. If M > N, each edge is allowed to link nonconsecutive rows, and the maximum allowed gap is M - N. (2) The two nodes linked by an edge must represent two different SSE-Ds; each SSE-D is not allowed to be assigned to an SSE-S twice. (3) A valid topology starts from the start node and ends at the end node, without passing the same SSE-D twice. The red dashed lines in Figure 25 represent an invalid topology

since it passes S_2 twice. The green solid lines represent a valid topology [(H_1, S_1, -1), (H_2, S_2, -1), (H_3, S_3, -1)]. The total weight for a topology is the sum of all edge weights in this topology. The best topology has the minimum weight among all the topologies. The topology search problem is converted into a shortest-path search problem. The algorithm reduces the complexity $O(N!2^N)$ to $O(N^2 2^N)$. The top-k topologies can be generated based on the best one.

Al Nasr's algorithm solves the top-k topologies search problem for the pure α-helix proteins. However, this algorithm cannot be applied to the protein containing β-sheets. The challenge in deriving the topology for the β-sheet is that the β-strands in the same β-sheet are fairly close, with about ~4.5-Å spacing [Figure 26 A]. The topology in Figure 26 B is the most common for the β-sheets of the known protein structures in the Protein Data Bank (PDB), for which the loop connects two adjacent nodes. However, with the binary constraints in Al Nasr's algorithm, the edge weight between these two nodes is greater than the weight between the two nodes representing two nonadjacent SSE-Ds. In other words, the true topology usually has the worst score. For instance, the topology in Figure 26 C has the best score; however, it is never observed in Dunbrack's database [203]. To solve this problem, several binary constraints [134] for β-strand nodes based on the statistical analysis have been added to the algorithm and has displayed the improvement for the proteins containing β-sheets.

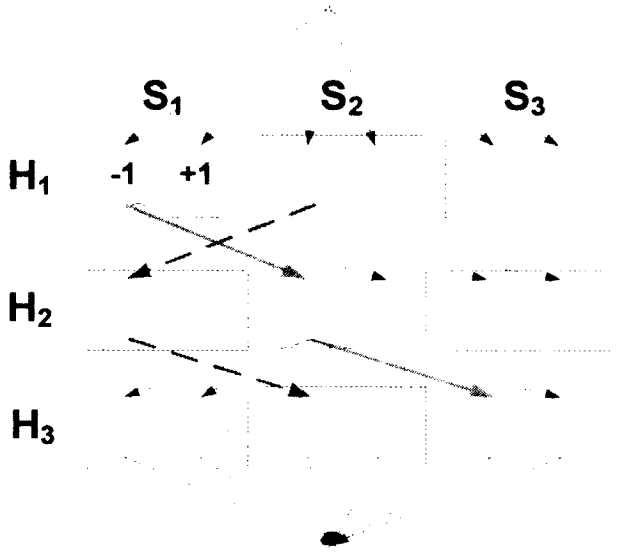


Figure 25. The graph of the pure α-helix proteins in the top-k topologies search algorithm. The red dashed lines represent an invalid topology; the green path is the true topology.

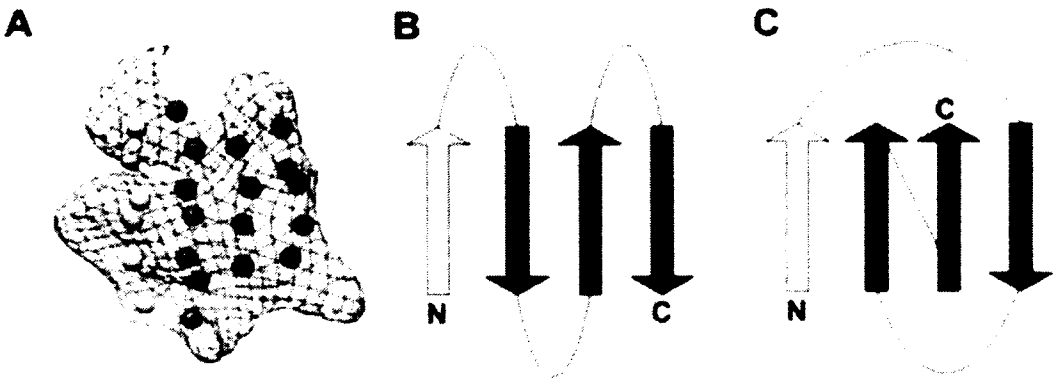


Figure 26. A 4-stranded β-sheet [1]. (A) Density map in β-sheet area; four sticks (SSE-Ds) with the different colors represent the four strands in the sheet. (B) General topology. (C) Rare topology.

The present work focuses on the top-k topologies search algorithm for β-sheets. More binary constraints for the β-sheet have been added to the algorithm instead of using only the basic length constraint for α-helices. We have translated the binary constraints for the β-sheet into the adjusted edge weight, using the probability information of β-sheet topologies. The topologies with low-occurrence probabilities have low probabilities to be the native topology and will be screened from the candidate topologies. The details of the algorithm are introduced in the method section. Several samples containing both α-helices and β-sheets have been used to evaluate the algorithm. The corresponding results are presented in the results section.

3.1 Method

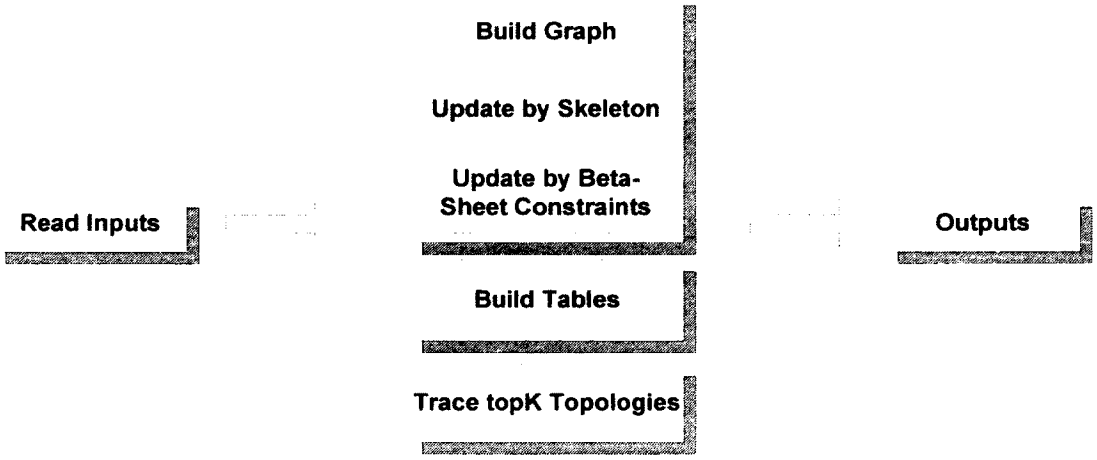


Figure 27. The flow chart of the top-k topology search program.

Figure 27 shows the flow chart of the top-k topology search program. Besides the input and output, the program includes the following five steps: (1) Build the graph, and calculate the edge weight between the valid node pairs by using the unary and binary constraints. (2) Update the edge weight if there are skeletons between two SSE-Ds. (3) Update the edge weight by the connection preference between two β-strands. (4) Build the node table that contains the information tracing the topology. (5) Trace the top-k topologies.

3.1.1 Inputs

The input information for the algorithm contains the SSE-Ss, SSE-Ds, and the skeleton predicted from the density maps. The SSE-Ss can be predicted from the amino acid sequence with about 80% accuracy. For test purposes, we have used the real SSE-Ss from the PDB file instead of the predicted SSE-Ss to avoid the intervention from the wrong prediction. Figure 28 A displays a sample of SSE-Ss'

A

H0(19, 24)

S0(28, 32)

S1(42, 47)

S2(50, 54)

H1(58, 75)

B

-6.684	-9.51933	-5.522	1
-6.758	-9.98967	-3.75367	1
-6.056	-9.71367	-2.48167	1
-6.208	-8.51867	-1.00133	1
-0.786	1.202	-9.97167	2
-1.89267	0.867	-8.50933	2
-2.59067	1.55533	-7.19467	2
-4.12033	2.32867	-6.90767	2
-5.579	1.87867	-6.42133	2
-6.589	1.69667	-5.00667	2
-7.837	2.17933	-3.92067	2
-9.34	1.85733	-3.66967	2
-10.0727	0.569	-2.68167	2
-10.597	0.360667	-1.00567	2
-11.963	0.462667	0.0956667	2
-13.0963	-0.573333	0.503	2
-12.6257	-2.73867	1.80367	2
-12.422	-4.13533	3.94333	2
-12.8477	-3.29567	5.75967	2
-12.879	-2.046	7.174	2

C

-2.138	-1.133	5.749	1	1
-0.46	1.105	3.21	1	1
2.968	2.235	4.332	1	1
4.668	5.114	2.584	1	1
8.397	4.798	2.163	1	1
8.007	3.978	-2.516	2	1
4.401	3.012	-1.902	2	1
4.574	-0.406	-0.337	2	1
1.526	-2.421	0.655	2	1
1.223	-4.037	4.048	2	1
-0.621	-7.332	4.004	2	1
3.834	-9.538	4.552	3	1
4.731	-6.281	2.822	3	1
5.317	-5.745	-0.893	3	1
7.192	-2.828	-2.456	3	1
5.419	-0.901	-5.189	3	1

Figure 28. The input information of 2KUM for the top-K topology search algorithm. (A) The sequence information. (B) The stick points of α SSE-Ds. (C) The stick points for β SSE-Ds.

input, in which H/S represents the helix/strand followed by an index. The start index and the end index of each SSE-S in the sequence are enclosed in parentheses. The

SSE-D input for the helices [Figure 28 B] contains the axis sticks of the helices. Each stick consists of many points on the axis. The first three columns in the SSE-D input for the helices list the coordinates of the points. The last column is the helix index. The SSE-D input for the β-strands [Figure 28 C] contains the axis sticks of the strands in the β-sheets. The first three columns present the coordinates of the strand axis. The fourth column lists the strand index in a specific β-sheet. The last column lists the index of each β-sheet. Each helix/strand is separated by an empty line. The SSE-D input files are from the outputs of SSETracer [9]. The skeleton of the density map has been generated with a skeleton detection tool developed by Al Nasr [139]. Although the skeletons from this tool have better quality than those from Gorgon, there are still many invalid traces. In other words, the skeletons from the intermediate resolution density maps cannot be used to trace the backbone directly. We have used the skeleton to obtain the more accurate edge weight between nodes.

3.1.2 Build the Graph

Let M_α and M_β be the number of helices and β-strands in the protein sequence, respectively. Let N_α and N_β be the number of helix sticks and β-sticks detected from the density map, respectively. Suppose that $M_\alpha \geq N_\alpha$, and $M_\beta \geq N_\beta$. The total number of possible matches between SSE-Ss and SSE-Ds is $\binom{M_\alpha}{N_\alpha} N_\alpha! \, 2^{N_\alpha} \binom{M_\beta}{N_\beta} N_\beta! \, 2^{N_\beta}$. Each possible match defines a possible topology. We have created the weighted directed graph $G_{Top} = (V, E, w)$ to represent the topology problem. Let the sequence segments of the secondary structure be (S_1, S_2, ..., S_M) and $M = M_\alpha + M_d$. Let the sePDcondary structure sticks detected from the density map be (D_1, D_2, ..., D_N) and $N = N_\alpha + N_\beta$. For convenience, we let $D_1, D_2, \ldots, D_{N_\alpha}$ be the helix sticks and $D_{N_\alpha+1}$, $D_{N_\alpha+2}, \ldots, D_{N_\alpha+N_\beta}$ be

the β-sticks. Let the set of columns C be {1, 2, ..., N}. Since a helix segment in the sequence will only be assigned to a helix stick and not a β-stick, V has $2M_\alpha N_\alpha + 2M_\beta N_\beta$ regular nodes and two special nodes, *START* and *END*. The indexes for the row and column of the nodes are i and j, respectively. The two ends of a stick are marked by $t = \pm 1$ to distinguish the two directions of each assignment. A node (i, j, t) represents an assignment of SSE-S_i to SSE-D_j in t direction. The G_{Top} graph is defined in equation (1). The graph for 2KUM is shown in Figure 30.

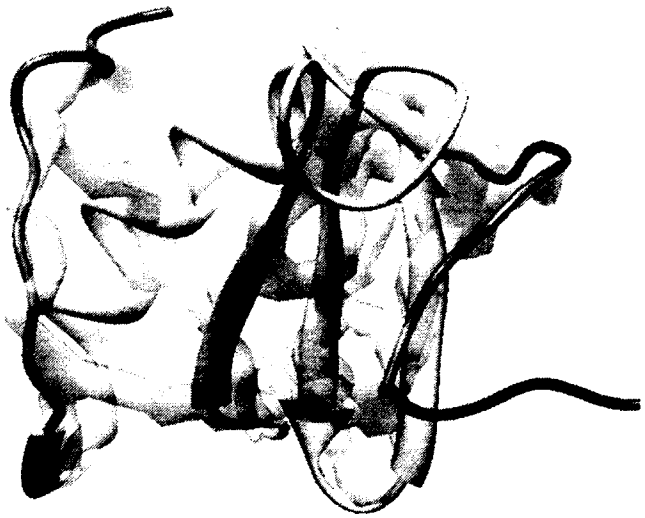


Figure 29. The protein 2KUM (colorful band) and the corresponding skeleton (gray).

$$V = \left\{ (i,j,t) \middle| \begin{matrix} 1 \le i \le M,\ t \in \{0,1\} \\ \begin{pmatrix} (1 \le j \le N_\alpha)\ AND\ (SS_i\ is\ a\ helix)) \\ OR((N_\alpha < j \le N) AND\ (SS_i is\ a\ \beta - strand) \end{pmatrix} \end{matrix} \right\} \cup \{START,\ END\}$$

$$E = \left\{ ((i,j,t),(i',j',t')) \middle| \begin{matrix} 1 \le i \le M-1, i < i', skip_\alpha(i,i') \le M_\alpha - N_\alpha, skip_\beta(i,i') \le M_\beta - N_\beta, \\ 1 \le j \ne j' \le N,\ t,\ t' \in \{0,1\} \end{matrix} \right\}$$

$$\cup \left\{ (START,(i,j,t)) \middle| skip_\alpha(0,i) \le M_\alpha - N_\alpha,\ skip_\beta(0,i) \le M_\beta - N_\beta, 1 \le j \le N,\ t \in \{0,1\} \right\} \quad (13)$$

$$\cup \{((i,j,t),\ END) \mid skip_\alpha(i,M+1) \le M_\alpha - N_\alpha,\ skip_\beta(i,\ M+1) \le M_\beta - N_\beta\ t \in \{\{0,1\}\}$$

Use the unary constraints to screen the invalid nodes. For each node, compare the lengths of SSE-S and SSE-D. If their length difference is over 60%, this node has a high possibility to be invalid and is removed from the graph. For a node, if L*LE*0.4 > LS [Figure 31 B] or L*LE < LS*0.4 [Figure 31 C], the node is invalid, in which L is the number of the amino acids of this SSE-S, LS is the length of SSE-D, and LE is the length of an amino acid in SSE, 1.5 for α-helix and 3.5 for β-strand. Figure 32 shows the graph after removing the invalid nodes, using the unary constraints.

Use the binary constraints to set up the weights for the edges. An edge from node (i,j,t) to (i',j',t') represents the assignment of $S_{i'}$ to $D_{j'}$ in direction t' right after the assignment of S_i to D_j in direction t. Since a protein sequence has its direction, all the edges in the graph point downward with $i' > i$. When $M = N, i' = i + 1$. When $M > N$, skipping edges exist. The maximum number of rows that an edge may skip should satisfy two rules, as follows: (1) The number of skipped helices [referred to as $skip_\alpha(i,i')$] is no more than $M_\alpha - N_\alpha$. (2) The number of skipped β-strands [referred to as $skip_\beta(i,i')$] is no more than $M_\beta - N_\beta$. Since each stick in the volume map can only be assigned to one sequence segment, there is no edge between the nodes in the same

column; similarly, there is no edge between the nodes in the same row. Special edges are drawn from the *START* node to each node on the top rows and are similarly drawn from each node on the bottom rows to the *END* node, as long as the skipping edges satisfy the above-mentioned two rules. The weight is zero for the special edges and nonnegative for others. Depending on the situation of the edge, three types of edge weights have been used, as follows: the ∞, the skeleton trace, and the penalized Euclidian distance. We have assigned ∞ as the edge weight to the two consecutive assignments that are impossible. An impossible situation arises when the length of the sequence segment is different from the length of the stick by 60%. Another impossible situation happens when the length of the loop is too short to make the connection of the two sticks. For example, the length of the loop between H_1 and H_2 is one amino acid [Figure 23 E]. Given the approximately 3.8-Å distance between two consecutive amino acids, the maximum distance between the two ends of the two sticks is about 3.8*(1+1) = 7.6Å. One extra amino acid has been added to estimate the length of half the amino acid at each end of the helix. Most of the edge weights in the graph have been assigned by tracing the skeleton. For any possible edge, the weight is calculated as follows: $w\big((i,j,t)(i',j',t')\big) = |l(i,i') - d(j,t,j',t')| + b$, in which $l(i,i')$ is 3.8 multiplied by the number of amino acids between S_i and $S_{i'}$, measured in the protein sequence, and $d(j,t,j',t')$ is the distance estimated along the skeleton trace between S_i and $S_{i'}$ when they are assigned to D_j at end t and $D_{j'}$ at end t', respectively. The skeleton voxels between two SSE-Ds have been used to track the traces, using the component labeling cluster. Even if the loop connections between β-strands are unclear in the skeleton density map, the skeleton trace is used to optimize the edge weight between two β-strands. If there is a continuous path or a gapped path along

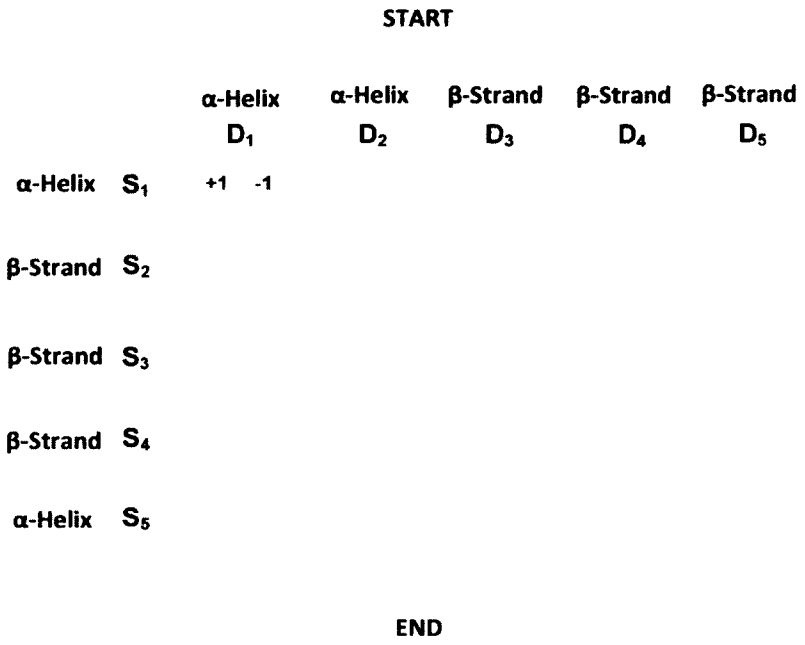


Figure 30. The graph of 2KUM, built with SSE-Ss and SSE-Ds. The solid nodes mean valid ones; the gray nodes represent nonexistent ones.

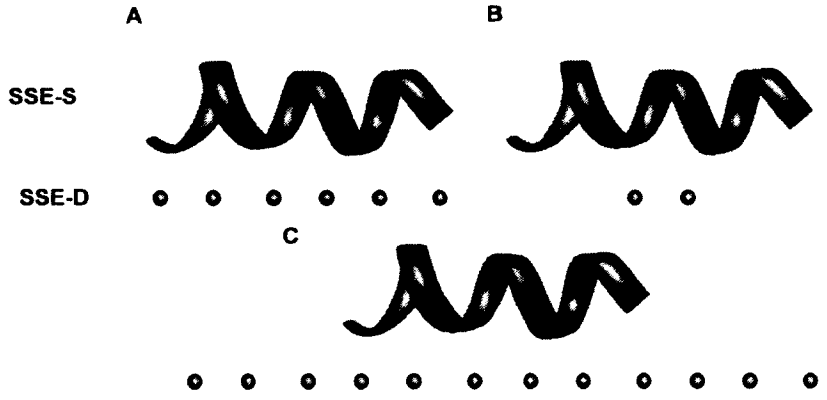


Figure 31. Comparison of the lengths of SSE-S and SSE-D for a node. (A) A valid SSE-S/SSE-D pair. (B) The length of SSE-S is too long. (C) The length of SSE-D is too long.

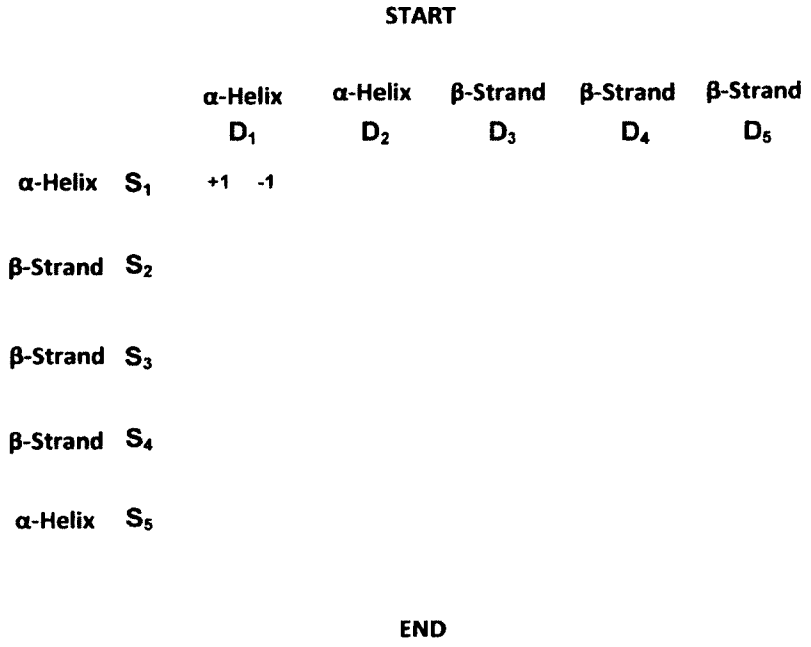


Figure 32. The graph of 2KUM with the unary constraints.

the skeleton, b is set to zero. Otherwise, $d(j,t,j',t')$ is estimated, using the Euclidian distance between t and t'. In this case, the penalty term is set to $b = 50$ unless the distance between t and t' is less than 7 Å. Since there are often multiple paths along the skeleton between two sticks, the path whose length best fit $l(i,i')$ is used for estimating $d(j,t,j',t')$.

3.1.3 Update Edge Weight, Using β-sheet Constraints

We have designed the following constraints to be biased toward the popular topologies, such as antiparallel strands with short loops.

Short loops and strand spacing. This constraint reflects the fact that two consecutive β-strands in the protein sequence are more likely to be neighboring strands in the density map. When the loop connecting two β-strands has less than five amino acids, this constraint applies. We require that $gap_{seq}(i,j) \geq gap_{stick}(k,l)$, in which $gap_{seq}(i,j) = |i-j|$, $1 \leq i \leq j \leq M$, and $gap_{stick}(k,l) = \lfloor (D(k,l) + \varepsilon)/4.5 \rfloor$, $1 \leq k < l \leq N_\beta$. ε be a tolerance parameter, where $D(k,l)$ is the measured shortest Euclidian distance between the two β-sticks D_k and D_l. As an example, the two consecutive β-strands are not likely to be assigned to strands 1 and 4 [Figure 33 B]. We set a penalty term of $50 * (gap_{stick} - gap_{seq})$ to the edge weight if two connected nodes have $gap_{seq} < gap_{stick}$.

Two-stranded antiparallel sheet. When two consecutive β-segments in the sequence are assigned to two β-sticks that are immediate neighbors, we create a bias toward antiparallel strands when the loop is not long enough to make a parallel relationship. When the loop is shorter than the length of the second β-stick, we require $D_{EE} > D_{ES}$ (Figure 33). A penalty term of 150 is charged for the violation.

Three strands. For most of the popular topologies, three consecutive strands form an antiparallel relationship. A penalty is imposed if $D_{ES} < D_{EE}$ and $mod(\text{gap}_{seq}, 2) = 0$ or if $D_{ES} > D_{EE}$ and $mod(gap_{seq}, 2) = 1$.

Neighboring strands. This constraint awards the assignment of two consecutive β-strands in the sequence toward two neighbors. When the loop between the β-strands is less than 5 amino acids, we set a reward of $-3.8 * 3$.

Long helix matching. If the length between a long helix in the sequence and that of the α-stick is less than 15% of the stick, a reward of -5 is given.

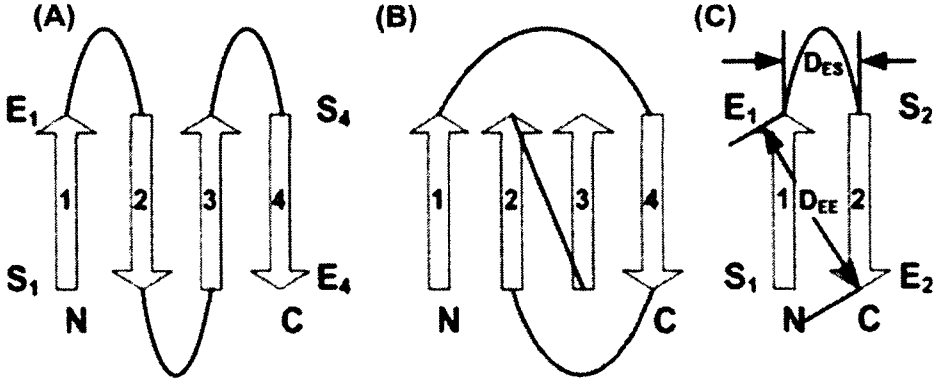


Figure 33. Popular topologies and β-sheet constraints [4]. (A) A popular antiparallel β-sheet topology. (B) A rare topology. (C) The diagonal D_{EE} is generally longer than the side of a rectangular D_{ES}. The start and end points are labeled for each strand in (A and C).

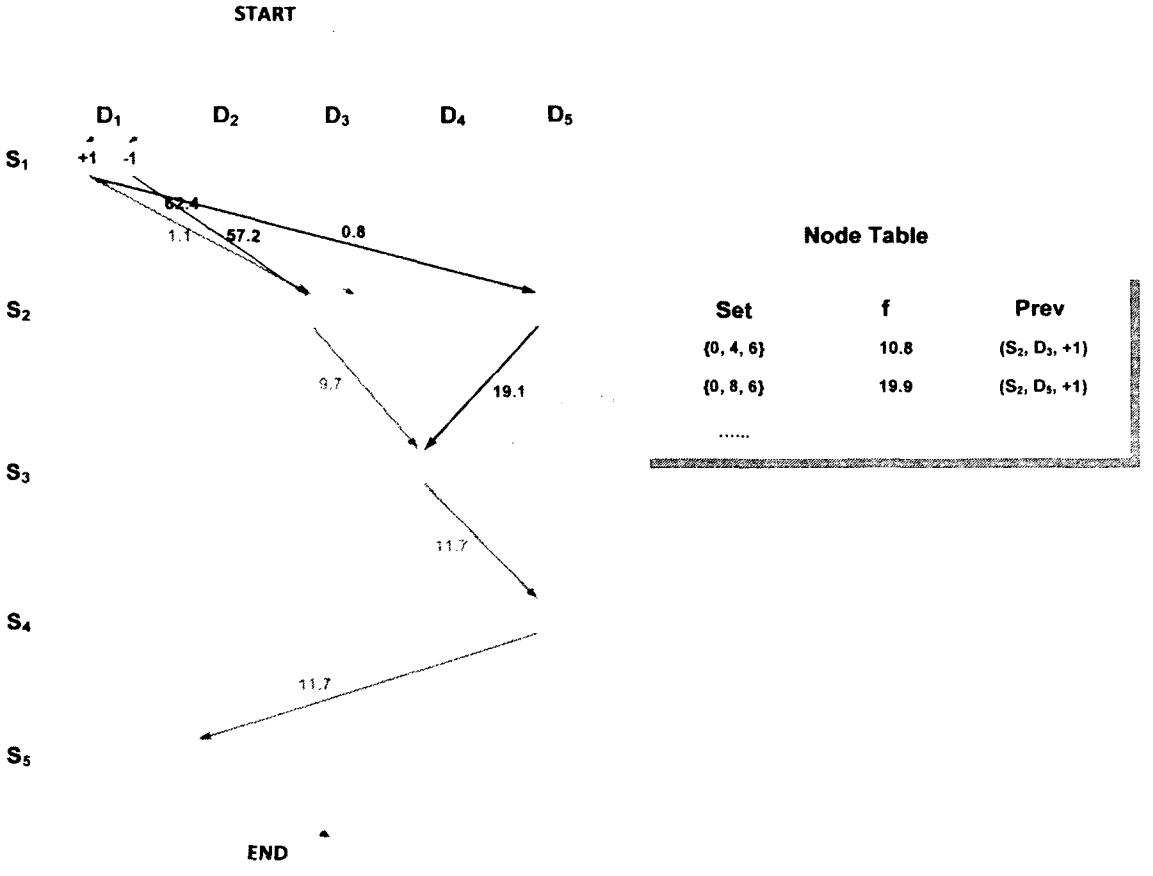


Figure 34. The graph of 2KUM with some of the edge weights. A true topology (shortest path) is shown by the thick red line; a wrong topology is shown by the thick blue line. A record table for node (S_3, D_4, +1) is shown on the right side.

Figure 34 shows the graph for 2KUM, using all constraints. Due to too many edge lines between the nodes, we only show some of the edges. The true topology is the shortest path in this sample, which is shown by the thick red lines. The edge weight (1.1) of [(S_1, D_1, +1), (S_2, D_3, +1)] is much less than the edge weight (57.2) of [(S_1, D_1, +1), (S_2, D_3, -1)], which means that entering the same stick from a different direction can be distinguished easily.

3.1.4 Generate the Node Table and Search the Shortest Path

A valid topology is a valid path (thick red line in Figure 34) from *START* to *END* and visits each column exactly once. The optimal path is one with the minimum cost, measured as the sum of the edge weights along the path. Al Nasr previously gave a dynamic programming algorithm to find the constrained shortest path [164]. We provide a dynamic programming algorithm in Algorithm 1 to find the shortest valid path in a general case when $M \geq N$.

Algorithm 1

/* Notations:

- $C = \{1,2,3,\dots,N\}$ $\quad U^{(i)} = \{U: U \subseteq C \text{ and } \max(1, i-(M-N)) \leq |U| \leq \min(i,N)\}, 2 \leq i \leq M \text{ and } M \geq N \geq 2.$
- $U_k^{(i)}$: the k^{th} element of $U^{(i)}$, $1 \leq k \leq |U^{(i)}|$.
- $v_{(j,t)}^i$: the node at i^{th} row, j^{th} column with t direction.
- $f\left((i,j,t), U_k^{(i)}\right) \leftrightarrow f(v_{(j,t)}^i, U_k^{(i)})$.

*/

input*: *G

output: The cost of the shortest path min_{cost}

$C \leftarrow \{1,2,3,\ldots,N\}$

$f\left(*, U_k^{(i)}\right) \leftarrow 0, |U_k^{(i)}| = 1, 1 \leq i \leq M - N + 1, 1 \leq k \leq |U^{(i)}|$

$f\left(*, U_k^{(i)}\right) \leftarrow \infty, |U_k^{(i)}| \neq 1, 2 \leq i \leq M, 1 \leq k \leq |U^{(i)}|$

for $i \leftarrow 2$ *to* M ***do***

for $k \leftarrow 1$ *to* $|U^{(i)}|$ ***do***

for each $p \in U_k^{(i)}, |U_k^{(i)}| > 1$ *and* $t \leftarrow 0$ *to* 1 ***do***

$U' \leftarrow U_k^{(i)} \backslash p$

for each $q \in U'$ *and* $t' \leftarrow 0$ *to* 1 ***do***

$$f\left(v_{(p,t)}^{i}, U_k^{(i)}\right) = \min_{\max(1,i-(M-N)-1) \leq i' < i} \left\{ f\left(v_{(q,t')}^{i'}, U'\right) + w\left(v_{(q,t')}^{i'}, v_{(p,t)}^{i}\right), f\left(v_{(p,t)}^{i}, U_k^{(i)}\right) \right\}$$

The idea of our method is to keep track of the columns visited along the path at each node, as well as the best score of all paths using these columns. A record table (Figur 34) is created for each node. Each record contains the set of columns U^i, the minimum cost f of the path to reach the current node, and the previous node; U^i represents all columns visited for a valid path. The value of f can be calculated by equation (2); the previous node lies before the current node in the shortest path passing all columns in U^i. Figure 34 illustrates the dynamic programming process for 2KUM at the node (S_3, D_4, +1). The first record in the table represents the red line path, which passes (D_1, +1), (D_3, +1), and (D_4, +1), or {0, 4, 6 represented by the SSE-D index.

This set represents the 3! path. The minimum cost of this 3! path is saved as the value of *f*. To track the path with the minimum cost, start from the current node and trace back to the previous node in the record. At this previous node, search the record with the set U^i/i, where i is the SSE-D index of the current node. Repeat the trace step until the start node is reached. All the traced nodes consist of the shortest path from the start node to the node (S_3, D_4, +1).

$$f((i,j,t),U) = f(v,U)$$
$$= \begin{cases} 0 & v =< START > \\ w(< START >, v) = 0 & skip_\alpha(0,i) \le M_\alpha - N_\alpha,\ skip_\beta(0,i) \le M_\beta - N_\beta,\ U = \{j\} \\ min_{j' \in U \setminus \{j\}, t' \in \{0,1\}} \left[f\left((i',j',t'), U \setminus \{j\}\right) + w\left((i',j',t'),(i,j,t)\right) \right] & \\ & i \in [2,M], skip_\alpha(i',i) \le M_\alpha - N_\alpha, \\ & skip_\alpha(i',i) \le M_\beta - N_\beta, j \in U \\ \infty & otherwise \end{cases} \quad (14)$$

2.2 and have the cost of the paths in nondecreasing order. Many algorithms have been developed to find the *K*-shortest paths without constraints. Yen {Yen, 1971 #217} proposed a classical deviation algorithm to find the *K*-shortest loopless paths. Due to the topology constraints, we cannot directly apply the available *K*-shortest path algorithm. Instead, we combine the concept of the "generalization of Yen's algorithm" with our dynamic programming method to find the constrained *K*-shortest paths.

The idea of finding the next shortest path is that the $(k+1)^{th}$ shortest path is not too different from the previous k shortest paths. It is at least one edge different from each of the previous k shortest paths. At each cycle, new candidates for the $(k+1)^{th}$ shortest path are generated in an edge deletion process and deposited in X, a set of the candidate paths. The $(k+1)^{th}$ shortest path is to be selected as the shortest path from X at iteration $k+1$.

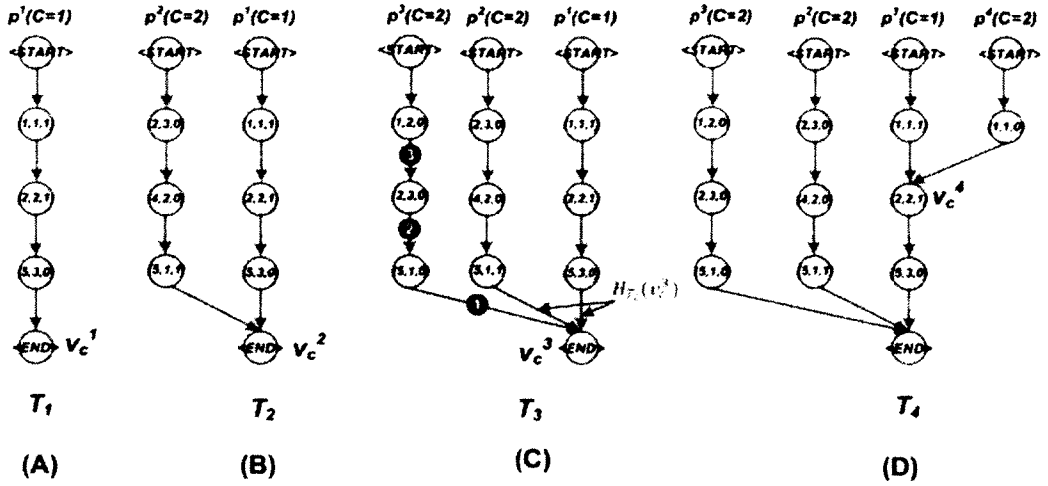


Figure 35. The reverse pseudo tree for the first four shortest paths [4]. The edges to be deleted to generate new candidates for p^4 are also marked in red [($H_2(v_c^3)$ + edges numbered 1, 2, and 3].

The edge deletion process generates new candidates for the next shortest path. A candidate for the second shortest path p^2 is generated by deleting one edge of p^1 at a time from the topology graph G_{Top}, starting from the last edge $e(v_N^1, END)$. Initially, we set the coinciding node of p^1 as $< END >$ and delete $e(v_N^1, END)$. This assumes that it is possible for p^2 to be the shortest path from $< START >$ to the coinciding node without using $e(v_N^1, END)$, which is the incoming edge to the coinciding node $< END >$. Generally, to obtain the $(k+1)^{\text{th}}$ shortest path, each new candidate is generated by deleting the head edges in $H_{T_{k-1}}(v_c^k)$ and an edge $e(v_{i-1}^k, v_i^k)$ of p^k, where $2 \leq i \leq v_c^k$. $H_{T_{k-1}}(v_c^k)$ is the set of edges whose head node is the coinciding node v_c^k of p^k in the reverse pseudo tree T_{k-1} [Figure 35 C]. The reason for deleting the head

edges is to avoid generating a new candidate that is the same as a previous shortest path. After all candidate paths have been generated from path p^k, the deleted edges are restored to the graph.

The k shortest path search algorithm uses our dynamic programming's constrained shortest path algorithm as the starting point. After certain edges are deleted from G_{Top}, a naïve way to find a candidate path that satisfies the constraints is to scan all the nodes v below the deleted edges to update $f(v, U)$ that was stored at each node. We provide Algorithm 2 for the top-k topologies search.

Algorithm 2: Finding constrained *K*-shortest paths

Notation:

- $p^k = < START = v_0^k, v_1^k, \dots, v_{N+1}^k = END >$: the k^{th} shortest path.
- $\mathcal{T}_k$: The reverse pseudo tree of the k shortest paths.
- X : A set contains candidate paths for the k shortest paths.
- $p_{i,j}^k$: The path from node v_i^k to node v_j^k in the k^{th} shortest path.
- $U(p_{i,j}^k)$: The subset of columns visited in path $p_{i,j}^k$.
- $H_{\mathcal{T}_k}(v)$: The set of edges in $\mathcal{T}_k$ whose head node is v.

***input*: G_{Top}, K.**

***output:* The reverse-pseudo-tree of K shortest paths, $\mathcal{T}_K$.**

$C \leftarrow \{1,2,3, \dots, N\}$

$k \leftarrow 1$

$p^k \leftarrow shortest\ path\ in\ G_{TOP}$ //the path with min cost

$\mathcal{T}_k \leftarrow p^k$

$X \leftarrow \{p^k\}$

While $(X \neq \emptyset \text{ and } k < K)$ ***do***

$X \leftarrow X - \{p^k\}$

$v_c^k \leftarrow$ the coinciding node of p^k

Remove edges $H_{\mathcal{T}_{k-1}}(v_c^k)$ from G_{Top}

for each $v_x^k \in p_{2,c}^k$

Remove edge (v_{x-1}^k, v_x^k)

$U' \leftarrow S \backslash U(p_{x,N}^k)$

$q \leftarrow$ the shortest path from $START$ to v_x^k for the set of columns in U'

//The path verifies $\min_{v' \in V}(f(v', U') + w(v', v_x^k))$

$q \leftarrow q \diamond p_{x,END}^k$

$X \leftarrow X \cup \{q\}$

End for

Restore removed edges to G_{Top}

$k \leftarrow k + 1$

$p^k \leftarrow$ shortest path in X

$\mathcal{T}_k \leftarrow \mathcal{T}_{k-1} + p^k$

End

return $\mathcal{T}_k$

3.2 Results and Discussion

The topology graph and the dynamic programming algorithm apply in principle to both α-proteins and α/β proteins. In practice, it is more challenging to derive topologies for proteins with β-sheets due to the close spacing of about 4.5 Å between two β-strands. We have applied additional constraints to be biased toward known popular topologies of β-sheets. We have used seven simulated density maps and two experimentally derived maps in the test. The β-strand locations were visually detected since there was no automatic tool to detect β-strands from a β-sheet when the work was performed. To evaluate the accuracy of the method, we have used the rank of the native topology on the list sorted by the score.

It appears that the framework of the top-k topology algorithm generally applies to the proteins with both α-helices and β-sheets. It was able to rank the native topology among the top 25 for seven out of nine proteins when no β-constraints were added for β-sheets (column 6 of Table 11). The β-sheet constraints are effective in identifying the native topology. For example, the protein extracted from the density map with EMDB ID 1733 has 5 α-sticks and 12 β-sticks. In this case, SSETracer has detected all five α-helices and three β-sheets. The native topology has not been found within the top 100 topologies without β-constraints, but it has been ranked 13^{th} out of 7.5e + 15 total possible topologies after using the constraints. Although there are $5!\,2^5 12!\,2^{12} \approx 7.5e+15$ different topologies, those that satisfy the density requirement and the β-sheet constraints can be quite limited. The results presented in this paper further support our previous finding [231] about the amazing properties of SSE topologies—the native topology is near the top of the entire topological space.

The results show improved accuracy and reduced memory and time in ranking the top 25 topologies. Although we have previously proven that the secondary structure topology problem is an NP-hard problem, with the computational approaches in this paper, we show that it is possible to use a generic desktop to derive the topology for a large protein with 5 helices and 20 β-strands. The results represent a major improvement in the ability to derive the secondary structure topology automatically for large and complicated density maps containing both α-helices and β-sheets.

Table 11. The rank of the native topology in α-β proteins [4].

IDEMDB	#Helices[a]	#Strand[b]	Sheet_ID[c]	#Total[d]	Rank_NC[e]	Rank[f]
5030	4/3	4/3	A	3.7e+04	1	1
1733	5/5	12/12	O,P,Q	7.5e+15	-/100	13
1OZ9	5/5	5/4	A	7.7e+05	25	7
2KUM	2/2	3/3	A	3.8e+02	5	1
2KZX	3/3	3/3	A	2.3e+03	10	10
2L6M	2/2	3/3	A	3.8e+02	6	6
1BJ7	5/1	9/9	A	1.9e+09	-/100	4
1ICX	6/3	7/7	A	6.2e+08	2	1
1JL1	4/4	5/5	A	1.5e+06	22	16

a. The number of α-helices in the protein sequence / the number of α-sticks detected from the density map.

b. The number of β-strands in the protein sequence / the number of β-strands visually detected.

c. β-sheet ID.

d. The total number of possible topologies.

e. The rank of the native topology without β-constraints; -/100: the native topology not found in top 100 topologies.

f. The rank of the native topology with β-constraints.

REFERENCES

1. Chen, L., Al Nasr, K. & He, J. (2013). Using Constraints in Modeling the Protein Beta-Sheet Topology. In *Capstone Conference*, Suffolk, VA.
2. Chen, L. & He, J. (2014). A Distance and Orientation Dependent Energy Function of Amino Acid Functional Blocks. *Biopolymers* **101**, 681-92.
3. Meyerson, J. R. e. (2011). Determination of Molecular Structures of HIV Envelope Glycoproteins using Cryo-Electron Tomography and Automated Sub-tomogram Averaging. *J. Vis. Exp.* **1**, 2770.
4. Al Nasr, K., Ranjan, D., Zubair, M., Chen, L. & He, J. (2014). Solving the secondary structure matching problem in de novo modeling using a constrained K-shortest path graph algorithm. *IEEE Transaction on Computational Biology and Bioinformatics* **11**, 419-29.
5. Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P. & Gasteiger, E. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**, 136.
6. Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourned, P. E. & Burley, S. K. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **43**, D345-56.
7. Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S. J., Newman, R. H., Oldfield, T. J., Rees, I., Sahni, G., Sala, R., Velankar, S., Warren, J., Westbrook, J. D., Henrick, K., Kleywegt, G. J., Berman, H. M. & Chiu, W. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456-D464.
8. Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). EMAN: Semi-automated software for high resolution single particle reconstructions. *J. Struct. Biol.* **128**, 82-97.

9. Si, D. & He, J. (2013). Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps. In *BCB'13: Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics.*

10. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605-1612.

11. Baker, M. L., Abeysinghe, S. S., Schuh, S., Coleman, R. A., Abrams, A., Marsh, M. P., Hryc, C. F., Ruths, T., Chiu, W. & Ju, T. (2011). Modeling protein structure at near atomic resolutions with Gorgon. *J. Struct. Biol.* **174**, 360-73.

12. Tropp, B. E. (2012). *Principles of Molecular Biology*, Jones & Bartlett Learning.

13. Alberts, B., Johnson, A., J, L. & al., e. (2002). *Molecular Biology of the Cell.* 4th edition edit, Garland Science, New York.

14. D. L. Nelson, M. C. (2008). *Lehninger Principles of Biochemistry 5th Edition.* 5 edit, M. W. H. Freeman.

15. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

16. Hvidsten, T. R., Lægreid, A., Kryshtafovych, A., Andersson, G., Fidelis, K. & Komorowski, J. (2009). A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity. *PLoS ONE* **4**, e6266.

17. Christiansen, C., Hachem, M. A., Friis, E., Baumann, M. J., Glaring, M. A., Viksø-Nielsen, A., Sigurskjold, B. W., Svensson, B. & Blennow, A. (2008). *Starch Recent Progress in Biopolymer and Enzyme Technology*, Polish Society of Food Technologists.

18. Gloster, T. M., Ibatullin, F. M., Macauley, K., Eklöf, J. M., Roberts, S., Turkenburg, J. P., Bjørnvad, M. E., Jørgensen, P. L., Danielsen, S., Johansen, K. S., Borchert, T. V., Wilson, K. S., Brumer, H. & Davies, G. J. (2007). Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *Journal of Biological Chemistry* **282**, 19177-19189.

19. Von Ossowski, I., Eaton, J. T., Czjzek, M., Perkins, S. J., Frandsen, T. P., Schülein, M., Panine, P., Henrissat, B. & Receveur-Bréchot, V. (2005). Protein disorder: Conformational distribution of the flexible linker in a chimeric double cellulase. *Biophys. J.* **88**, 2823-2832.

20. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-230.

21. White, S. H. & Wimley, W. C. (1999). Memberane Protein Folding and Stability: Physical Principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319-65.

22. Ridley, M. (2000). *Genome*, Harper Perennial, New York.

23. Wagner, I. & Musso, H. (1983). New Naturally Occurring Amino Acids. *Angewandte Chemie International Edition in English* **22**, 816-828.

24. Hausman, R. E. & Cooper, G. M. (2004). *The cell: a molecular approach*, ASM Press, Washington, D.C.

25. Carl, B. & Tooze, J. (1997). *Introduction to Protein Structure*, Garland Publishing, Inc., New York and London.

26. Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of Polypeptides and Proteins. In *Advances in Protein Chemistry* (C.B. Anfinsen, M. L. A. J. T. E. & Frederic, M. R., eds.), Vol. Volume 23, pp. 283-437. Academic Press.

27. Brocchieri, L. & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390-3400.

28. Zhang, J. (2000). Protein-length distributions for the three domains of life. *Trends Genet.* **16**, 107-9.

29. Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A. & Kolchanov, N. A. (2005). PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* **33**, D183-7.

30. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. D. & Shore, V. C. (1960). Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature* **185**, 422-427.

31. Voet, D. & Voet, J. G. (2004). *Biochemistry*. 3rd ed. edit, Wileys, Hoboken, NJ.

32. Pauling, L. & Corey, R. B. (1951). The structure of hair, muscle, and related proteins. *Proc. Nat. Acad. Sci. USA* **37**, 261-271.

33. Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci.* **37**, 205-211.

34. Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425-36.

35. Dunitz, J. D. (2001). Pauling's Left-Handed α-Helix. *Angewandte Chemie International Edition* **40**, 4167-4173.

36. Guzzo, A. V. (1965). The Influence of Amino Acid Sequence on Protein Structure. *Biophys. J.* **5**, 809-822.

37. Lewis, P. N., Go, N., Go, M., Kotelchuck, D. & Scheraga, H. A. (1970). Helix Probability Profiles of Denatured Proteins and Their Correlation with Native Structures. *Proc. Natl. Acad. Sci.* **65**, 810-815.

38. Schiffer, M. & Edmundson, A. B. (1967). Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential. *Biophys. J.* **7**, 121-135.

39. Richardson, J. S. & Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci.* **99**, 2754-2759.

40. Némethy, G. & Printz, M. P. (1972). The -Turn, a Possible Folded Conformation of the Polypeptide Chain. Comparison with the β-Turn. *Macromolecules* **5**, 755.

41. Alexander, M. (1989). Macromolecular Crystals: The Growth of Crystals Is Now the Key to Deducing the Structure of Large Molecules. *Scientific American* **260**, 62-69.

42. Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol* **145**, 215-250.

43. Janin, J. & Chothia, C. (1980). Packing of alpha-helices onto beta-pleated sheets and the anatomy of alpha/beta proteins. *J. Mol. Biol.* **143**, 95-128.

44. Bonneau, R. & Baker, D. A. (2001). Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173-189.

45. Kendrew, J. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181**, 662-666.

46. Kendrew, J. C. (1961). The three-dimensional structure of a protein molecule. *Sci. Am.* **205**, 96-110.

47. All Nobel Prizes, Vol. 2015.

48. Klug, A. (2010). From virus structure to chromatin: X-ray diffraction to three-dimensional electron microscopy. *Annu Rev Biochem* **79**, 1-35.

49. Blow, D. M. & Rossmann, M. G. (1961). The single isomorphous replacement technique. *Acta Crystallogr.* **14**, 1195-1202.

50. Rossmann, M. G. (1984). Constraints on the assembly of spherical virus particles. *Virology* **134**, 1-11.

51. Berman, H. M. (2008). The Protein Data Bank: a historical perspective. *Acta Crystallographic A.* **64**, 88-95.

52. Langridge, R. (1974). Interactive Three-Dimensional Computer Graphics in Molecular Biolog. *Fed. Proc.* **33**, 2332-2335.

53. Ferrin, T. E. & Langridge, R. (1980). Interactive Computer Graphics with the UNIX Time-Sharing System. *Computer Graphics* **13**, 320-331.

54. Wagner, G. & Wuthrich, K. (1978). Dynamic model of globular protein conformations based on NMR studies in solution. *Nature* **275**, 247-248.

55. Richardson, J. S. (2000). Early ribbon drawings of proteins. *Nature Structural Biology* 7, 624-625.

56. Adrian, M., Dubochet, J., Lepault, J. & McDowall, A. W. (1984). Cryo-electron microscopy of viruses. *Nature* **308**, 32-36.

57. Callaway, E. (2015). The revolution will not be crystallized: a new method sweeps through structural biology. *Nature* **525**, 172-174.

58. Abad-Zapatero, C., Abdel-Meguid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossmann, M. G., Suck, D. & Tsukihara, T. (1980). Structure of southern bean mosaic virus at 2.8 Å resolution. *Nature* **286**, 33-39.

59. Chandonia, J. & Brenner, S. E. (2006). The Impact of Structural Genomics: Expectations and Outcomes. *Science* **311**, 347-51.

60. Huber, R., Deisenhofer, J. & Colman, P. M. (1976). Crystallographic structure studies of an IgG molecule and an Fc fragment. *Nature* **264**, 415-20.

61. Casjens, S. & Hendrix, R. (1988). Control mechanisms in dsDNA bacteriophage assembly. In *In: The Bacteriophages* (Calendar, R., ed.), Vol. 1, pp. 15-91. Plenum Press, New York.

62. Dorn, M., Silva, M. B., Buriol, L. S. & Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry* **53**, 251-276.

63. Frenkel, D. & Smit, B. (1996). *Understanding Molecular Simulation from Algorithms to Applications*. Computational Science Series, Academic Press, San Diego.

64. Zhang, Y. (2008). Progress and Challenges in Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **18**, 342-348.

65. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. (2012). Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proc. Natl. Acad. Sci.* **109**, 17845-17850.

66. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science* **334**, 517-520.

67. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y. & Wriggers, W. (2010). Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330**, 341-346.

68. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. (2009). Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **19**, 1-18.

69. Lindert, S., Alexander, N., Wotzel, N., Karaka, M., Stewart, P. L. & Meiler, J. (2012). EM-Fold: De Novo Atomic-Detail Protein Structure Determination from Medium-Resolution Density Maps. *Structure* **20**, 464-478.

70. Feynman, R. P. The Feynman Lectures on Physics. Addison-Wesley, USA.

71. Segre, E. (1980). *From X-Rays to Quarks: Modern Physicists and Their Discoveries*, W. H. Freeman and Company, New York.

72. Dickinson, R. G. & Raymond, A. L. (1923). The Crystal Structure of Hexamethylene-Tetramine. *J. Ams. Chem. Soc.* **45**, 22.

73. Pauling, L. (1929). The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **51**, 1010-1026.

74. Suryanarayana, C. & Norton, M. G. (1998). *X-Ray Diffraction: A Practical Approach*, Springer Science+Bussiness Media, LLC, New York.

75. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18**, 581-586.

76. Guentert, O. J. & Cvikevich, S. (1964). Preferred orientation and its effect on the (hk) reflections in X-ray patterns of pyrolytic graphites. *Carbon* **1**, 309-313.

77. Woolfson, M. (1997). *An Introduction to X-ray Crystallography*. 2nd edition edit, Cambridge University Press.

78. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. (2007). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal* **275**, 1-21.

79. Stewart, G. W. (1931). X-Ray Diffraction in Water: The Nature of Molecular Association. *Phys. Rev.* **37**, 9.

80. Wong, K. C. (2014). Review of NMR Spectroscopy: Basic Principles, Concepts and Applications in Chemistry. *J. Chem. Educ.* **91**, 1103-1104.

81. The Nobel Prize in Chemistry 1962, Vol. 2015.

82. Moudgil, K. D., Rao, D. N. & Narang, B. S. (1985). Nuclear magnetic resonance and its applications in medicine. *The Indian Journal of Pediatrics* **52**, 231-241.

83. Silverstein, R. M., Webster, F. X. & Kiemle, D. J. (2005). *Spectrometric Identification of Organic Compounds*. 7th edition edit, Wiley.

84. Ferentz, A. E. & Wagner, G. (2000). NMR spectroscopy: a multifaceted approach to macromolecular structure. *Quarterly Reviews of Biophysics* **33**, 29-65.

85. Hobbie, R. K. (1988). *Intermediate Physics for Medicine and Biology*. 2nd Ed. edit, Wiley.

86. Kaiser, R. (1962). Use of the Nuclear Overhauser Effect in the Analysis of High - Resolution Nuclear Magnetic Resonance Spectra. *J. Chem. Phys.* **39**, 2435.

87. Ni, F. & Scheraga, H. A. (1994). Use of the Transferred Nuclear Overhauser Effect To Determine the Conformations of Ligands Bound to Proteins. *Accounts of Chemical Research* **27**, 257-264.

88. Guntert, P. (2004). Automated NMR Structure Calculation With CYANA. *Methods Mol. Biol.* **278**.

89. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* **160**, 65-73.

90. Zheng, Y. & Yang, D. (2005). STARS: statistics on inter-atomic distances and torsion angles in protein secondary structures. *Bioinformatics* **21**, 2925-6.

91. Yu, H. (1999). Extending the size limit of protein nuclear magnetic resonance. *Proc. Natl. Acad. Sci.* **96**, 332-334.

92. Pervushin, K. V., Riek, R., Wider, G. & Wuthrich, K. (1997). Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci.* **94**, 12366-71.

93. Hefke, F., Schmucki, R. & Guntert, P. (2013). Prediction of peak overlap in NMR spectra. *J. Biomol. NMR* **56**, 113-23.

94. Glaeser, R. M. & Hall, R. J. (2011). Reaching the Information Limit in Cryo-EM of Biological Macromolecules: Experimental Aspects. *Biophys. J.* **100**, 2331-2337.

95. Kourkoutis, L. F., Plitzko, J. M. & Baumeister, W. (2012). Electron Microscopy of Biological Materials at the Nanometer Scale. *Annual Review of Materials Research* **42**, 33-58.

96. Milne, J. L., Borgnia, M. J., Bartesaghi, A., Tran, E. E., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A. & Subramaniam, S. (2013). Cryo-electron microscopy--a primer for the non-microscopist. *FEBS Journal* **280**, 28-45.

97. Kong, Y., Zhang, X., Baker, T. S. & Ma, J. (2004). A Structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps. *J. Mol. Biol.* **339**, 117-30.

98. Dubochet, J. (2012). Cryo-EM—the first thirty years. *Journal of Microscopy* **245**, 221-224.

99. Cheng, Y. (2015). Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161**, 450-457.

100. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. (2015). A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **161**, 438-449.

101. Frank, J. (2006). *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*, Oxford University Press, Oxford.

102. Lucic, V., Rigort, A. & Baumeister, W. (2013). Cryo-electron tomography: The challenge of doing structural biology in situ. *J. Cell Biol.* **202**, 407-19.

103. Boekema, E. J., Folea, M. & Kouril, R. (2009). Single particle electron microscopy. *Photosynth Res* **102**, 189-196.

104. Mancini, E. J., Clarke, M., Gowen, B. E., Rutten, T. & Fuller, S. D. (1999). Cryo-Electron Microscopy Reveals the Functional Organization of an Enveloped Virus, Semliki Forest Virus. *Molecular Cell* **5**, 255-266.

105. Zhang, X., Jin, L., Fang, Q., Hui, W. H. & Zhou, Z. H. (2010). 3.3 Å Cryo-EM Structure of a Nonenveloped Virus Reveals a Priming Mechanism for Cell Entry. *Cell Biochemistry and Biophysics* **141**, 472-482.

106. Cong, Y., Baker, M. L., Jakana, J., Woolford, D., Miller, E. J., Reissmann, S., Kumar, R. N., Redding-Johanson, A. M., Batth, T. S., Mukhopadhyay, A., Ludtke, S. J., Frydman, J. & Chiu, W. (2010). 4.0-Å resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. *Proc. Natl. Acad. Sci.* **107**, 4967-4972.

107. Zhang, R., Hryc, C. F., Cong, Y., Liu, X. G., Jakana, J., Gorchakov, R., Baker, M. L., Weaver, S. C. & Chiu, W. (2011). 4.4 angstrom cryo-EM structure of an enveloped alphavirus Venezuelan equine encephalitis virus. *Embo Journal* **30**, 3854-3863.

108. Soejima, T., Sherman, M. B., Schmid, M. F. & Chiu, W. (1993). 4-Å projection map of bacteriophage T4 DNA helix-destabilizing protein (gp32*I) crystal by 400-kV electron cryomicroscopy. *J. Struct. Biol.* **111**, 9-16.

109. EMDB, Vol. 2015.

110. Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., Ginkel, G. v., Devkota, B., Lagerstedt, I., Ludtke, S. J., Newman, R. H., Oldfield, T. J., Rees, I., Sahni, G., Sala, R., Velankar, S., Warren, J., Westbrook, J. D., Henrick, K., Kleywegt, G. J., Berman, H. M. & Chiu, W. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456-64.

111. Suck, D., Rayment, I., Johnson, J. E. & Rossmann, M. G. (1978). The structure of southern bean mosaic virus at 5 Å resolution. *Virology* **85**, 187-97.

112. Smiley, I. E., Koekoek, R., Adams, M. J. & Rossmann, M. G. (1971). The 5 Å resolution structure of an abortive ternary complex of lactate dehydrogenase and its comparison with the apo-enzyme. *J. Mol. Biol.* **55**, 467-75.

113. Suzanne, C. (2008). RNA Splicing: Introns, Exons and Spliceosome. *Nature Education* **1**, 31.

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

115. Hark Gan, H. & al., e. (2002). Analysis of Protein Sequence/Structure Similarity Relationships. *Biophysical journal* **83**, 2781-2791.

116. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

117. Pearson, W. R. (2000). Protein sequence comparison and Protein evolution. In *International Conference Intelligent Systems for Molecular Biology*, San Diego, California.

118. Kopp, J. & Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics* **5**, 405-416.

119. AFiser, A. & Sali, A. (2003). *Comparative protein structure modeling. In protein structure: Determination, Analysis, and Application for drug discovery*, Marcel Dekker, Inc., New York.

120. Choo, Y., Castellanos, A., Garcia-Hernandez, B., Sanchez-Garcia, I. & Klug, A. (1997). Promoter-specific activation of gene expression directed by bacteriophage-selected zinc fingers. *J. Mol. Biol.* **273**, 525-32.

121. Daga, P. R., Y., P. R. & Doerksen, R. J. (2010). Template-based protein modeling: recent methodological advances. *Curr. Top Med. Chem.* **10**, 84-94.

122. Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* **153**, 1027-42.

123. Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507-33.

124. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.

125. Topf, M., Baker, M. L., Marti-Renom, M. A., Chiu, W. & Sali, A. (2006). Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *J. Mol. Biol.* **357**, 1655-68.

126. M. L. Baker, M. R. B., C. F. Hryc, adn F. Dimaio. (2010). Analyses of Subnanometer Resolution Cryo-EM Density Maps. *Methods Enzymol* **483**, 1-29.

127. Murzin, A. G., Hubbard, T. J. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

128. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-108.

129. FitzGerald, P. G. & Graham, D. (1991). Ultrastructural localization of alpha A-crystallin to the bovine lens fiber cell cytoskeleton. *Current Eye Res.* **10**, 417-436.

130. Zhang, Y. & Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci.* **102**, 1029-34.

131. Kihara, D. & Skamnaki, V. (2003). The PDB is a covering set of small protein structures. *J. Mol. Biol.* **334**, 793-802.

132. Lindert, S., Staritzbichler, R., Wotzel, N., Karakas, M., Stewart, P. L. & Meiler, J. (2009). EM-fold: De novo folding of alpha-helical proteins guided by

intermediate-resolution electron microscopy density maps. *Structure* **17**, 990-1003.

133. P. Bradley, K. M. M., D. Baker. (2005). Toward high-resolution de novo structure prediciton for small proteins. *Science* **309**, 1868-1871.

134. Ruczinski, I., Kooperberg, C., Bonneau, R. & Baker, D. (2002). Distributions of beta sheets in proteins with application to structure prediction. *Proteins: Structure, Function, and Genetics* **48**, 85-97.

135. Yen, J. Y. (1971). Finding the K Shortest Loopless Paths in a Network. *Management Science* **17**, 712-716.

136. Kong, Y. & Ma, J. (2003). A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *J. Mol. Biol.* **332**, 399-413.

137. Dal Palu, A., He, J., Pontelli, E. & Lu, Y. (2006). Identification of alpha-helices from low resolution protein density maps. *Comput. Syst. Bioinformatics Conf.*, 89-98.

138. Lasker, K., Dror, O., Shatsky, M., Nussinov, R. & Wolfson, H. J. (2007). EMatch: Discovery of High Resolution Structural Homologues of Protein Domains in Intermediate Resolution Cryo-EM Maps. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **4**, 28-39.

139. Al Nasr, K., Liu, C., Kwebangira, M., Burge, L. & He, J. (2013). Intensity-Based Skeletonization of CryoEM Grayscale Images Using a True Segmentation-Free Algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **10**, 1289-1298.

140. Rossmann, M. G. (2000). Fitting atomic models into electron-microscopy maps. *Acta Crystallogr. D Biol. Crystallogr.* **56**, 1341-9.

141. Tama, F., Miyashita, O. & Brooks, C. L., 3rd. (2004). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* **147**, 315-26.

142. Cowtan, K. (2008). Fitting molecular fragments into electron density. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 83-899.

143. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673-83.

144. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* **392**, 181-90.

145. Vasishtan, D. & Topf, M. (2011). Scoring functions for cryoEM density fitting. *J. Struct. Biol.* **174**, 333-43.

146. Woetzel, N., Lindert, S., Stewart, P. L. & Meiler, J. (2011). BCL::EM-Fit: rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. *J. Struct. Biol.* **175**, 264-76.

147. Pandurangan, A. P. & Topf, M. (2012). RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. *Bioinformatics.* **28**, 2391-3.

148. Topf, M. & Sali, A. (2005). Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.* **15**, 578-85.

149. Lu, Y., He, J. & Strauss, C. E. (2008). Deriving topology and sequence alignment for the helix skeleton in low-resolution protein density maps. *J. Bioinform. Comput. Biol.* **6**, 183-201.

150. Chivian, D. & Baker, D. (2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res.* **34**, e112.

151. Velazquez-Muriel, J. A. & Carazo, J. M. (2007). Flexible fitting in 3D-EM with incomplete data on superfamily variability. *J. Struct. Biol.* **158**, 165-81.

152. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033-44.

153. Dror, O., Lasker, K., Nussinov, R. & Wolfson, H. (2007). EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 42-9.

154. Baker, M. L., Ju, T. & Chiu, W. (2007). Identification of secondary structure elements in intermediate-resolution density maps. *Structure* **15**, 7-19.

155. Ludtke, S. J., Baker, M. L., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2008). De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure* **16**, 441-8.

156. Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci.* **90**, 7558-62.

157. Meiler, J., Muller, M., Zeidler, A. & Schmaschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular Modeling Annual* **7**, 360-369.

158. Meiler, J. & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci.* **100**, 12105-12110.

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

160. Pollastri, G. & McLysaght, A. (2005). Porter: a new accurate server for protein secondary structure prediction. *Bioinformatics* **21**, 1719-20.

161. Wu, Y., Chen, M., Lu, M., Wang, Q. & Ma, J. (2005). Determining protein topology from skeletons of secondary structures. *J. Mol. Biol.* **350**, 571-86.

162. Klepeis, J. L. & Floudas, C. A. (2003). Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* **24**, 191-208.

163. Fonseca, R., Glennie, H. & Pawel, W. (2011). Ranking Beta Sheet Topologies with Applications to Protein Structure Prediction. *Journal of Mathematical Modelling and Algorithms* **10**, 357-369.

164. Al Nasr, K., Ranjan, D., Zubair, M. & He, J. (2011). Ranking valid topologies of the secondary structure elements using a constraint graph. *J. Bioinform. Comput. Biol.* **9**, 415-30.

165. Fleishman, S. J., Harrington, S., Friesner, R. A., Honig, B. & Ben-Tal, N. (2004). An Automatic Method for Predicting Transmembrane Protein Structures Using Cryo-EM and Evolutionary Data. *Biophys J.* **87**, 3448-3459.

166. Cengel, Y. B., M. (2001). *Thermodynamics: An Engineering Approach*. 4th Edition edit, Mcgraw-Hill College, Boston.

167. Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York.

168. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187-217.

169. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R. Z. & Merz, K. M. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668-88.

170. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.

171. Sippl, M. J. (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**, 859-83.

172. Tanaka, S. & Scheraga, H. A. (1976). Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* **9**, 945-950.

173. Wu, Y., Lu, M., Chen, M., Li, J. & Ma, J. (2007). OPUS-Cα: A knowledge-based potential function requiring only Cα positions. *Protein Sci.* **16**, 1449-1463.

174. Russ, W. P. & Ranganathan, R. (2002). Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* **12**, 447-452.

175. Ravikant, D. V. S. & Elber, R. (2011). Energy design for protein-protein interactions. *J. Chem. Phys.* **135**, -.

176. Zhou, H. & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714-2726.

177. Shen, M. Y. & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507-2524.

178. Nanias, M., Chinchio, M., Pillardy, J., Ripoll, D. R. & Scheraga, H. A. (2003). Packing helices in proteins by global optimization of a potential energy function. *Proc. Natl. Acad. Sci.* **100**, 1706-10.

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

180. Sun, W. & He, J. (2009). Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topologies. *Proteins* 77, 159-73.

181. Sun, W. & He, J. (2011). From isotropic to anisotropic side chain representations: comparison of three models for residue contact estimation. *PLoS ONE* **6**, 0019238.

182. Lu, M., Dousis, A. D. & Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* **376**, 288-301.

183. Zhang, J. & Zhang, Y. (2010). A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLoS ONE* **5**, e15386.

184. Vendruscolo, M., Najmanovich, R. & Domany, E. (2000). Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins: Structure, Function, and Bioinformatics* **38**, 134-148.

185. Carter Jr, C. W., LeFebvre, B. C., Cammer, S. A., Tropsha, A. & Edgell, M. H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.* **311**, 625-638.

186. Krishnamoorthy, B. & Tropsha, A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics* **19**, 1540-1548.

187. Feng, Y., Kloczkowski, A. & Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins: Structure, Function, and Bioinformatics* **68**, 57-66.

188. Zhou, Y., Zhou, H., Zhang, C. & Liu, S. (2006). What is a desirable statistical energy functions for proteins and how can it be obtained? *Cell Biochemistry and Biophysics* **46**, 165-174.

189. Lu, H. & Skolnick, J. (2001). A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection. *Protein: Structure, Function, and Genetics* **44**, 223-232.

190. Samudrala, R. & Moult, J. (1998). An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. *J. Mol. Biol.* **275**, 895-916.

191. Buchete, N.-V., Straub, J. E. & Thirumalai, D. (2004). Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* **13**, 862-874.

192. Wu, S., Szilagyi, A. & Zhang, Y. (2011). Improving Protein Structure Prediction Using Multiple Sequence-Based Contact Predictions. *Structure* **19**, 182-91.

193. Zhou, H. & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophys. J.* **101**, 2043-2052.

194. Kolinski, A. & Bujnicki, J. M. (2005). Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* **61**, 84-90.

195. Gopal, S. M., Mukherjee, S. K., Cheng, Y. M. & Feig, M. (2010). PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins* **78**, 1266-81.

196. Gniewek, P., Leelananda, S. P., Kolinski, A., Jernigan, R. L. & Kloczkowski, A. (2011). Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. *Proteins.* **79**, 1923-9.

197. Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent Rotamer Library for Proteins: Application to Side-chain prediction. *J. Mol. Biol.* **230**, p543-74.

198. Misura, K. M. S., Morozov, A. V. & Baker, D. (2004). Analysis of Anisotropic Side-chain Packing in Proteins and Application to High-resolution Structure Prediction. *J. Mol. Biol.* **342**, 651-664.

199. Tuncbag, N., Salman, F. S., Keskin, O. & Gursoy, A. (2010). Analysis and network representation of hotspots in protein interfaces using minimum cut trees. *Proteins: Structure, Function, and Bioinformatics* **78**, 2283-94.

200. Belnap, D. M., Grochulski, W. D., Olson, N. H. & Baker, T. S. (1993). Use of radial density plots to calibrate image magnification for frozen- hydrated specimens. *Ultramicroscopy* **48**, 347-58.

201. Thorn, K. S. & Bogan, A. A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284-5.

202. Fischer, T. B., Arunachalam, K. V., Bailey, D., Mangual, V., Bakhru, S. & Russo, R. (2003). The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* **19**, 1453-4.

203. Wang, G. & Jr, R. L. D. (2003). PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.

204. Mukherjee, A., Bhimalapuram, P. & Bagchi, B. (2005). Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.* **123**, 014901.

205. Miyazawa, S. & Jernigan, R. L. (2005). How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins. *J. Chem. Phys.* **122**, 024901.

206. Samudrala, R. & Levitt, M. (2000). Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **9**, 1399-1401.

207. John, B., Sali, A. & Journals, O. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* **31**, 3982-3992.

208. Feng, Y., Kloczkowski, A. & Jernigan, R. L. (2010). Potentials 'R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics* **11**, 92.

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

210. Simons, K. T., Kooperberg, C., Huang, E. & 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**, 209-25.

211. Kim, D. E., Yi, Q., Gladwin, S. T., Goldberg, J. M. & Baker, D. (1998). The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* **284**, 807-15.

212. Gu, H., Kim, D. E. & Baker, D. (1997). Contrasting roles for symmetrically disposed β-turns in the folding of a small protein. *J. Mol. Biol.* **274**.

213. Chiu, W. (1986). Electron microscopy of frozen, hydrated biological specimens. *Annu. Rev. Biophys. Biophys. Chem.* **15**, 237-57.

214. Chiu, W., Avila-Sakar, A. J. & Schmid, M. F. (1997). Electron crystallography of macromolecular periodic arrays on phospholipid monolayers. *Adv. Biophys.* **34**, 161-72.

215. Zhou, Z. H., Dougherty, M., Jakana, J., He, J., Rixon, F. J. & Chiu, W. (2000). Seeing the herpesvirus capsid at 8.5 A. *Science* **288**, 877-80.

216. Ludtke, C. D., Song, S. J., Chuang, D. T. & Chiu, W. (2004). Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy. *Structure* **12**, 1129-36.

217. Chiu, W., Baker, M. L., Jiang, W. & Zhou, Z. H. (2002). Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. *Curr. Opin. Struct. Biol.* **12**, 263-9.

218. Yu, X., Jin, L. & Zhou, Z. H. (2008). 3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature* **453**, 415-9.

219. Del Palu, A., He, J., Pontelli, E. & Lu, Y. (2006). Identification of Alpha-Helices from Low Resolution Protein Density Maps. *Proceeding of Computational Systems Bioinformatics Conference(CSB)*, 89-98.

220. Si, D., Ji, S., Al Nasr, K. & He, J. (2012). A machine learning approach for the identification of protein secondary structure elements from cryoEM density maps. *Biopolymers* **97**, 698-708.

221. Zeyun, Y. & Bajaj, C. (2008). Computational Approaches for Automatic Structural Analysis of Large Biomolecular Complexes. *Computational Biology and Bioinformatics, IEEE/ACM Transactions* **5**, 568-582.

222. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110-9.

223. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126-32.

224. Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-235.

225. Cole, C., Barber, J. D. & Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197-W201.

226. McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.

227. Al Nasr, K., Ranjan, D. & He, J. (2010). Enumeration of the geometrically constrained assignment of the secondary structures using the constraint graph. In *Proceeding of the 3rd international conference on bioinformatics and system biology,* , Chongqing, China.

228. Ju, T., Baker, M. L. & Chiu, W. (2007). Computing a family of skeletons of volumetric models for shape description. *Comput. Aided Des.* **39**, 352-360.

229. Abeysinghe, S., Ju, T., Baker, M. L. & Chiu, W. (2008). Shape modeling and matching in identifying 3D protein structures. *Comput. Aided Des.* **40**, 708-720.

230. Grimson, W. L. & Lozano-Perez, T. (1987). Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-9**, 469-482.

231. Al Nasr, K., Ranjan, D., Zubair, M. & He, J. (2011). Ranking valid topologies of the secondary structure elements using a constraint graph. *J. Bioinform. Comput. Biol.* **9**, 15.

VITA

Lin Chen

Department of Computer Science, Old Dominion University, Norfolk, VA 23529

Education

Jan. 2010 – Current, Ph.D. candidate, Computer Science, Old Dominion University
Jul. 2004 – Dec. 2009, Master of Science, Physical Chemistry, New Mexico State University
Sep. 1997 – Jul. 2001, Bachelor of Science, Chemistry, Lanzhou University, P.R. China

Employment

Big Data Analytics Team, NASA Langley Research Center, Hampton VA
Software Developer — Feb. 2014 – Current

Tidewater Community College, Norfolk VA
Adjunct Faculty — Aug. 2013 – Dec. 2013

Old Dominion University, Norfolk VA
Teaching Assistant & Research Assistant — Jan. 2010 – Dec. 2013

New Mexico State University, Las Cruces NM
Teaching Assistant & Research Assistant — Aug. 2004 – Dec. 2009

Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences
Research Assistant — Sep. 2001 – Jun. 2004

Research Experience

Non-destructive evaluation, NASA Langley Research Center — Feb. 2014 - Current
Optimization of the protein top-K topologies search algorithm, Old Dominion University — Jan. 2012 – Dec. 2013
Protein energy function design, Old Dominion University — Jan. 2010 – Dec. 2011
Simulation of water purification by carbon nanotube, New Mexico State University — Sep. 2005 – Dec. 2009
Novel catalyst synthesis, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences — Sep. 2001 – Jun. 2004