Doctoral Thesis

Improving *de novo* model quality and its application in *ab initio* phasing

(de novo構造予測の改善とそのab initio位相決定への応用)

Rojan Shrestha

ロジャン　シュレスタ

A Dissertation Presented

By

Rojan Shrestha

Submitted to

The Graduate School of Frontier Sciences of the
University of Tokyo in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

August 2014

Department of Computational Biology

Abstract

De novo models are computationally predicted three-dimensional models of the given proteins using only amino acids sequence information. The key components of *de novo* modeling are the methods responsible for conformational space searching and the evaluation of each conformation accurately using energy function. The conformational space is astronomically large due to the degrees of freedom associated with each residue, which creates the challenge to develop the efficient method for searching the conformational space. Another challenge in *de novo* modeling is to devise an accurate energy function to evaluate the conformers. Despite these challenges, the *de novo* modeling has succeeded to generate accurate models for small and single domain proteins. Fragment assembly is an effective and efficient method for *de novo* modeling. This method assembles the fragments from known structures under the guidance of energy function. This concept was practically implemented in Rosetta, which achieved a number of break-through successes. Rosetta has two major stages, which are termed as coarse-grained sampling and all-atom refinement, to generate the final model from the input sequence. At the initial stage, three-residue and nine-residue fragments obtained from known structures are assembled to generate full-length coarse-grained models. These models contain only backbone atoms and the centroid of side-chain atoms. Subsequently, side-chain atoms were packed to construct all-atom models followed by energy minimization in all-atom refinement. However, there exist many challenges in the prediction of accurate models needed for practical use such as solving the crystallographic phase problem. To address these issues, I have focused on method development – biased conformation sampling and fragment quality improvement to enhance the quality of predicted models. Furthermore, I have developed the method to use *de novo* fragments for phasing and to assemble these fragments after phasing when full-length model is difficult to predict accurately for phasing.

First, I have developed a method to improve the conformational space search for accuracy improvement. This method first generated coarse-grained models using Rosetta. Second, an ensemble of lowest energy coarse-grained models was selected and deviation for each model from other models of the ensemble was calculated. The deviation for each residue was also computed and this score was called as average pair-wise residue distant score. The score correlated with the accuracy of predicted

residues in the model. When the predicted residues had larger scores, the residues were considered as less accurate and vice versa. Lastly, conformational search was biased using the score as residues with larger scores were given higher frequency for sampling. This procedure rebuilt selected coarse-grained models and then packed the side-chain atoms followed by energy minimization. Molecular replacement was run on these all-atom models and the entire simulation was terminated after a few correct solutions were obtained. This method was tested on 10 difficult targets, which were failed to achieve the success in previous studies using other methods - Rosetta and RosettaX. The rebuilding procedure improved the accuracy of coarse-grained models from 4.93 Å to 4.06 Å on average. Seven out of ten protein targets showed successful molecular replacement solution using rebuilt models.

The second method focused on improving the fragment quality to generate the better quality model. In this study, the method was developed to generate new fragment libraries using a resampling process. Therefore, the lowest energy all-atom models were selected after generation of models using Rosetta. These models were broken into overlapping fragments of three-residue and nine-residue. Average pair-wise residue deviation score was computed for three-residue and nine-residue fragments to remove distant fragments. The resultant fragments were clustered and then twenty-five fragments were randomly selected from the top five clusters. These new fragments were used for the second round of prediction. The performance of the method was tested on a benchmark set of 30 different proteins. The accuracy of new fragments and predicted models was evaluated. The result showed that the new fragment library contained better fragments and enriched with many high-quality fragments. In order to evaluate the performance, the lowest energy models and one of best from top five models were taken as the best prediction and computed their root mean square deviation of C-alpha atom (CA-RMSD), template modeling score (TM-score), and global distance test total score (GDT-TS) to the native structures. In all these assessment criteria, this method performed significantly better than Rosetta for lowest energy models and best in top five models. On average, this method improved CA-RMSD from 5.99 Å to 5.03 Å when lowest energy models were selected as the best predicted models. Similarly, it improved both the TM-score and GDT-TS by 7%.

Lastly, a new method was developed to tackle the phase problem using fragmentation and fragment reassembly approach when the full-length model was

inaccurate to use as the template model in molecular replacement. In this method, *de novo* model were fragmented, independently phased, and reassembled. A lowest energy all-atom models produced using Rosetta were chosen for fragmentation. For each residue position, constant-length overlapping fragments were constructed. These fragments were clustered and two hundred candidate fragments were randomly selected for each residue position. The selected fragments were independently used as search model in molecular replacement. The fragments were assembled together after molecular replacement. To reassemble, one fragment was selected as a seed fragment and one low-energy *de novo* model was taken as a reference model. The reference model was superposed to the seed fragment. Using the seed fragment and the reference model, position and orientation of other fragments were determined in the crystallographic unit cell and partial model was obtained. The combinations of permissible origins and symmetry operators of space group with unit cell translation were computed to identify the location of other fragments. The combination that gave the smallest distant between the reference model and the candidate fragment was taken as the correct location. In this way, all the fragments were reassembled in the asymmetric unit. This method was tested in ten difficult proteins with three different fragments – thirteen-residue, seventeen-residue and twenty-one-residue. Ten targets were considered as difficult because the best predicted full-length models of these targets, which showed average CA-RMSD 3.97 Å, were unable to provide the phase angles after molecular replacement experiment. The crystal structures of eight protein targets were solved from a total of ten using seventeen-residue fragment and their average CA-RMSD is 1.25 Å.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Kam Y. J. Zhang. It has been an honor to be his first Ph.D. student at The University of Tokyo and RIKEN. He has provided me a fantastic academic and research environments where I have had the chance to develop logical thinking, creativity, research skills, and to become an independent and collaborative research professional. I appreciate all his efforts to make my PhD study productive and enjoyable.

I would also like to thank Professor Masahiro Kasahara for stimulating discussions about programming. The discussion with him about programming was very fruitful for the study. I am also grateful to Professor Kasahara for being the jury member of thesis evaluation committee. I would also like to thank Professor Yutaka Suzuki and Professor Koji Tsuda from The University of Tokyo for being the judge for thesis evaluation committee. Similarly, I would like to thank Professor Min Yao from Hokkaido University for being the judge as external referee of my PhD thesis committee.

I like to thank all of my co-workers, Dr. Asuhtosh Kumar, Dr. Arnout Voet, Dr. David Simoncini, Dr. Muhammad Muddassar, Dr Kamlesh Sahu, Dr. Taeho Jo, Dr. Yong Zhou, and Dr. Ryo Takahashi, Mr. Francois Berenger and Ms. Xiao Yin Lee, for professional and personal supports. Their supports have made the PhD study enjoyable and interesting. I would also like to thank the secretary Ms. Hiroko Kani for her support in many aspects of life in Japan.

I gratefully acknowledge the RIKEN, Japan for many things. First, RIKEN funded me for three years to study PhD. The financial support, International Program Associate (IPA), provided from RIKEN was tremendous to spend good life in Japan during the PhD study. Without support from RIKEN, I would not have reached to write this PhD dissertation. I would also thank Graduate School of Frontier Sciences, The University of Tokyo for different research grants. RIKEN has also provided highly sophisticated facilities required for the research from workstation to supercomputer. I appreciate the supercomputing power provided by RIKEN Integrated Cluster of Clusters and would acknowledge Advance Center of Computing and Communication, RIKEN. All experiments I have presented in this thesis were carried out at RIKEN Integrated Cluster of Clusters.

I appreciate the open source community that has freely provided source code written in different programming languages that saved my time and effort tremendously. Especially, I would like to thank the researchers and developers of Rosetta software team from University of Washington, Phaser program group from University of Cambridge, and Kevin Cowtan developer of clipper from University of York.

Lastly, I sincerely thank my family for their all time supports, love, and encouragement. I am grateful to my parents (Min Bahadur Shrestha and Dropati Shrestha) who raised me with a love of science and supported me in all my pursuits. I thank my sister, Roj, and brother, Ujjan, for all their supports. Finally, I appreciate my wife, Shalu, for her love, supports, and encouragements during the period of this PhD. Thank you!

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Protein and its structure

Proteins are macromolecules performing numerous biochemical functions in the living cell. Protein regulates DNA transcription together with ribonucleic acids (Kornberg, 1974), maintains the integrity of genomic information (van Gent, et al., 2001), performs enzymatic reactions in metabolic pathways (Desnick and Schuchman, 2002), synthesizes and degrades other proteins (Glotzer, et al., 1991), metabolize xenobiotic (Geffeney, et al., 2002). These varieties of functions are performed by specific sequence of proteins that contain different amino acids from the twenty natural amino acids. These amino acids in proteins are linearly connected through covalent bond formed between carboxyl and amide groups of amino acids. The bond is known as peptide bond and the linear chain of amino acids is termed as primary structure. Primary structure starts from N-terminus and ends at C-terminus. Each amino acid contains the functional group, which is known as side-chain that determines the property of amino acids. Primary structure determines the three-

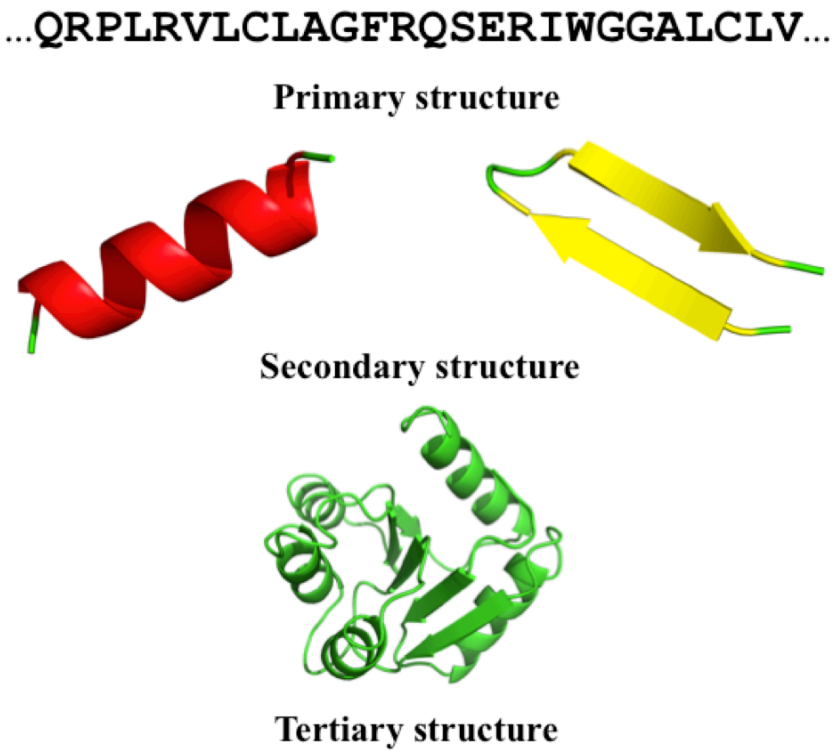

Figure 1.1 Different level of protein structure

dimensional structures (3D) or tertiary structure of proteins according to protein folding principle (Anfinsen, 1972). Secondary structure elements alpha-helices and beta-strands, are formed locally that are generally form first during protein folding process (Pauling, et al., 1951). Subsequently, the spatial arrangement of different secondary structure elements determines the tertiary structure of proteins following biophysical principle (Anfinsen, 1972). The arrangement of number of folded polypeptide chains, which also referred as subunits, defines the quaternary structure. These subunits in quaternary structures associate through non-covalent interaction (Jones and Thornton, 1996) and, in some cases, disulfide bonds (Sela and Lifson, 1959).

The geometry of tertiary structure of proteins is defined using bond length, bond angle, and torsion or dihedral angles. Bond length and bond angle require two and three atoms to compute. Torsion angle needs four consecutive atoms and it is the angle between two normal vectors of the planes. The sequence of three torsion angles defines the backbone conformation of protein. These three torsion angles, phi (ϕ), psi (φ), and omega (ω), are the only degree of freedom for the polypeptide backbone conformation. Certain combinations of ϕ and φ are only allowed in backbone conformation because of the strong repulsive van der Waals interaction. The

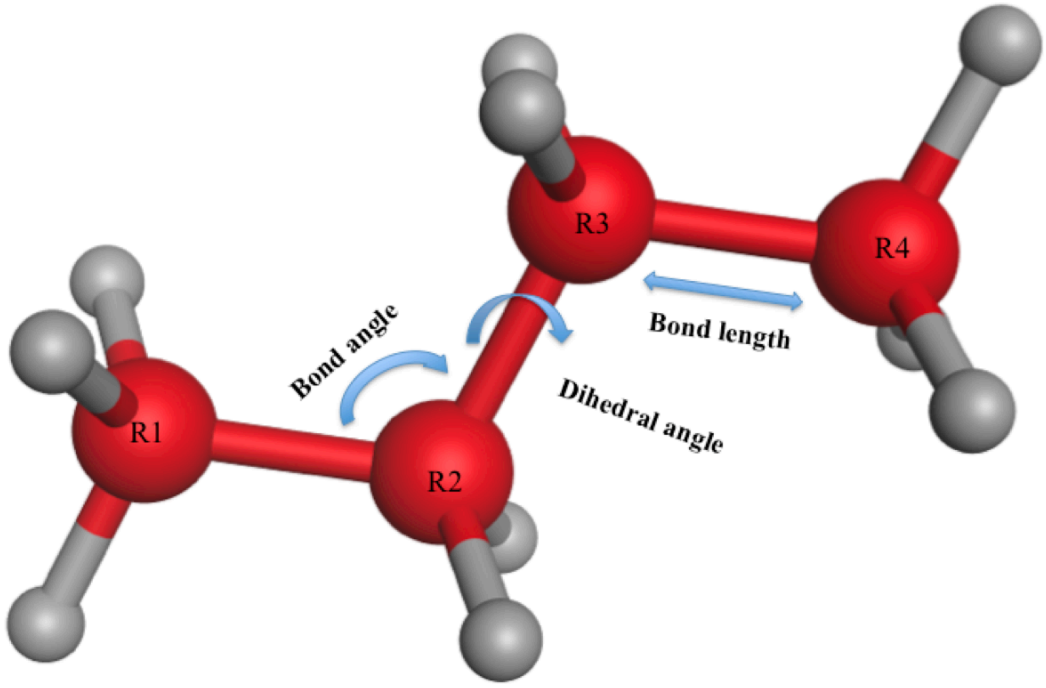

Figure 1.2 Bond length, bond angle, and dihedral angle

favorable backbone torsion angle distribution of ϕ and φ has been shown in Ramachandran plot (Ramachandran, et al., 1963). The planar peptide bond is highly restrained to ω angles and the ω angles are around 180° for trans- and 0° for cis-peptides (Weiss and Hilgenfeld, 1999).

The folded proteins contain energetically stable torsion angles for local structures forming the regular pattern that is the secondary structures (Eisenberg, 2003). The secondary structures are helices (Pauling, et al., 1951), sheets (Pauling and Corey, 1951), and turns (Rose, et al., 1985). These local structures are frequently appeared and optimally satisfy the torsion angle restraints and hydrogen bonding patterns simultaneously. Carbonyl oxygen atoms as acceptors and nitrogen atoms of subsequent residue as donors form hydrogen bond in backbone conformation of alpha helices and turns. However, the formation of hydrogen bond in beta strands is different. The beta stands were also formed by hydrogen bonds, which also determine orientation of beta sheet. It can be parallel, anti-parallel, or mixed of parallel and anti-parallel. In addition to hydrogen bonding in backbone conformation, side-chain interactions also play an important role in the protein folding and for other intra-atomic interactions. Weak non-covalent interactions, electrostatic, van der Waals, and non-polar, play principle role in the protein folding.

1.2. Computational methods for protein structure prediction

Computational protein structure prediction method mainly includes comparative modeling (Blundell, et al., 1987; Marti-Renom, et al., 2000; Sanchez and Sali, 1997) and *de novo* or *ab initio* approaches (Bradley, et al., 2005). Comparative modeling requires the sequence similarity of target sequence with at least one known structure but *de novo* or *ab initio* structure prediction approach is free from this constraint.

Comparative modeling builds the three-dimensional model for target sequence using the known structures as the template model on the basis of sequence similarity between the target sequence and the sequences of known structures (Blundell, et al., 1987; Marti-Renom, et al., 2000). Comparative modeling needs a two major steps to build the final model (Baker and Sali, 2001). A first step requires searching of suitable template. The correct templates can be found by sequence alignment methods, such as PSI-BLAST(Altschul, et al., 1997) , or by threading or

fold recognition methods (Bowie, et al., 1991; Jones, et al., 1992). Threading methods incorporates structural information in addition to the knowledge from sequences to assess the sequence-structure relationship therefore sometimes it can reveal more distantly related proteins which cannot be detected using sequence comparison methods alone. Comparative modeling builds final models of target sequence using one or more protein structures based on sequence alignment in the second step. The widely used methods for comparative modeling are based on rigid body assembly (Blundell, et al., 1987), segment matching (Levitt, 1992), and satisfaction of spatial restraints using either by distance geometry or real-space optimization (Sali and Blundell, 1993). In recent year, many robust and accurate computational methods have been developed for comparative modeling such as Modeller (Sali and Blundell, 1993), I-TASSER (Roy, et al., 2010), RosettaCM (Song, et al., 2013), HHPred (Soding, et al., 2005) and other methods recombine multiple templates.

The 3D structures of the protein sequences are often likely to be at global free-energy minima (Fleishman and Baker, 2012; Lazaridis, et al., 1995) with few major exceptions(Sohl, et al., 1998). The *de novo* protein structure prediction problem searches the vast number of conformations to find the lowest free-energy structure for a given amino acid sequence. Therefore, The key components are the methods responsible for conformational space searching (Levintha.C, 1968) and the evaluation of each conformation accurately using energy function (Bradley, et al., 2005). The conformational space is astronomically large due to the degrees of freedom associated with each residue, which creates the challenge to develop the efficient method for searching the conformational space. Another challenge in *de novo* structure prediction is to develop an accurate energy function to calculate the protein conformation in the solvent. Despite these challenges, the *de novo* structure prediction has succeeded to generate the accurate models for small and single domain proteins (Bradley, et al., 2005). Many research works have focused to develop the efficient methods for conformational search (Liwo, et al., 2008) and the accurate free energy function (Bradley, et al., 2005; Fleishman and Baker, 2012).

Fragment assembly is an effective, practical and efficient approach for *de novo* structure prediction. The method utilizes the fragments from experimentally determined structures in order to reduce the conformation space to be searched. In this approach, the target sequence is broken into small and overlapping fragments. The

similar fragment sequences are searched in the Protein Data Bank (PDB) to identify the known substructures, which are then assembled into full-length tertiary structures under the guidance of energy function. This energy function is underlying on the thermodynamic hypothesis (Lazaridis and Karplus, 2000). The effective energy functions has been derived from physics-based (Brooks, et al., 1983) and knowledge based (Simons, et al., 1999) potentials. Hydrogen bonding (Mirsky and Pauling, 1936), van der Waals interactions, backbone angle preferences, electrostatic interactions, hydrophobic interactions, and chain entropy (Dill, 1990; Dill and MacCallum, 2012) are the principle components of the energy function (Chothia, 1984) .

Fragment assembly has been practically implemented in the Rosetta protein structure prediction program (Rohl, et al., 2004; Simons, et al., 1997) although this concept was initially proposed in the study by Bowie and Eisenberg (Bowie and Eisenberg, 1994). They initially assembled the nine-residue fragments to construct the tertiary structures (Bowie and Eisenberg, 1994). Subsequently, Baker and co-workers has matured the program and implemented in Rosetta program suite (Rohl, et al., 2004).

Rosetta has started the protein structure prediction simulation using the amino acids sequence and the constant-length fragments queried from the PDB. The fragments are overlapping and size of three and nine residues. These are called as three-residue and nine-residue fragments. It divides entire simulation into two stages, coarse-grained sampling and all-atom refinement, to generate the full model from amino acids sequence. These two fragments of the protein chain are assembled; native state of the protein occurs when these fragments are oriented such that low free energy interactions are made throughout the protein (Rohl, et al., 2004; Simons, et al., 1997). The coarse-grained sampling involves the rapid and efficient searching of conformational space with backbone atoms and centroid of side-chain atoms. The conformation generated at this stage is called as coarse-grained model. Therefore, energy functions must include terms that reflect the averaged-out effects of the omitted atoms and solvent molecules (Baker and Sali, 2001). Due to the large errors introduced by the missing atoms in the true free energy, coarse-grained sampling locates a large number of local minima while searching the global minima. Each coarse-grained model is packed with side-chain atoms followed by energy

minimization in all-atom refinement. The all-atom refinement uses the realistic all-atom physics-based force fields with Metropolis Monte Carlo (Li and Scheraga, 1987) to optimize the models (Bradley, et al., 2005). The all-atom forces used in all-atom energy function consist with short-range interactions such as van der Waals packing, hydrogen bonding (Kendrew, et al., 1960; Perutz, et al., 1960), and desolvation (Tsai, et al., 2003).

Rosetta has been shown to be one of the best performing methods for *de novo* structure prediction (Bradley, et al., 2005; Das, et al., 2007) in Critical Assessment of Techniques for Protein Structure Prediction (Karplus, et al., 2003) competitions (Moult, et al., 2014) although many approaches have been developed for structure prediction (Fujitsuka, et al., 2006; Hamelryck, et al., 2006; Jones and McGuffin, 2003; Karplus, et al., 2003; Lee, et al., 2004). *De novo* models predicted using Rosetta have been used for practical utility such as in solving the crystallographic phase problem (Das and Baker, 2009; Qian, et al., 2007) and protein design (Kuhlman, et al., 2003). Similarly, *de novo* protein structure prediction methodology is also used for nuclear magnetic resonance structure refinement to improve the phasing power by moving it closer to its X-ray crystal structure counterpart (Mao, et al., 2011; Ramelot, et al., 2009).

Instead of constant-length fragments, another *de novo* structure prediction program, Quark (Xu, et al., 2011; Xu and Zhang, 2013), a top performer in recent CASPs, has been introduced. Global models are generated by assembling the variable-length continuous fragments of different sizes from 1 to 20 residues with replica-exchange Monte Carlo simulation. These fragments are assembled to generate the full models, which are guided by a composite knowledge-based force field. These are semi-reduced models that contain the backbone atoms and center of mass of side-chain atoms. The representative semi-reduced models from the top-five largest clusters (Zhang and Skolnick, 2004) are selected and sent for packing of side chain atoms. These atoms are packed in the models using ModRefiner and the final models were minimized with physics-based energies (Xu, et al., 2011).

Protein structure provides meaningful insights about how atomic interactions occur in the molecules. Therefore, protein structures are highly demanding among chemists and biologists. Computationally predicted protein structures are adequate to understand many biological functions when experiment information are combined.

The utility of proteins structure depends on its accuracy (Baker and Sali, 2001). High-quality protein structures with atomic resolution are necessary in understanding catalytic mechanism (Barford, et al., 1998; Rajagopalan, et al., 2014; Trievel, et al., 2002), designing and improving ligands (Blundell, et al., 1987; Procko, et al., 2014), selection of ligands in drug designing using virtual screening(Blundell, et al., 2002; Blundell and Patel, 2004; Carvalho, et al., 2009), docking of macromolecules (Strynadka, et al., 1996), understanding protein-protein interactions (Zhang, et al., 2012) and designing novel proteins including enzymes and vaccines (Correia, et al., 2014; Kuhlman, et al., 2003). Protein structures are also useful in structure determination such as solving crystallographic phase problems using MR (DiMaio, et al., 2011; Qian, et al., 2007), refining NMR structures (Mao, et al., 2011; Mao, et al., 2014; Ramelot, et al., 2009), interpreting low-resolution electron density map(Schroder, et al., 2010), and structure from sparse experimental restraints (Thompson, et al., 2012).

1.3. X-ray crystallography for protein structure determination

X-ray crystallography is a principle method in the study of biological systems. It provides atomic resolution information to understand the fundamentals of life. The structure of the double helix of deoxyribose nucleic acid (Watson and Crick, 1953) and the high-resolution structure of eukaryotic 80S ribosome (Yusupova and Yusupov, 2014) were solved using X-ray crystallography. This method has also become central to the development of new therapeutics for human disease (Blundell and Patel, 2004; Carvalho, et al., 2009; Rowland, 2002). The technique has become robust since Kendrew and Perutz solved the structures of myoglobin (Kendrew, et al., 1960) and hemoglobin(Perutz, et al., 1960). As a result, more than 100,000 protein structures to date have been deposited into the PDB (Berman, et al., 2002). The consistent advancement in the technology for protein production, crystallization, data collection, and data analysis increases the remarkable success in macromolecule structure determination. In the last decade, the success in technology development has been achieved by worldwide structural genomics efforts (Collins, et al., 2003; Collins, et al., 1998; Joachimiak, 2009; Ueno, et al., 2006). In addition, the advancement in hardware and software for crystallographic data collection, structure determination, refinement, bioinformatics (tools and databases), robotics and automation improved and accelerated the many processes in structure determination. Currently available

computational methods for the analysis of diffraction data (Adams, et al., 2009) have improved the crystallographic process and also introduced automation. An algorithm for phasing with molecular replacement (MR) (Rossmann and Blow, 1962) becomes robust and improves the generation of structure factor phases using maximum likelihood (McCoy, et al., 2007). Automated model building methods have reduced manual efforts to generate the initial models for many crystallographic projects and also work at higher and lower resolution limits (Langer, et al., 2008; Terwilliger, et al., 2008). Atomic models after automatic or manual methods must be further optimized to best fit the experimental diffraction data and prior chemical information. Because an initial model is often incomplete, refinement is iteratively carried out to improve the phases that can then be used to obtain a more accurate electron density map (Afonine, et al., 2012; Murshudov, et al., 1997). These refinement programs optimize models with diffraction data even when only low-resolution (lower than 3Å) data are available. The refined models are validated to detect errors in the models before the deposition. This process has improved to a point that many errors in models are readily detectible and can be corrected early (Chen, et al., 2010).

1.4. Phase problem

The crystallographic experiment aims to obtain a three-dimensional map of the electron density in the macromolecular crystal. Fourier synthesis using complex numbers derived from the diffraction experiment computes the distribution of electron density in the crystal. Each complex number contains the amplitude and an associated phase angles. However, diffraction experiment measures the amplitude but cannot obtain the phase angles. Many methods have been developed to obtain phases that include experimental and computational methods (Adams, et al., 2013). Mathematically, the electron density in a crystal can be obtained by calculating the Fourier summation:

$$\rho(x\ y\ z) = \frac{1}{V}\sum_{h\ k\ l} |F(h\ k\ l)| \exp[-2\pi i(hx + ky + lz) + i\alpha(h\ k\ l)] \qquad \ldots (1.1)$$

where, $|F(h\ k\ l)|$ is the structure factor amplitude of reflection ($h\ k\ l$) including the temperature factor, and $\alpha(h\ k\ l)$ is the phase angle. x, y, and z are coordinates in the unit cell. The amplitude $|F(h\ k\ l)|$ can be obtained. However, the phase angles $\alpha(h\ k\ l)$ are not available.

The isomorphous replacement and anomalous scattering methods are the experimental ways to solve the phase problem. These methods obtain phases using information derived from small differences between diffraction datasets. Both methods located the place of the heavy atoms or anomalous scatters in the crystallographic asymmetric unit (Adams, et al., 2009). Perutz, for the first time, successfully applied the multiple isomorphous replacement method to solve the protein structure of hemoglobin (Perutz, et al., 1960). In this method, diffraction pattern of the target protein crystal is compared with that of a crystal that contains at least one heavy atom. Apart from attached heavy atoms, other parameters are same for both crystals. The intensity differences between native and other patterns are mainly due to attached heavy atoms. The attached heavy atoms played the roles for determining the position of other heavy atoms (delaFortelle and Bricogne, 1997; Terwilliger and Berendzen, 1999). This can be done either manually or by an automatic Patterson search procedure. This experiment discarded anomalous scattering effect. However, current experimental methods depend on anomalous scattering alone - in the form of multi-wavelength anomalous diffraction(delaFortelle and Bricogne, 1997) and single-wavelength anomalous diffraction (McCoy, et al., 2004; Wayne A. Hendrickson, 1981) because anomalous scattering is sensitive to the X-ray wavelength.

The difference in intensity between Bijvoet pairs due to the anomalous scatters can be exploited for the phase angle determination in the proteins. The reflections (h, k, l) and (-h, -k, -l) are called Bijvoet pairs and intensities of these two reflections are equal that rise a center of symmetry in the diffraction pattern. Here, h, k, and l are reflection indexes in the reciprocal space. The wavelength dependence of the anomalous scattering is used in the multiple-wavelength method. Therefore, protein should contain an element that gives a strong anomalous signal. The presence of selenium atoms in protein is sufficient for successful structure determination using multiple-wavelength anomalous diffraction (Hendrickson, et al., 1990; Leahy, et al., 1992). The method that stably and reproducibly incorporates intrinsic anomalous scatters by replacing methionine residues with selenomethionine (Hendrickson, et al., 1990) has been widely used in recent years. The multi-wavelength anomalous dispersion faces the disadvantages such as the collection of data at multiple wavelengths, the long exposure time, and danger of radiation damage to the crystal.

These pitfalls are less sensitive when the crystal structure can be solved by the data collection on a single crystal with one wavelength only. Using single wavelength anomalous dispersion for structure determination, the crystal must contain anomalous scatters that provide strong anomalous signal. The application of maximum likelihood methods to multi-wavelength anomalous diffraction (delaFortelle and Bricogne, 1997) and single-wavelength anomalous diffraction phasing (McCoy, et al., 2004) exploits small phasing signals robustly. Finally, density modification methods significantly improve the weak phase information obtained from anomalous scattering methods.

MR is a widely used computational method to solve the phase problem. This method needs a homologous structure, which is already known, to solve the unknown structures. Almost two-thirds of protein structures deposited in the PDB are solved using MR (Long, et al., 2008). Although the database of known structures in the PDB grows, the number of new folds reduces and the proportion of structures solved by MR increases. In this method, the phases are calculated from a similar structure placed in the position of the unknown molecule in the crystallographic unit cell. Placement of the molecule in the target unit cell requires its proper orientation and precise position. This step involves rotation and translation searches in the unit cell. Therefore, MR requires the six degrees of freedom for searching in the unit cell. The spatial orientation of known and unknown structure with respect to each other is determined by rotation function whereas translation vector finds correct position of correctly oriented structure. Since known and unknown structures are in Cartesian and Fourier space, rotation and translation functions cannot be computed in a straightforward way.

MR has used the Patterson map to identify the orientation of molecule i.e. rotation function. Patterson map is the vector map that is calculated from distance between atoms. The self-Patterson peak all lie in a volume around the origin with a radius equal to the dimension of the molecule. Two different Patterson maps must be superimposed to maximum overlap by a rotation of one of the two maps. When a number of identical molecules lie within one asymmetric unit, the self-Patterson vector distribution is exactly the same for all of these molecules, except for a rotation that is the same as their non-crystallographic rotational symmetry in real space. After correct rotation, the translation function determines the translation vector required to overlap one molecule onto the other in the real space. The known molecule is

translated through the asymmetric unit. Structure factors are calculated and compared with the observed structure by calculating and R-factor. R-factor is calculated:

$$R = \frac{\sum \left| |F_{obs}| - |F_{cal}| \right|}{\sum |F_{obs}|} \qquad \ldots (1.2)$$

In the equation, F_{obs} is structure factor observed from diffraction data and F_{cal} is calculated structure factors using model. The rise of MR is mainly due to improvement in methodology for rotation and translation functions (Kissinger, et al., 1999; McCoy, et al., 2007; Navaza, 2001). MR was implemented with six-dimensional search using evolutionary approach (Kissinger, et al., 1999) whereas the performance of the search is better in three-dimensional rotation search followed by translation (McCoy, et al., 2007; Navaza, 2001). Mathematically elegant fast rotation function was first introduced to replace the conventional procedure. Advanced procedure is the introduction of maximum likelihood targets for MR(McCoy, et al., 2007) that has increased the signal in MR searches, thus allowing structures to be solved with more distant homologs. These target functions also exploit the information from partial solutions, which improves success in solving structures of complexes or crystals containing multiple copies. Automation is another key for increasing the utility of MR. Here is an example for automation implemented in Phaser. The ability to test multiple models for multiple choices of possible space group allows problems to be solved without the manual intervention. MR pipelines extend the power of automation even further, by testing alternative approaches for model preparation (Claude, et al., 2004; Keegan and Winn, 2007) or building up models using automated domain databases (Long, et al., 2008).

MR becomes more difficult for targets with low-homology templates (or with no identifiable homologs). Therefore, MR provides a highly stringent and practical challenge in structure modeling. Phasing experiments have been carried out with the models generated in CASP experiments (Giorgetti, et al., 2005; MacCallum, et al., 2011) to assess whether models have achieved a high enough accuracy to be practically useful. However, successful MR solutions are for very few cases that are high symmetry molecules (Kratzner, et al., 2005; Szep, et al., 2003). Furthermore, homologous model refined with all-atom energy used for *de novo* protein structure prediction has achieved the solution in MR (Qian, et al., 2007). The all-atom

refinement has also worked successful to improve the NMR models for MR trials (Qian, et al., 2007). More recently, using the sampling methods and force field from *de novo* modeling has solved the difficult MR targets. The iterative process of conformation sampling with force field used in protein modeling have improved the phases in the density map obtained from the ambiguous MR solutions (DiMaio, et al., 2011).

Likelihood target functions in MR have increased sensitivity for MR searches and thus have found the correct position and orientation of smaller fragments, as small as single helices. ARCIMBOLDO (Rodriguez, et al., 2009), a computer program, has especially placed the alpha-helices fragments in the correct position and orientation and with the help of automated model building program, it has also built the complete structure started from a few helices. Furthermore, generalized algorithm has been recently developed to solve the structure starting from the fragments (Sammito, et al., 2013). This algorithm is not only limited to the alpha helical fragments and also works with other secondary structure elements. The program identifies the suitable fragments, places the fragments and subsequently constructs the complete model.

1.5. *Ab initio* phasing with *de novo* models

MR cannot be used if the suitable search model for target sequence is not available. However, the advancement in protein structure modeling with sequence information only has provided the novel view to employ the *de novo* model as the search model. Therefore, *ab initio* phasing with *de novo* model becomes recently emerging and challenging problem in protein crystallography. Recent progress in *de novo* protein structure prediction has generated highly accurate *de novo* models (Bradley, et al., 2005; Kuhlman, et al., 2003). These high-quality models predicted using only amino acid sequences has created new possibilities for *ab initio* phasing using MR (Qian, et al., 2007). This approach is called as '*ab initio* phasing with *de novo* models'.

Initially, *ab initio* phasing techniques showed some success for targets with simple folds of high symmetry in the cases where structurally similar experimental models were not available (Strop, et al., 2007). Because of the generation of high

quality models using amino acids sequence, the *ab initio* phasing was expanded with *de novo* models.

Ab initio phasing with *de novo* models begins with generation of pools of three-dimension structures of the given amino acids sequence without providing any experimental constraints. Many programs (Rohl, et al., 2004; Xu and Zhang, 2012) have been developed for generation of high-quality 3D structures of small-sized proteins using amino acids sequence. In *de novo* prediction, the models are minimized under the guidance of all-atom energy to identify the global minima. The models with low-energy are considered as the most accurate in the absence of native structure and selected as the represented models for different purposes. Sometimes, clustering method is also useful for model selection because of ruggedness in the energy landscape. Since a large number of low-energy conformations surrounding the correct fold than low-energy incorrect folds in randomly sampled energy landscape, clustering can select the most accurate models. Therefore, the representative models are generally selected from the pool of generated models either using all-atom energies (Bradley, et al., 2005) or after clustering of the models (Shortle, et al., 1998; Zhang and Skolnick, 2004). Similarly, the representative models selected either of using these methods are used as template for MR for solving phase problem.

A *de novo* model generated by Rosetta for a set of proteins were successfully phased using MR (Das and Baker, 2009). In this study, a set of diffraction data was phased with the represented models that were selected using all-atom energy after generation of full models. The success rate in MR trial was increased when all-atom models were used instead of coarse-grained models and huge computational power was spent to search the larger conformational space. This study has identified the reasons for the difficulty in MR using *de novo* models as template models. Model accuracy and computational time has appeared as the primary bottleneck. The computational time was significantly reduced by incorporating the MR program, Phaser, into the structure prediction program, Rosetta (Shrestha, et al., 2011). This procedure increased the success rate of phasing as well as efficiently managed the computation time required for phasing. Conformations generated at each trajectory were phased at most five times in the course of all-atom minimization. The models, which were very bad and very good at first trial of phasing, were escaped from time-intensive refinement procedure. This procedure saves huge computing time without

degrading the success rate. Similar procedure was also employed to the top-ranked models produced by Foldit game in all-atom refinement for phasing (Khatib, et al., 2011). Recently, the ensemble of selected models was used differently as the MR template (Bibby, et al., 2014). The models were selected after clustering and then errors at local regions were estimated. These local regions that contain errors were truncated from the models before used in MR as the search model (Bibby, et al., 2014).

Chapter 2. Objective of the study

The most critical factor to achieve the success in *ab initio* phasing using *de novo* models was the accuracy of search model (Das and Baker, 2009; Shrestha, et al., 2011). The accuracy improvement is also a challenging task in *de novo* modeling due to insufficient conformational sampling, inaccurate energy function, and lack of better quality fragments generation. The improvement in these areas would have increased the accuracy of prediction. Therefore, I was interested to carry out the research work focusing on the development of efficient conformational sampling method and better quality fragment generation to improve the accuracy of predicted models in my PhD study. In addition, when model accuracy was difficult to improve to the quality required for crystallographic phasing, the substructures from the predicted models were used to solve the phase problem.

The algorithm was developed to employ the information obtained from the ensembles of low-energy models to improve the sampling strategy. In this procedure, coarse-grained models were first generated using fragment assembly method implemented in Rosetta. Second, the ensemble of these models was selected to find the inaccurately predicted residues or regions and then conformation sampling was biased so that inaccurate regions were frequently sampled using short fragments (three-residue). Third, these residues or regions were rebuilt to improve the overall accuracy of these models. The rebuilt models were first converted into all-atom models and further optimized using the all-atom energy function. This algorithm was implemented in MORPHEUS (Shrestha, et al., 2012) and increased the success rate of difficult targets (Das and Baker, 2009) of small-sized globular proteins.

Second, the algorithm was purposed to improve the fragment quality. The information exploited from the ensembles of low-energy all-atom models were employed to generate the new fragments that were most likely adopt in native structures. The major goal was to gather a diverse set of fragments that contributes to generate low-energy models. These fragments were used to generate high quality all-atom models. The selected models were broken down into continuous nine-residue and three-residue fragments. Distant fragments were removed using the score computed using residues in ensemble of models, which is known as average pairwise residue distance score, and then clustering was performed in the resultant fragments.

Representative fragments were taken from top five clusters, which were used to generate new set of models.

When full-length *de novo* models necessary for phasing cannot be predicted accurately, *ab initio* phasing using *de novo* models is impossible. Instead of using global models as the search models, local substructures (fragments) can be considered as search models for a MR program, Phaser. The maximum likelihood target functions used in Phaser increases the sensitivity of searching; smaller fragments can be used to locate at correct place in asymmetric unit. In this algorithm, the *de novo* models were broken into numerous constant-length overlapping fragments. The representative fragments for each residue position were chosen after the fragments were clustered. The selected fragments were independently given as the search model for Phaser. Since all the fragments are independently phased and scattered at different locations of crystallographic space, the goal is to assemble them in the same asymmetric unit making same reference point. Indeed, they are interrelated by permissible origins, crystallographic symmetry, and unit cell translation. These fragments were assembled together using a real-space strategy although it can be assembled using a reciprocal-space method and it is computationally challenging due to many combinations. The assembly procedure begins with selection of one of the phased fragments that is termed as seed fragment and low-energy model as reference model. For other fragments, the positions were searched using permissible origins, crystallographic symmetry, and unit cell translation and picked the position that gives the minimum distant to the reference model and seed fragment. The assembled fragments built the partial models and reduces error existed in the full-length model.

Chapter 3. MORPHEUS – error-estimation-guided rebuilding of *de novo* models increases the success rate of *ab initio* phasing

3.1. Objective

X-ray crystallography is the principal method for the structure determination of macromolecules, including proteins, to atomic detail. Protein structures have been alternatively solved by computational methods such as MR (Rossmann and Blow, 1962). It requires template models derived from the structures of homologous proteins so that it is impossible to obtain the phases without at least one homologous protein of target sequence. However, recent improvements in computational methods for the prediction of protein structures using only amino acid sequences, known as *de novo* modeling, have opened a new frontier in structure determination. One of the practical applications of these computationally predicted *de novo* models has been shown to be the search model for the solution of the crystallographic phase problem for new folds (Qian, et al., 2007), which can be considered as *ab initio* phasing. These models have extended the utility of MR in the absence of known starting homologous structures.

A successful *de novo* modeling method was inspired by the fragment-assembly approach, in which fragments of known structures are combined under the guidance of scoring functions. The scoring functions combine the major energy terms for protein stability (Bowie and Eisenberg, 1994; Rohl, et al., 2004). Rosetta (Rohl et al., 2004) is the one of the most successful fragment-assembly methods for protein structure prediction. Rosetta has demonstrated the ability to predict the high-quality models necessary for solving the phase problem by MR(Qian, et al., 2007). Rosetta uses first a coarse-grained model that contains only the main-chain and the centroids of side-chain atoms for wider conformational space searching. Second, it refines the all-atom models derived from coarse-grained models with limited main-chain conformational searches and full side-chain packing and optimization. The success of all-atom refinement depends highly on the quality of the coarse-grained models generated in coarse-grained sampling.

The predicted model must have the correct fold as present in the target structure in order to be the successful search model for MR. Furthermore significant

portion of the atomic scatters should spatially match those of the underlying target structure. Many studies have accelerated the development in *ab initio* phasing with *de novo* models recent years. The MR method has generally been executed on selected *de novo* models after generating a large number of models (Das and Baker, 2008). Coarse-grained models or polyalanine models were also used for MR (Das and Baker, 2009). However, the absence of many atoms in the model appears as bottleneck to achieving successful solutions. The number of success in MR trials was significantly increased with all-atom models optimized using Rosetta all-atom energy (Das and Baker, 2009). The success rates were further increased when huge computing power spent on conformational sampling (Das and Baker, 2009) and the highly flexible loop regions in the predicted models were trimmed off (Bibby, et al., 2014). In addition, phasing with intermediate all-atom models during optimization managed the computational time as well as increased the success rate of MR experiments (Shrestha, et al., 2011).

Many crystallographic factors such as resolution, solvent content, non-crystallographic symmetry in the unit cell and others could have significant impact on obtaining successful MR solutions. However, these factors showed poor correlations with the MR success rate using *de novo* models (Shrestha, et al., 2011). In contrast, accurate models have enabled success in all tested cases (Das and Baker, 2009; Shrestha, et al., 2011). Therefore, highly accurate *de novo* models are necessary to predict for suitable search model for MR. The accuracy of *de novo* model can be improved in different ways such as the improved identification of high-quality fragments, more accurate energy functions, and more efficient sampling of conformational space.

The errors in a template *de novo* model are not uniformly distributed. Removing regions with large errors can produce a template that is closer to the target and increase the chances of success in MR provided that the remaining structure still constitutes a significantly large portion of the scattering matter with respect to the target. Many approaches have been traditionally employed to increase the success rate of MR from a given template, which is typically from a structural homologue. These include trimming off loops or terminal regions to create a compact core structure, removing side-chain atoms to generate a polyalanine model, deleting highly flexible regions identified by high temperature factors in the coordinates and pruning off the

side-chain atoms of residues that are non-conserved between the template sequence and the target sequence (Stein, 2008). Searching multiple domains or multiple templates simultaneously can also be very powerful in solving difficult cases of MR (McCoy et al., 2007). To take advantage of the ever-increasing number of structures that are being deposited in PDB (Berman et al., 2000), automated pipelines have been created in order to relieve users of the burden of the manual curation of templates for MR, resulting in an increased success rate (Keegan & Winn, 2008; Long et al., 2008). In this work, the focus is on improving the entire template model for MR.

One way to improve the *de novo* models accuracy is to identify the loop regions and then focus the conformational sampling on these regions (Canutescu and Dunbrack, 2003; Mandell, et al., 2009). The loops are identified from the secondary structure assignment of the predicted models (Kabsch and Sander, 1983). Many algorithms have been developed to carry out the extensive resampling of these loop regions. Although loop regions are often less accurately predicted, some loops are intrinsically disordered or can adopt multiple conformations. In this scenario, extensive conformational resampling in order to find one energetically most stable conformation may not be fruitful. Moreover, errors in predicted models exist not only in loop regions but also in regions of regular secondary structure.

One important step in improving the *de novo* model quality for phasing could be to initially identify the less accurately predicted regions in the model and then to perform rigorous sampling on these regions. There have been extensive efforts to develop methods that can assess the quality of computationally predicted models (Kryshtafovych & Fidelis, 2009). These model-quality assessment (MQA) methods have been shown to be very useful in identifying good-quality models and ranking them (Levitt & Gerstein, 1998; Zemla, 2003; Zhang & Skolnick, 2004b). Qian et al. (2007) used such a strategy to improve the success rate of MR by rebuilding the most variable regions within an ensemble of structural models. After identifying regions of high conformational variability using a principle similar to PCons (Wallner & Elofsson, 2006), an aggressive sampling was conducted on these regions and the cyclic coordinate-descent method (Canutescu & Dunbrack, 2003) was used to maintain the chain connectivity. The conformational variation has also been exploited as colony energy for loop prediction (Xiang et al., 2002). When the electron density map guided the rebuilding, the success rate of MR was further improved and many

challenging cases could be solved (DiMaio et al., 2011), although an approximate MR solution was required in this case.

Each step for improving the quality of *de novo* models is described. Firstly, this method identified the local regions or residues in the coarse-grained models with large errors. These errors were estimated by the average pairwise geometric distance per residue computed among selected lowest energy coarse-grained models. Secondly, this method rebuilt these more error-prone residues in the coarse-grained models. Lastly, these rebuilt coarse-grained models were converted into all-atom models and refined with Rosetta all-atom energy. These all-atom *de novo* models were used as the search model for the MR. Score used for error estimation and to guide the conformation sampling is similar to many MQA methods. However, per residue score is calculated in order to identify residues or regions where large errors exist instead of a global score of the entire protein model. More than 50% of the targets were tested that were not able to succeed in MR trials primarily owing to a lack of sufficiently accurate models in the previous study (Shrestha, et al., 2011). The results showed that the coarse-grained models were first rebuilt and then refined to closer to the native structures. Second, these models after all-atom refinement significantly increased the success rate of phasing.

3.2. Methods

This method aimed to reduce the distance between the coarse-grained models and the native structures. A geometric distance score for each residue of the selected coarse-grained models was calculated; Rosetta3.2 generated the coarse-grained models. Each residue was superposed to corresponding residue of remaining models from the selected pool and then a pairwise average root-mean-square deviation was calculated for C-alpha (CA) atoms. The score was termed as the average pairwise residue distance scores (APRDS) and defined as

$$APRDS(i,j) = \frac{1}{n-1}\sum_{k=1(k\neq j)}^{n-1}\left[\left(X_{ij}-X_{ik}\right)^2 + \left(Y_{ij}-Y_{ik}\right)^2 + \left(Z_{ij}-Z_{ik}\right)^2\right]^{\frac{1}{2}} \quad \ldots(3.1)$$

In the above equation, i represents the residue number, j represents the model number, k represents all of the other models except model j, n represents the total number of models and X, Y, Z represent the Cartesian coordinates of each CA atom in a residue. The least-squared method known as Kabsch algorithm (Kabsch, 1976) was used to compute the root mean square deviation for all decoys to native structure. This study

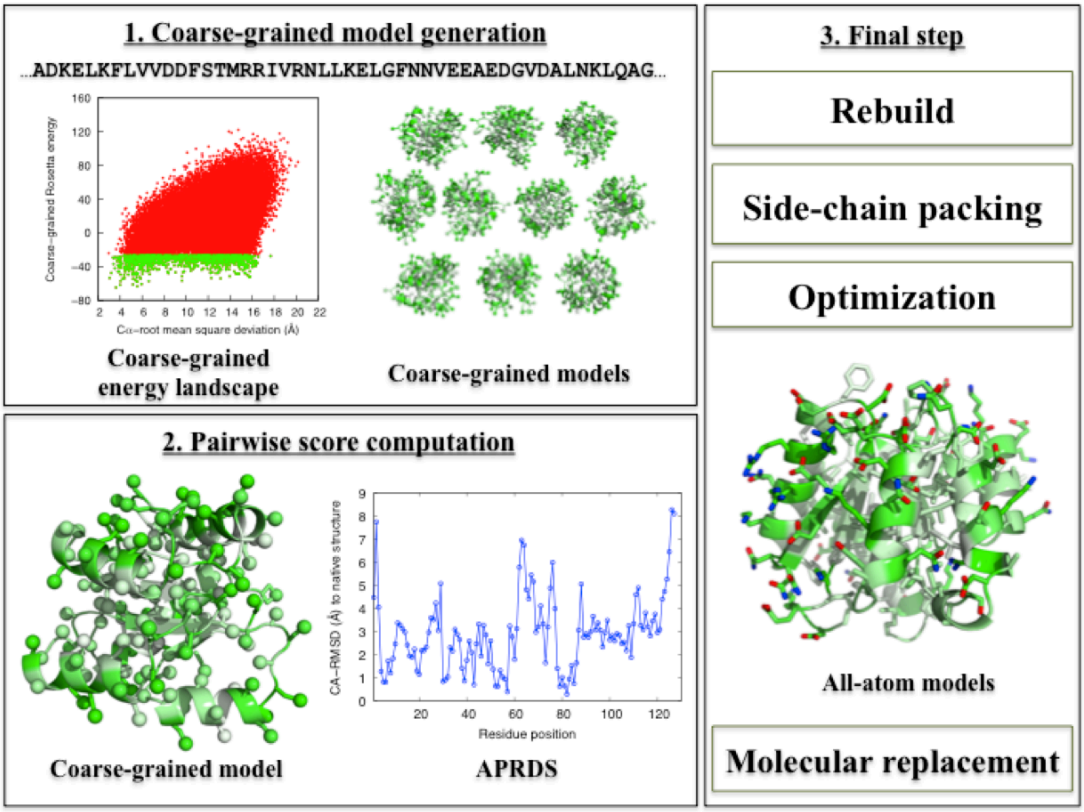

Figure 3.1 Schematic diagram of MORPHEUS program

used the root mean square deviation as the measurement metric because of its simplicity and generality although many different geometric distances were already existed for measurement, such as GDT (Zemla, 2003; Zemla, et al., 1999), MaxSub (Siew, et al., 2000), TM-score (Zhang and Skolnick, 2004), Q-score (Ben-David, et al., 2009) and percentile-based spread (Pozharski, 2010). APRDS was computed in the similar way to that implemented in PCons-local (Wallner and Elofsson, 2006). However, this score was used not only to identify residue errors but also to estimate the sampling frequency for each residue during rebuilding. The APRDS guided conformation space searching during rebuilding so that residues with higher scores are sampled more often than those with lower scores. The rebuilt coarse-grained models are converted to all-atom models using the Rosetta fast relax algorithm (Tyka, et al., 2011). All rebuilt models after all-atom refinement were tested with the diffraction data for their suitability as templates for solution of the phase problem using the Phaser program (McCoy, et al., 2007).

This approach was implemented using the C++ programming language. The program was developed using Rosetta and Phaser as libraries in the program and was

referred as MORPHEUS (**MO**del **R**ebuilding for **PH**asing with **E**nhanced s**U**cce**S**s) and schematic diagram is shown in Figure 3.1. This program has determined the continuation or termination of the simulation using the number of successful MR solutions. The lower and upper bounds for the Phaser score were used from the previous study (Shrestha, et al., 2011) in order to control the simulation. The program stopped the entire simulation once a few good models have been obtained with high confidence Phaser score. However, MORPHEUS used all *de novo* models requested for generation in the worst case for phasing.

3.2.1. Benchmark dataset and initial model generation

Ten difficult targets that were unproductive in RosettaX experiment (Shrestha, et al., 2011) were selected. Indeed, these targets were difficult cases for RosettaX approach. In addition, two more targets were also included, which were also solved using RosettaX program. Rosetta3.2 (Rohl, et al., 2004; Tyka, et al., 2011) generated 3.0E+05 initial coarse-grained models for each target sequence in the RIKEN Integrated Cluster of Clusters (RICC). Robetta server (Chivian, et al., 2003) generated two different types of fragments (nine-residue and three-residue). Fragments from the target structure and structures with homologous sequences were excluded from the fragment libraries in order to mimic a blind prediction.

3.2.2. Determine incorrectly predicted residues or regions

One thousand lowest energy coarse-grained models were selected from the pool of 3.0E+05 models generated by Rosetta. Kabash algorithm (Kabsch, 1976) was employed to superimpose each model with all other selected models using rigid-body transformation with an optimal translation vector and a rotation matrix that minimizes the sum of the squared distances between two coordinate sets of corresponding atoms (Kabsch, 1976). After optimal transformation, the APRDS was calculated by taking an average of the CA atom root mean square deviations (CA-RMSDs) computed between one model and all other models. This value was also assigned for each residue of the model. The correlation between the APRDS and the CA-RMSD from the native structure was calculated and used to assess the capability of the APRDS to estimate error in each residue. Furthermore, each coarse-grained model was assigned a score that was the average of the CA-RMSDs between this model and other models

covering the entire sequence, which was defined as the average pairwise model distance score (APMDS),

$$APMDS(i,j) = \frac{1}{n-1}\sum_{k=1(k\neq j)}^{n-1}\left[\frac{1}{m}\sum_{i=1}^{m}\left(X_{ij}-X_{ik}\right)^2+\left(Y_{ij}-Y_{ik}\right)^2+\left(Z_{ij}-Z_{ik}\right)^2\right]^{\frac{1}{2}} \quad \ldots (3.2)$$

where j represents the model number other than model k, i represents the residue number, m represents the total number of residues in the model and n represents the number of models. X, Y, Z represent the Cartesian coordinates of each c-alpha atom in a residue. The APMDS can be used to assess the overall quality of coarse-grained models.

3.2.3. Rebuilt inaccurately predicted residues

A subset of coarse-grained models, which were selected from the group of 1000, were allowed for further model rebuilding. APMDS score selected the coarse-grained models, which was 65% of the models, for the subsequent model rebuilding. Although APRDS and APMDS scores were calculated with a relatively large set of decoys, there is no need to subject all of these models to further rebuilding. This is because rebuilding only a subset of these models with the lowest APMDS scores will enable the inclusion of the majority of high-quality models, with a substantial saving of computational time.

The model selection criteria was difficult to optimize therefore it was hard to optimize the choices of selecting the 1000 lowest energy models for APRDS and APMDS calculation and the subsequent selection of 65% models for rebuilding. Instead, these parameters were empirically obtained by testing on the first target and they seemed to work well. Subsequently, they were used for all of the other targets. Owing to the extensive computing time needed to complete the calculation for the entire test set of targets, alternative choices cannot be exhaustively tested and compared in order to come up with an optimum combination of parameters.

The error-prone residues obtained the more frequency for conformational space searching according to the APRDS during the rebuilding process. Each rebuilding simulation has contained a total of 5000 rebuilding steps and these rebuilding steps were distributed to each residue based on its APRDS. Roulette-wheel procedure provided non-uniform sampling based on APRDS. Three-residue fragment

library was only provided as the source for rebuilding in order to reduce large changes in global conformations. Subsequent rebuilding run generated 300 trajectories for each selected coarse-grained model with different random seeds. The allowed trajectories were sufficient to explore the conformational space within a reasonable computational time. The models generated during each rebuilding trajectory were evaluated using the Rosetta coarse-grained scoring function. However, temperature factor was adjusted in the Monte Carlo simulated-annealing procedure to make the acceptance rate of high-energy models proportional to the residue error using the equation

$$T_{cur} = T_{min} + \frac{T_{max} - T_{min}}{D_{max} - D_{min}} (D_{cur} - D_{min}) \qquad \dots (3.3)$$

3.2.4. Molecular replacement with rebuilt models

The program converted each model from 300 independent rebuilding trajectories into an all-atom model using Rosetta all-atom refinement program. MORPHEUS employed the Rosetta fast relax algorithm (Tyka, et al., 2011) to pack

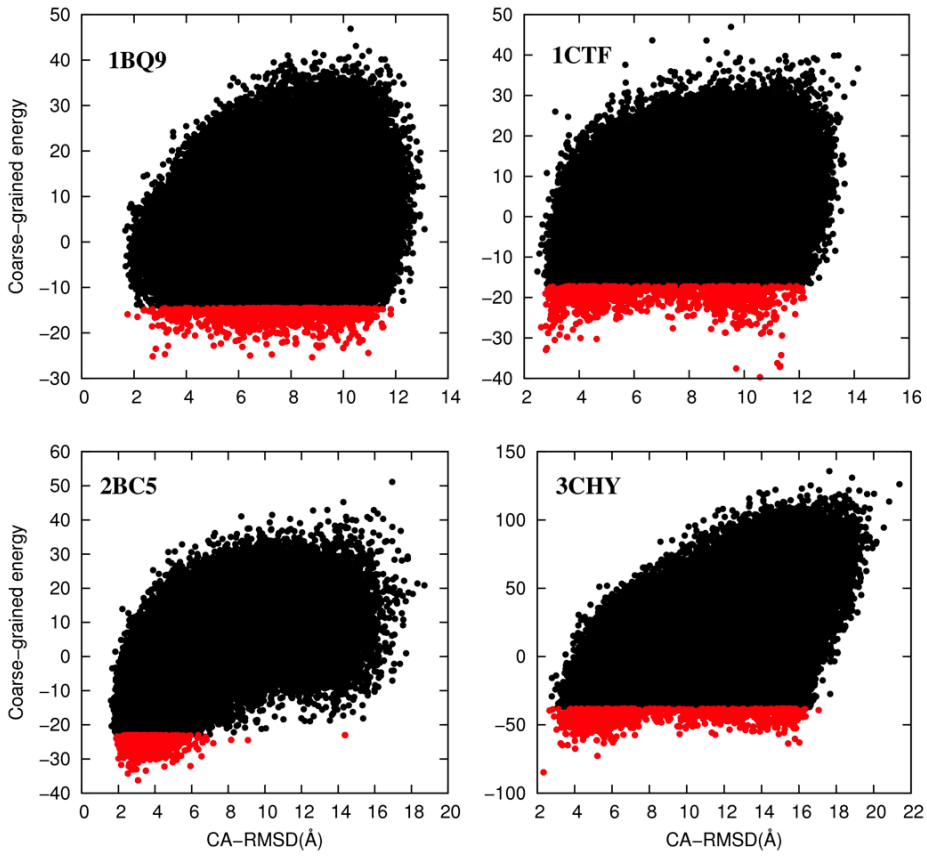

Figure 3.2 Scatter plot between coarse-grained energy and accuracy of the models

the side-chain rotamers and then performed the all-atom refinement through energy minimization. All full-length all-atom models after all-atom refinement were sent for MR using Phaser (McCoy, et al., 2007) to assess its quality for phasing. MORPHEUS used Phaser scores as a criterion to terminate the entire rebuilding process after a few successful *de novo* models for phasing had been obtained (Shrestha, et al., 2011).

3.3. Results

3.3.1. Model accuracy correlated with their divergence

When all models were considered, the coarse-grained energy of the models poorly correlates with their accuracy because those models cover a wide range of distances from the native structure and there is a high degree of degeneracy in energy for less accurate models. This can be seen from the energy landscape in the form of a scatter plot of the coarse-grained energies for all models generated versus their CA-RMSDs from the native structure (Figure 3.2).

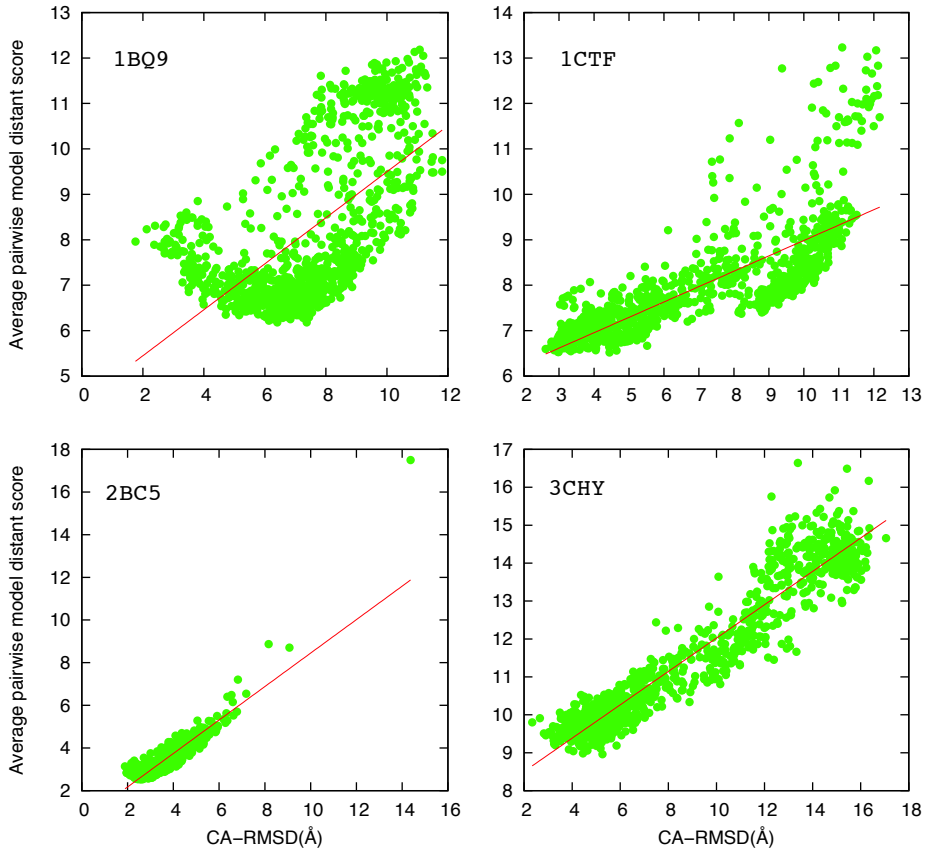

Figure 3.3 Correlation between APMDS and model accuracy

The coarse-grained energy landscape also showed multiple local minima for all targets. This experiment revealed that the models that were nearest to the native structure always were not the lowest energy models. However, the distribution of these low-energy coarse-grained models may encode information about their accuracy. Therefore, one thousand low-energy *de novo* models were selected with the coarse-grained energy in order to exploit the information about their accuracy. The APMDS of these selected models showed the good correlation with the prediction accuracy (Figure 3.3). This APMDS seems to be a useful measure for the assessment and can be used as selection of *de novo* models for model rebuilding.

MORPHEUS aimed to identify residues that are predicted to have large errors. Therefore, the APRDS was calculated for each residue from the selected lowest energy coarse-grained models to indicate how these residues were inaccurate. Importantly, the CA-RMSD of each residue from the native structure also correlated to the APRDS in this experiment (Figure 3.4). The APRDS showed the better correlation (>0.5) with the model accuracy of residues of the molecules (1BQ9, 1CTF, 2BC5, and 3CHY). However, APRDS failed to maintain a same correlation for another molecule 1OPD. This is because many low-energy models in the selected pool were generated as noise. Indeed, APRDS is highly relied on the energy and their distribution with model accuracy. When many low-energy models used in APRDS computation are inaccurate and low-quality, the information in APRDS is either weak or inaccurate. An absolute threshold was not defined to discriminate the correctly predicted residues from that of incorrectly predicted. Instead, all residues were subjected to rebuilding with the sampling frequency proportional to the estimated residue error based on the APRDS. The lower APRDS indicates the more accurately predicted residues and that of higher represents less accurately predicted residues (Figure 3.4). Therefore, the APRDS of each residue can provide the knowledge of the accuracy of that residue in the predicted model.

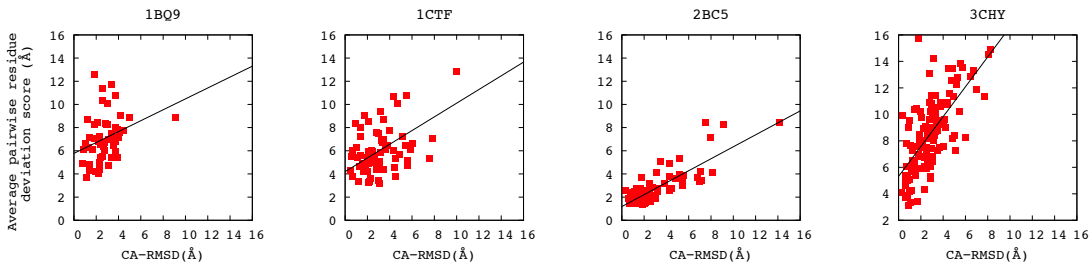

Figure 3.4 Correlation between APRDS and CA-RMSD of the residue in the sequence

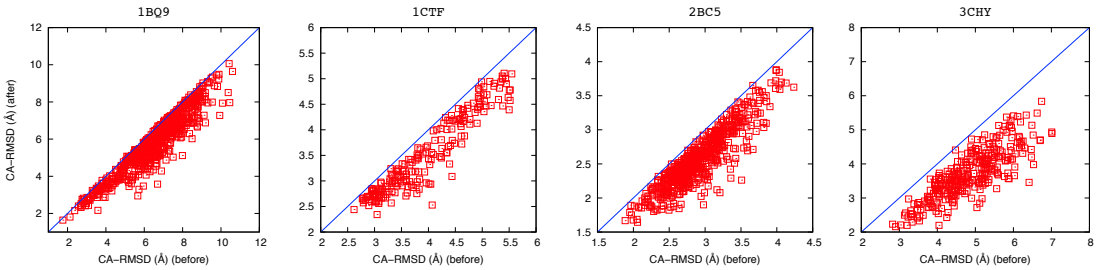

Figure 3.5 Comparison of accuracy of models before and after rebuilding

3.3.2. Accuracy improvement after rebuilding

The coarse-grained models were optimized in the rebuilding procedure to improve the accuracy of all-atom models for successful phasing by MR. This rebuilding procedure improved the coarse-grained models for most targets. These models were only intermediates because these rebuilt models were allowed to all-atom energy optimization before MR. The best models were analyzed after rebuilding and compared to their corresponding input coarse-grained models in order to assess the potential improvement. In this experiment, rebuilt models appeared more accurate than the initial input models (Figure 3.5).

The improvement was also observed for most of the residues in the model (Figure 3.6). MORPHEUS improved the CA-RMSD of coarse-grained models on average from 4.93 to 4.06 Å (Figure 3.7). Importantly, the goal was to improve the *de novo* models to make the suitable search model for MR. The rebuilt models can only be productive when their CA-RMSD is better than 3.0 Å to the target structure because these models might have higher probability to become suitable templates for MR after Rosetta all-atom optimization. Therefore, the model improvement was carefully inspected when their accuracy has CA-RMSD better than 3.0 Å after rebuilding. On average, this method improved a CA-RMSD from 3.38 to 2.60 Å (Figure 3.7). This improved accuracy indicated that these coarse-grained models could be potential candidates for MR. The improvements were observed not only in

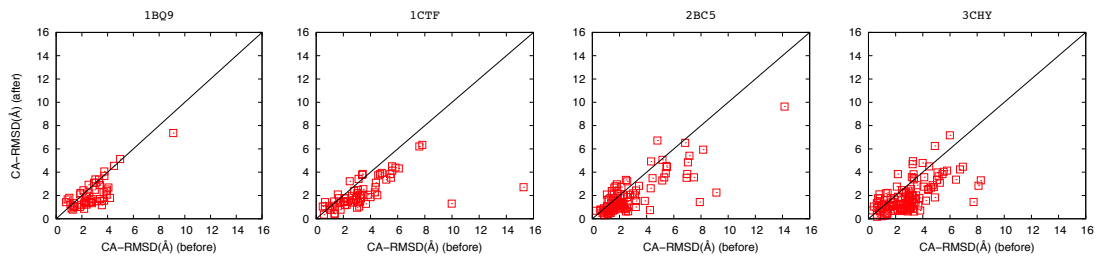

Figure 3.6 Comparison of accuracy of residues before and after rebuilding

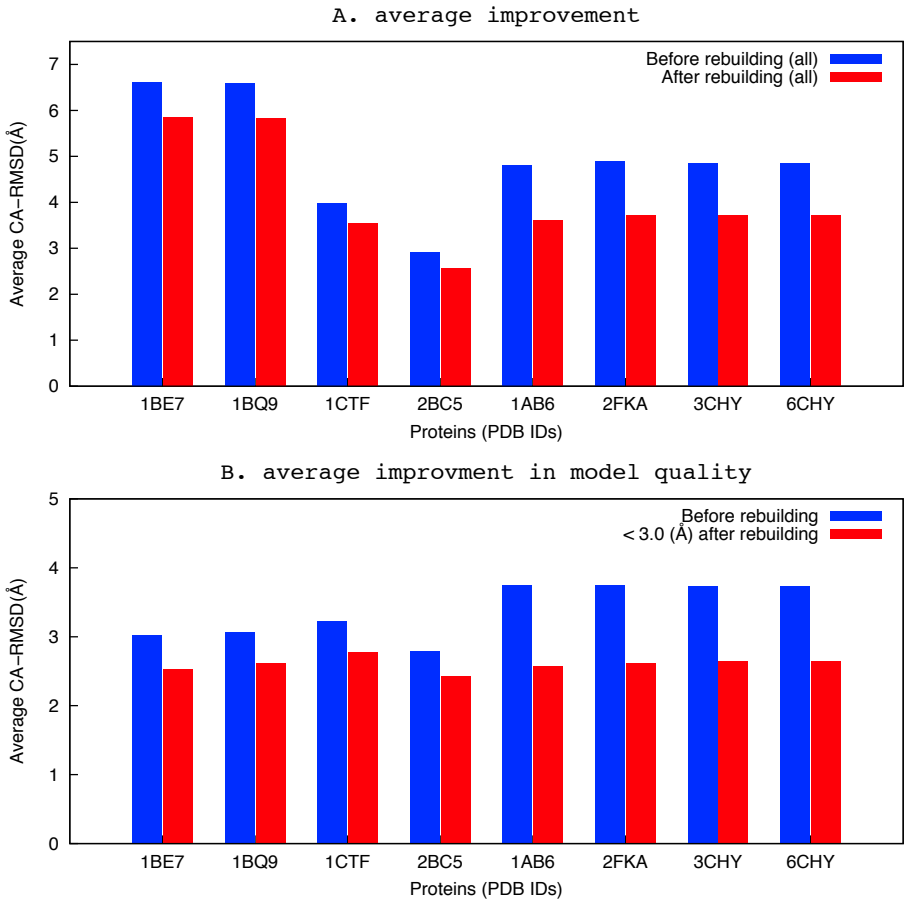

Figure 3.7 Comparison of average improvement in models before and after rebuilding

the loops and termini but also in buried core regions significantly. This method improved α-helical region (residues 65G–74A) in the core of 3CHY by 2.5 Å (Figure 3.8). Similarly, The improvement was observed in α-helical region (residues 38L–44T) in the core of 2BC5 by 0.5 Å. The improvement was further observed in α-helical region (residues 29D–33I) in 1BE7 by 0.8 Å (Figure 3.8).

This rebuilding procedure improved the accuracy for each residue for four targets and one of the best models was selected for each target. MORPHEUS accurately rebuilt a large portion of the residues in the coarse-grained models (Figure 3.8). Furthermore, the improvement was not only seen in particular regions or secondary structures but observed throughout the entire structure (Figure 3.8). This method significantly improved the accuracy of the N- and C-terminal residues because these regions were sampled more frequently owing to their high APRDS (Figure 3.8). Despite all the improvement, it was also observed that some residues with higher APRDS were harder to optimize for the few targets.

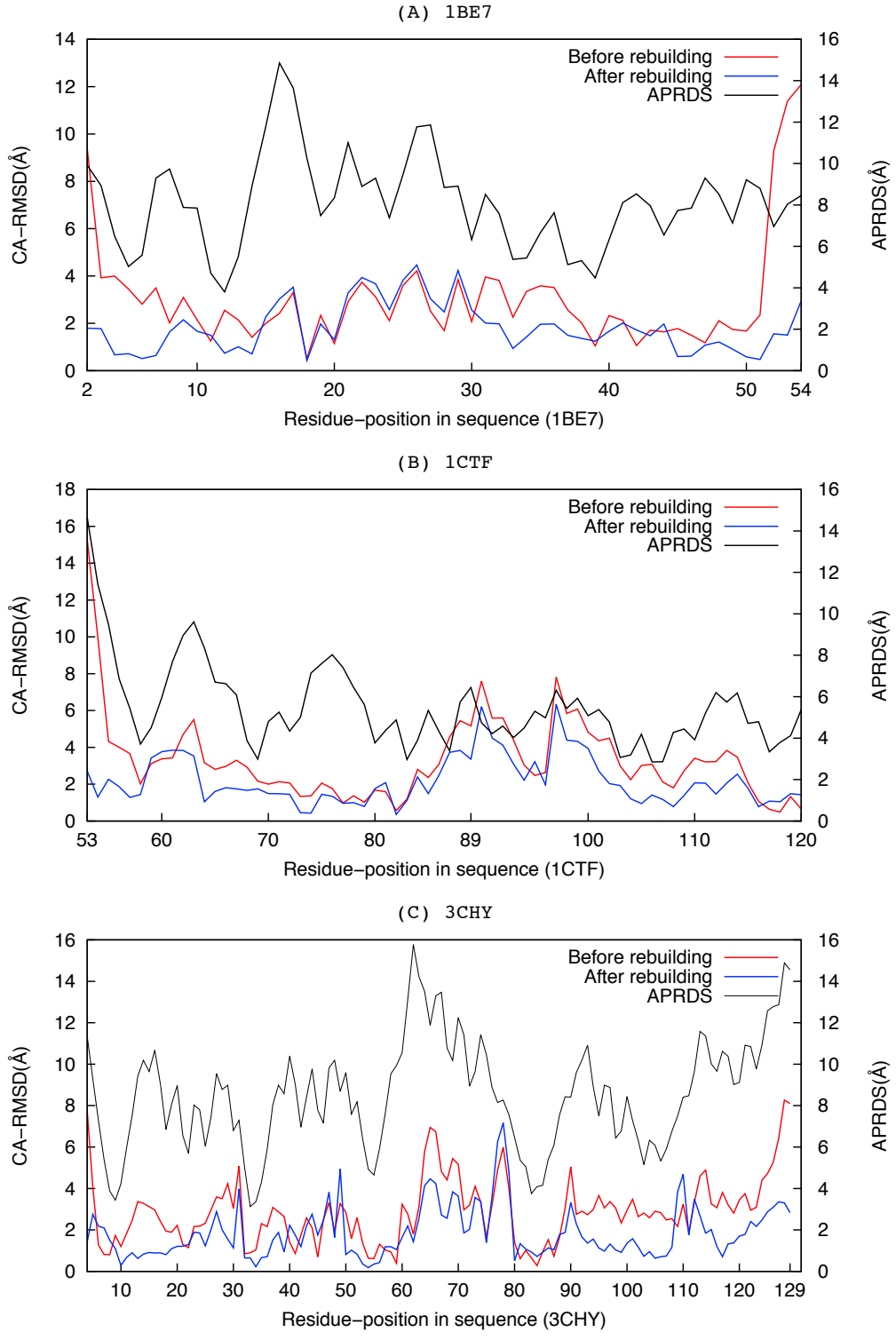

Figure 3.8 Distribution of APRDS of model before and after rebuilding with their accuracy

3.3.3. *Ab initio* phasing with rebuilt *de novo* models

Ten data sets that were unsuccessful using RosettaX (Shrestha et al., 2011) were selected. In addition, two molecules that were already successful were also included. However, MORPHEUS generated accurate search model for seven cases

necessary for successful phasing. These targets were difficult to phase in the previous experiment because model accuracy of search model necessary for MR was not adequate for phasing (Shrestha et al., 2011). All the tested molecules were listed that includes the accuracy of rebuilt models, the phasing statistics, and other relevant information (Table 3.1).

MORPHEUS generated the models better than 2.0 Å in terms of CA-RMSD from the native structure for seven targets. These models produced high TFZ scores after MR indicating the successful MR. Phaser solutions were further evaluated using an MR validation tool on the models with TFZ values of greater than 5.8 however this procedure cannot be used in the absence of the crystal structure (Shrestha et al., 2011). The CA-RMSD was calculated in two different ways in MR validation tool. First, the CA-RMSD was calculated using rigid-body transformation with an optimal translation vector and a rotation matrix that minimized the sum of the squared distances between two sets of coordinates (Kabsch, 1976). Second, the CA-RMSD was computed by applying crystallographic symmetry operators with all permissible origins of the space group. This was called as SYM-RMSD. The model that was placed correctly in asymmetric unit must show the similar CA-RMSDs using two different methods. Therefore, *de novo* models that show the small CA-RMSD difference were considered as successful cases in MR experiment. Therefore, the CA-RMSD difference of 1.0 Å was used in the experiment for molecular replacement solution verification. The CA-RMSD was showed in column 7 of Table 3.1. The models with small differences (around 1.0 Å) between the CA-RMSD (the first number in column 7) and the SYM-RMSD (the third number in column 7) were validated as successful in MR. The ability of successful models were further validated in construction of electron density maps using the initial phases. Low R factor and R

Table 3.1 Summary of MORPHEUS experiment

Targets	Space group	No. of copies in ASU	Sequence Length	Resolution (Å)	RFZ, TFZ	RMSD (CA, All, SYM)	(R-, R-free) factors
1BE7	H3	1	53	1.67	4.4, 6.5	1.33, 1.63, 2.48	0.18, 0.20
1BQ9	$P2_12_12_1$	1	53	1.20	3.8, 6.9	1.61, 2.28, 1.77	0.22, 0.22
1CTF	$P4_32_12$	1	68	1.70	3.0, 6.1	2.67, 3.20, 27.82	-
1OPD	P1	1	85	1.50	4.9, 100.0	9.52, 10.03, 15.93	-
1CM3	$P2_1$	1	85	1.60	4.6, 3.9	9.46, 10.05, 16.94	-
2BC5	$P2_12_12_1$	4	106	2.25	2.5, 9.3	1.26, 2.07, 1.47	0.26, 0.31
3CHY	$P3_1$	2	128	2.20	4.4, 7.9	1.72, 2.25, 1.81	0.20, 0.26
3CHY	F432	1	128	2.00	3.6, 7.0	1.88, 2.60, 1.92	0.24, 0.25
3CHY	$P2_12_12_1$	1	128	1.66	4.0, 8.3	1.80, 2.44, 1.93	0.18, 0.22
3CHY	$P2_12_12$	2	128	2.33	4.0, 9.7	1.96, 2.45, 1.99	0.22, 0.28
1IG5	$P4_32_12$	1	75	1.50	3.7, 6.8	1.59, 2.49, 1.62	0.21, 0.26
256B	P1	2	106	1.40	9.0, 8.6	1.16, 1.82, 1.18	0.24, 0.36

free generated by PHENIX AutoBuild using these MR models further confirmed the successful phasing.

The rebuilding procedure not only improved the quality of models for targets that were unsuccessful for MR in the previous work but also improved the quality of the models that were successful for MR. MORPHEUS achieved the success on two previously successful targets, which were selected for test. Indeed, MORPHEUS improved the CA-RMSD and all-atom root mean square deviation for 1IG5 from 2.36 to 1.59 Å and from 3.13 to 2.49 Å respectively. Similarly, MORPHEUS also improved the CA-RMSD and all-atom root mean square deviation for 256B from 2.60 to 1.16 Å and from 2.90 to 1.82 Å respectively (Table 3.1).

The phases obtained from successful MR models were used in model building and refinement using automated model building programs. The model building and refinement was carried out using the AutoBuild protocol implemented in PHENIX v.1.3 (Adams et al., 2002) with default parameters. The electron density maps constructed using phases from the *de novo* models successfully led to complete three-dimensional protein structures for the seven targets with good R-factor and R free values. These models that were successful in MR were significantly improved

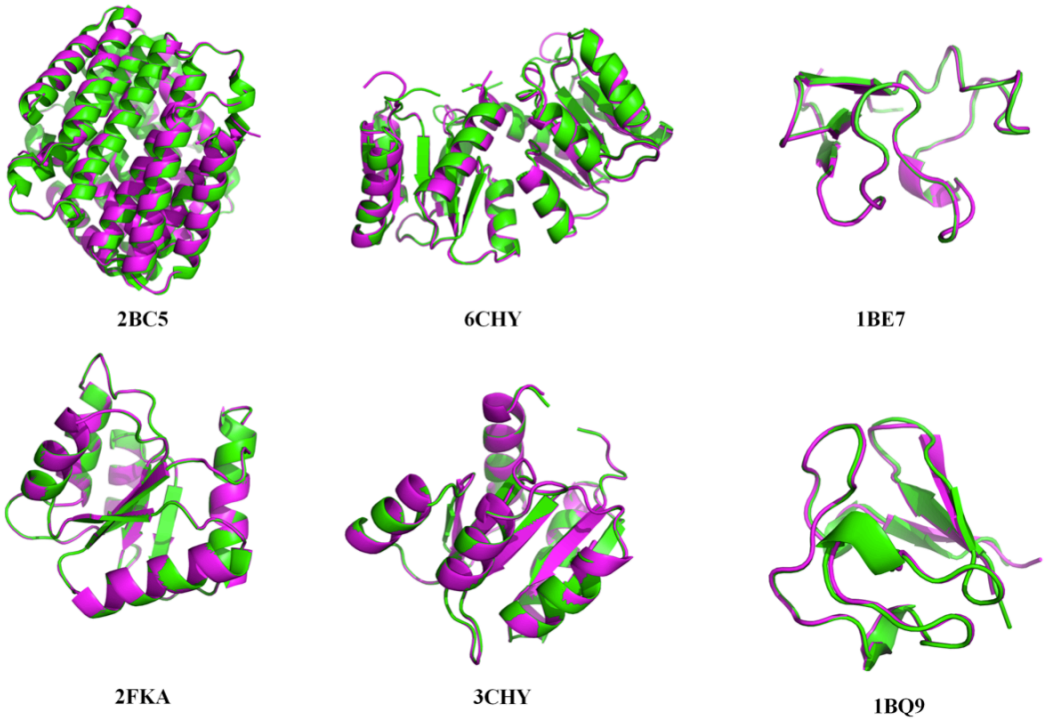

Figure 3.9 Superposition of models after rebuilding to the native structures

after automated refinement (Figure 3.9).

MORPHEUS failed to find MR solution for three diffraction data sets in this experiment. This could be because the energy landscape is far from ideal. This could also arise due to the insufficient sampling. The energy landscape of 1OPD showed the lowest energy models around 12 Å away from the native structure resulting in an anti-correlation between APMDS and CA-RMSD among the selected low-energy models. The APRDS is distributed over a wide range even for the best rebuilt model and there was small improvement. The best rebuilt models for histidine-containing proteins from Escherichia coli (PDB IDs 1OPD and 1CM3) were far away from the native structure, with the largest improvement being from 12 to 8 Å; they were almost impossible to use for phasing in this study. MORPHEUS rebuilt and improved the average CA-RMSD of potential models of the ribosomal protein L7/L12 (PDB ID 1CTF) by 0.45 Å. However, the program was unsuccessful in MR trial for this protein using the rebuilt *de novo* models in this study. In this case, Rosetta all-atom energy minimization did not yield models with sufficient accuracy to solve the phase problem.

3.3.4. Performance measurement

This method was compared with Rosetta 100 CPU day, Rosetta large-scale CPU time (Das & Baker, 2009) and RosettaX (Shrestha et al., 2011) in terms of success rate (Table 3.2) and the computation time required (Figure 3.10). MORPHEUS succeeded on seven out of ten tested cases but RosettaX failed on all these cases. Rosetta 100 CPU-day generated the accurate models required for phasing for two targets out of ten. The success rate was increased from twenty to thirty percentages using the increased computing power in Rosetta large-scale CPU time.

Table 3.2 Comparison of success and failure cases by different methods

Targets	Rosetta, 100 CPU days	Rosetta large-scale	RosettaX	Rosetta3.2	MORPHEUS
1BE7	0	0	0	1	1
1BQ9	0	0	0	1	1
1CTF	0	0	0	0	0
1OPD	0	0	0	0	0
1CM3	0	0	0	0	0
2BC5	1	0	0	0	1
1AB6	0	1	0	0	1
2FKA	1	1	0	0	1
3CHY	0	0	0	0	1
6CHY	0	1	0	0	1

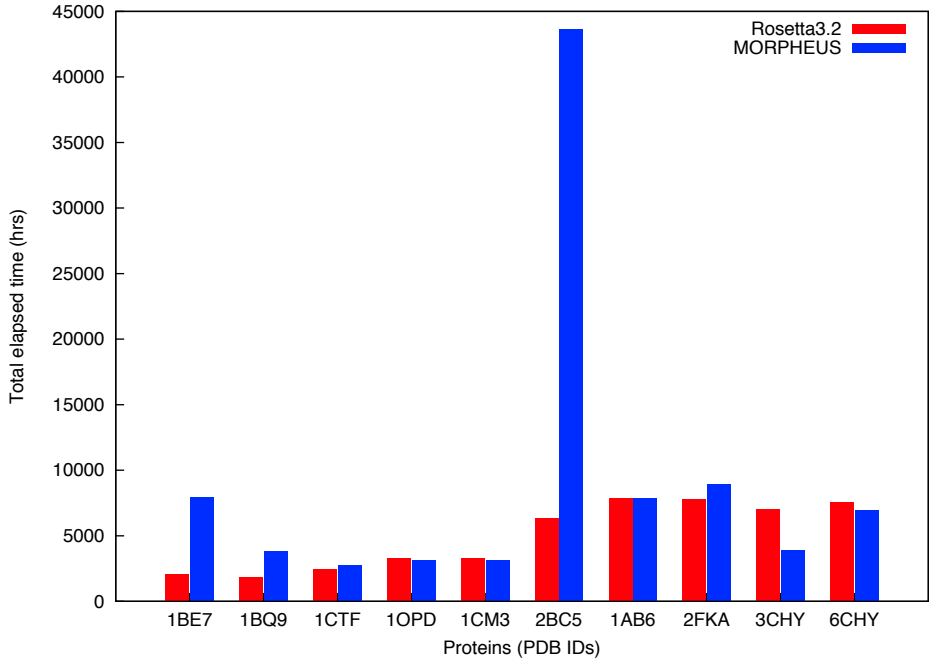

Figure 3.10 Total elapsed time spent by Rosetta3.2 and MORPHEUS

MORPHEUS showed better results than the previously published methods (Das & Baker, 2009; Shrestha et al., 2011) however this comparison might be unfair because of the periodic improvements in Rosetta methods and differences in fragment libraries arising from the increased pool of new structures deposited in the PDB. MORPHEUS result was compared with phasing result obtained from Rosetta3.2 using the same fragment library to measure the impact of this method on model rebuilding more objectively. Three hundred thousands all-atom *de novo* models were generated using Rosetta3.2 for all targets and then ten thousands lowest energy all-atom models was selected for MR experiments. All best models generated by Rosetta3.2 for MR was included (Table 3.3). Rosetta3.2 and MORPHEUS produced the sufficiently accurate model for MR experiment for rubredoxin (PDB IDs 1BQ9 and 1BE7).

Table 3.3 Comparison of best model produced and their result in MR experiment

Sequence	Rosetta3.2				MORPHEUS		
	RFZ, TFZ	RMSD (CA, ALL)	Rank	Best RMSD (CA, ALL)	RFZ, TFZ	RMSD (CA, ALL)	Best RMSD (CA, ALL)
1BE7	4.8, 7.1	1.18, 1.70	1	1.18, 1.70	4.4, 6.5	1.33, 1.63	1.29, 2.01
1BQ9	4.4, 6.8	1.11, 1.41	1	1.11, 1.41	3.8, 6.9	1.61, 2.28	1.39, 2.09
1CTF	3.3, 3.7	2.46, 2.96	1	2.46, 2.96	3.0, 6.1	2.67, 3.20	2.31, 3.03
1OPD	3.7, 100	3.09, 3.99	3	2.89, 3.85	4.9, 100.0	9.52, 10.03	8.33, 9.18
1CM3	3.8, 2.5	3.09, 3.97	3	2.96, 3.85	4.6, 3.9	9.46, 10.05	8.38, 9.13
2BC5	-	1.11, 1.78	2	1.09, 1.86	2.5, 9.3	1.26, 2.07	1.04, 1.68
1AB6	-	2.30, 2.84	1	2.30, 2.84	4.4, 7.9	1.72, 2.25	1.72, 2.25
2FKA	3.9,4.1	2.44, 3.12	1	2.44, 3.12	3.6, 7.0	1.88, 2.60	1.85, 2.61
3CHY	4.2, 4.7	2.37, 3.05	1	2.37, 3.05	4.0, 8.3	1.80, 2.44	1.68, 2.28
6CHY	-	2.37, 2.93	1	2.37, 2.93	4.0, 9.7	1.96, 2.45	1.78, 2.27

However, Rosetta3.2 generated better quality models for these PDBs than MORPHEUS (Table 3.3). These models easily succeeded in MR in both cases. Indeed, MORPHEUS did not include all of the best models for 1BQ9 in the selection of 1000 lowest energy coarse-grained models.

This led to MORPHEUS generating less accurate models than Rosetta3.2. Aside from these two data sets, Rosetta3.2 was unable to predict accurate models for the other eight data sets for phasing. Both RosettaX and Rosetta3.2 predicted the high accurate *de novo* models but the best predicted *de novo* models could not pass the MR test for cytochrome c- (PDB ID 2BC5). MORPHEUS predicted slightly better models than these two methods but these models failed to provide the accurate phase information. However, MORPHUES yielded an MR solution for this target with slightly less accurate models. Many highly accurate models were examined with the diffraction data set during the simulation and this could be the one reason for achieving success. In addition, identical CA-RMSDs could arise from very different structures. The success in MR of a structure with a relatively large CA-RMSD from the native structure could conceivably arise from the errors in the residues being unevenly distributed. Some residues with large errors might have made the overall CA-RMSD relatively high although most of the residues were probably more accurately predicted. For 1OPD and 1CM3, most accurate model was about 9.0 Å CA-RMSD and MORPHEUS was not able to significantly improve the prediction accuracy to make useful as search model.

Total elapsed time spent by the both methods was considered as another factor for comparison. The total elapsed time spent by both methods was monitored although MORPHEUS and Rosetta3.2 are feasible using currently available moderate computing resources. The elapsed time spent in coarse-grained model generation, energy minimization using all-atom models and MR for Rosetta3.2 was accumulated. The elapsed time for MORPHEUS includes the model-rebuilding time in addition to time spent for coarse-grained model generation and all-atom optimization. The elapsed time was measured on the same computing resource. Both methods spent on average equal total elapsed time except for two targets (PDB IDs 2BC5 and 1BE7). The large differences in elapsed time for these two proteins are primarily owing to MR. These are very likely to be the worst-case scenario for MORPHEUS, which needs the generation of all the requested models. At the other extreme, MORPHEUS

can obtain an MR solution with fewer models generated compared with Rosetta3.2. The experiment showed the total number of models generated for phasing varied from 1.0E+04 to 1.8E+05. The total number of models to be generated changed the required elapsed time. MORPHEUS needed less elapsed time when rebuilt models provided MR solution very early in the simulation, such as for the targets 3CHY and 6CHY. Rosetta3.2 missed the best models for *ab initio* phasing when models were selected based on energy although suitable models had already been predicted. However, MORPHEUS executed MR program on all generated *de novo* models MR until successful solutions are found. Therefore, MORPHEUS does not suffer from drawback of missing suitable models, if the accurate models are generated and the energy based selection misses.

3.4. Discussion

3.4.1. Coarse-grained energy landscape

The coarse-grained energy function is designed to enable the sampling of a larger conformational space for simplified protein models, which contain only the main chain and the centroids of side-chain atoms. The coarse-grained energy function aims to search and find the global fold of a target protein by maximizing the burial of hydrophobic side chains and the exposure of hydrophilic side chains. The coarse-grained models generated can be from conformations trapped in multiple minima in a complex energy landscape. This energy function is less accurate due to the missing side-chain atoms. Therefore, it has often less discriminative power to identify the near-native models than its all-atom counterpart. Despite being less accurate and having less discriminative power, the coarse-grained energy can be used to generate near-native models (Das & Baker, 2008).

It is generally assumed that in a randomly sampled energy landscape there should be more models generated that correspond to lower energy than models whose conformations correspond to higher energy. This principle has provided the foundation for the use of clustering methods to identify native-like protein models (Shortle et al., 1998; Zhang & Skolnick, 2004a; Berenger et al., 2011). The geometric similarity among the low-energy models was chosen; this is in principle similar to the clustering methods to identify native-like models. The coarse-grained model is expected to have the small APMDS when this model has more neighbours because

they sample a lower energy level in the energy landscape. Similarly, this principle can also be used to reason for APRDS. The APRDS will be small for a residue that has more neighbours in the generated models because this residue contributes the lower energy to sample a lower energy conformation. The data also showed similar trend (Figure 3.2 and Figure 3.3). In homology modeling, similar concept was employed to generate hybrid models with the best residues from selected templates (Wallner & Elofsson, 2006). AP-MDS or APRDS often correlates to the model quality and it depends on the coarse-grained energy landscape pertains to the protein target. This correlation holds true for most of the targets tested but sometimes it breaks down such as for protein target 1OPD. In this case, the coarse-grained model holds the lowest energy of the models that appeared around 12–14 Å CA-RMSD from the native structure. Therefore, the model rebuilding procedure failed to improve the model for successful phasing. As the CA-RMSD is calculated by comparing the corresponding atoms between the model and the native structure, it may appear to be very large for the purpose of assessing the suitability of a model for MR since it is the spatial matching of the scatters rather than the order with which the atoms are connected that is important for MR. The CA-RMSD was used in the study to measure the quality of a predicted model owing to the critical dependence of the method that was used to generate the model on the connection order of all of the atoms in a protein.

3.4.2. Biased conformational space searching

The APRDS guided the conformation sampling in the rebuilding procedure. This biased sampling was effective for potentially incorrect local residues in a coarse-grained model. As the result, this sampling strategy increased the sampling rate of residues with larger error. Specific threshold was not defined to discriminate correctly and incorrectly predicted residues. Instead, MORPHEUS performed the conformation sampling in the rebuilding procedure at a relative rate proportional to the APRDS.

Wrongly predicted segments are non-uniformly distributed in the model. Terminal and loop segments often contain the largest structural diversity therefore many algorithms have been developed to improve these segments of the protein models. The most effective algorithm for sampling the loop segments is by fixing two anchor points in a protein without changing the entire conformation (Canutescu & Dunbrack, 2003). However, MORPHEUS has sampled the conformational space of the entire protein structure non-uniformly using the APRDS regardless of the

secondary structure and terminals. The non-uniform sampling using APRDS employed in this study may be an alternative strategy to random sampling or sampling of loop regions only.

MORPHEUS improved the model quality over the entire region of a protein in the tested proteins. It significantly improved the input models at the C- and N-terminals, as residues in these segments showed a higher APRDS. It also improved the local segments on the exterior and in the interior of the input models regardless of secondary structure. These residues or segments contain larger APRDS. Some local segments were difficult to improve in the protein structure despite having a higher APRDS. The potential responsible reasons were explored. First, dihedral angles in the conformations represented by the short three-residue fragments could not provide the adequate information to sample near-native regions for these residues. Second, the coarse-grained energy may be inaccurate for these targets to guide the sampling.

3.4.3. Molecular replacement with rebuilt models

Successful MR needs the accurate *de novo* models, which is a critical factor in achieving accurate phases. Therefore, accuracy of *de novo* models is necessary to be improved. In practice, loop regions are mostly hard to predict due to the flexibility and the presence of alternative states. N- and C-terminal segments were also considered to be flexible and harder to predict accurately. Therefore, these segments are often trimmed off in template models prior to the MR experiment when homologous structures were used as search models for MR. This method improved these regions in the model instead of trimming off and can be an alternative strategy to make accurate search model for successful MR.

The errors in the residues having larger deviation with the native structure were reduced and then the overall accuracy of *de novo* models was improved. It is advantageous to rebuild these residues at the coarse-grained stage because the coarse-grained sampling method more effectively explores a large conformational space using cheap computing time. Short fragments (three-residue) were more accurate and its usage in rebuilding of coarse-grained modeling can be fruitful to generate more accurate models. In principle, all-atom models can be also rebuilt. However this is difficult because the current search protocol is designed to avoid drastic changes to the global conformation at the all-atom optimization stage. It is time consuming too.

The higher order non-crystallographic symmetry existed in the crystal tends to require more accurate input models for successful MR. In this case, the successful location of each monomer in the asymmetric unit depends on the solution for the previous monomer and the errors tend to accumulate. This was seen for protein cytochrome c-b562 (PDB ID 2BC5), which is an alpha-helical bundle. *De novo* modeling generated highly accurate models for this target but phasing with these models was not successful mainly owing to the presence of four molecules in the asymmetric unit. The larger loop segments existed in rubredoxin (PDB ID 1BE7) increase the difficulty in phasing. The success of *de novo* models for phasing depends not only on the accuracy of backbone atoms but also that of side-chain atoms. The main chain conformation is improved for ribosomal protein L7/L12 (PDB ID 1CTF), which is insufficient when all-atom optimization cannot lead to improve the accuracy required for phasing. The subsequent all-atom models did not appear to be sufficiently accurate for phasing although the CA-RMSD of the coarse-grained models was significantly improved. Therefore, improvement in all-atom modeling is also important in addition to rebuilding coarse-grained models.

The success of MR relies on the model quality, which has been shown to be as important determinant. However, it has paradoxically been observed that two models with very similar root mean square deviation to the native structure could have opposite outcomes in MR, as in the case of 2BC5. This might be owing to the use of the root mean square deviation as a single measure of the structural differences between two models. The root mean square deviation cannot distinguish various scenarios of structural differences between two models because it is degenerated. The root mean square deviation is quadratic in nature that gives a higher weight to the region that differs the most, whereas for MR the matched regions between the template and target give rise to signal while the mismatched regions generate noise. It is conceivable that alternative measures such as GDT (Zemla et al., 1999; Zemla, 2003), MaxSub (Siew et al., 2000), TM-score (Zhang & Skolnick, 2004b), Q-score (Ben-David et al., 2009) or percentile-based spread (Pozharski, 2010) might be used as a better predictor of success in MR for a given model.

Earlier studies have exploited the conformational variation in an ensemble of predicted models not only for model-quality assessment but also for model rebuilding. Although conformational variation has been employed as a post-filtering measure for

loop prediction (Xiang et al., 2002), which was termed as a 'colony energy', it is used here as both a global (APMDS) and a local (APRDS) measure not only for the estimation of errors but also to guide the sampling and rebuilding of the entire region in a model. Another study has used the local structural variation of models to identify regions that are most likely to be in error and subsequently these regions were aggressively sampled and refined to improve model quality (Qian et al., 2007). Several differences between this method and that of Qian and coworkers will be described. Firstly, MORPHEUS uses coarse-grained models to estimate errors and rebuilds coarse-grained models before subjecting them to all-atom refinement, whereas Qian and coworkers use all-atom models to identify error-prone regions and their rebuilding procedure also uses all-atom models. Secondly, MORPHEUS uses local variation to estimate errors and subsequently uses this variation to guide the sampling proportional to the estimated errors. MORPHEUS did not use the threshold to identify a particular region for rebuilding and used non-uniform sampling in the entire model. In contrast, Qian and coworkers used local variation to identify regions that were most likely to contain errors and then aggressively sampled these regions uniformly regardless of the actual amount of variation within and among the regions. Thirdly, MORPHEUS rebuilt the entire model with sampling proportional to the structural variation. The rebuilding process in MORPHEUS did not create the chain break. MORPHEUS accepted the large conformational changes that cause an increase in energy during the Monte Carlo sampling because a modified acceptance criterion proportional to the structural variation in the form of a temperature factor is introduced. However, Qian and coworkers fixed the C- and N-terminal ends adjacent to the region to be rebuilt and the chain break was closed using the cyclic coordinate procedure.

3.5. Conclusion

In this study, the error-prone residues in a coarse-grained model are rebuilt in order to generate more accurate models with side-chain atoms. These models could be suitable search models for phasing by MR. The number of targets that were unsuccessful in the previous study were tested to evaluate this method. The accuracy of potential coarse-grained models for MR (less than 3.0 Å CA-RMSD from the native structure) was improved from 3.4 Å to 2.6 Å (CA-RMSD) on average. This method significantly reduced the large errors present in the N- and C-terminal

segments. Since *de novo* modeling faced difficulty in the prediction of terminal segments accurately, the rebuilding methodology may be a method for improving the accuracy of terminal segments. Moreover, this method reduced the local errors in the protein models regardless of secondary structure. MORPHESU improves not only in the termini but also in the core regions.

This method increased the success rate of MR when the rebuilt coarse-grained models after all-atom optimization were used. MR succeeded in 70% of the tested cases primarily owing to the improved model quality. Moreover, the phase angles obtained after successful MR were sufficient to generate high-quality electron-density maps for automated model building and refinement.

Chapter 4. NEFILIM – improving fragment quality for *de novo* structure prediction

4.1. Objective

A major challenge in computational structural biology is to predict the atomic-level 3D structures of proteins using their amino acid sequences. One problem is the vast number of conformations to be searched to find the correct structure. Another problem is the lack of an accurate energy function to identify the near-native models. Many methods have been proposed to tackle conformation sampling problems (Liwo, et al., 2008) along with energy function development (Bradley, et al., 2005; Fleishman and Baker, 2012) assuming the principal that native-like models are in the global energy minimum. The most effective strategy to date for searching conformation space efficiently was the usage of fragments from experimentally determined structures (Rohl, et al., 2004).

The fragment assembly approach for *de novo* prediction has been practically implemented in Rosetta program suite (Rohl, et al., 2004; Simons, et al., 1997). It uses two types of fragment (three-residue and nine-residue) queried from experimentally solved structures to generate final models (Bradley, et al., 2005). These models have reached at atomic level accuracy for small-sized globular proteins (Bradley, et al., 2005; Das, et al., 2007). With all these successes, these exist many challenge in *de novo* modeling using the fragment assemble method. The performance of this method principally depends on the conformation sampling strategy and energy function with noticeable exceptions (Sohl, et al., 1998). However, conformation sampling is a major problem in this approach (Kim, et al., 2009). Stochastic Monte Carlo methods are mostly used in conformational sampling to explore the vast conformational space (Liwo, et al., 2008; Simons, et al., 1997; Xu and Zhang, 2012). Conformational search in it is more restricted since it uses substructures provided for each sequence region. Hence the overall prediction quality is certainly dependent on the quality of provided fragments (Hegler, et al., 2009). Therefore, improvement in fragment quality can be potentially used to increase the overall accuracy of structure prediction.

Mostly, fragments for structure prediction are generated given the target sequence. However, optimal fragment generation using the sequence is difficult and

challenging because restraints (set of torsion angles) provided by fragments has to be maintained for uniform sampling (not sampling too broadly and too narrowly). Fragment generation method needs to include the fragments representing the entire distribution of conformations that each sequence segment most probably adopts in the protein structures (Gront, et al., 2011). These fragments are obtained using sequence profiles of aligned sequences with position-specific information on amino acid pattern (Han and Baker, 1996). The position-specific fragment generation using gapless threading was implemented to create continuous and dynamic fragment libraries (Xu and Zhang, 2013). Similarly, an HHM-based method was also introduced for fragment generation (Kalev and Habeck, 2011). Sequence-based fragment generation is not adequate to get precise fragments to predict good quality models because local sequence-structure relationship does not have a one-to-one mapping for all the protein fragments.

The resampling approach was proposed to improve conformational search by providing constraints in fragment selection. Predicted structures of target sequence were used to find out most selective fragments under the given energy function; these fragments are selected again for next-generation of structure prediction (Blum, et al., 2010). This requires at least two rounds of simulations. In this method, an initial round of prediction was executed for learning critical features that resemble the native-like features from models located at the local minima. The next-round of simulation utilizes this information to guide the search towards regions of the landscape corresponding to the native-like structure. This method importantly attempts to search the conformational space that has been sampled more in the initial round. This concept was also used differently in a model-based search for protein structure prediction (Brunette and Brock, 2005). There are methods developed to integrate the information from an initial to next round of sampling to date in fragment assembly approach. Most of these methods differ from each other in information retrieval – the principle features that can be used to predict the properties of native structure. Rosetta resampling method (Blum, et al., 2010) used secondary structures, torsion angles and beta contacts to estimate the native-like properties in the predicted models. Subsequently, these properties were used to improve the conformation searching. The algorithm implemented in Edafold (Simoncini, et al., 2012) estimated the probability of occurrence for native-like fragments in the lowest energy models

and used it for improved sampling iteratively. Indeed, both methods computed the probability of occurrence of particular properties or nine-residue fragments from the lowest energy models to estimate the native-like features.

In this work, a method was designed to generate new fragments from *de novo* models to increase the sampling efficiency near the native region. This method is termed NEFILIM (**NE**w **F**ragments **I**n **L**ibrary **I**mprove **M**odels). The hypothesis is based on that fragments adopted in the lowest energy models are most probably the native-like fragments because these fragments are responsible for minimizing the energy. However, these native-like fragments are scattered in many models located at the local minima. The fragments from the lowest energy models for each residue position were clustered in order to identify native-like fragments. After new fragments were selected, they were used for a new round of prediction. The experiment shows that these new fragments, which are better in quality, increase the sampling near the native region of conformation space. Consequently, NEFILIM predicted more accurate models with the energies closer to the native structure. Moreover, better models were produced with higher concentration with lower energies that makes easier to use energy-based criteria to identify the best models.

4.2. Methods

This approach consists of three major steps as shown in Figure 4.1. First, Rosetta was used to generate a batch of all-atom models giving the target sequence and fragment libraries obtained from Robetta server (Chivian, et al., 2003). This step was termed as an initial run. In second step, the representative models were selected from the pool of predicted models using Rosetta all-atom energy and then the average pairwise residue distance score (APRDS) were computed for each model (Shrestha, et al., 2012). These APRDS were normalized and then locally averaged for three- and nine-residue fragments respectively. The goal of calculating APRDS was to remove distant fragments and then the resultant fragments were clustered according to their respective window sizes. Twenty-five fragments were randomly picked from the top five clusters for each residue position. This is the step for the new fragment library generation. Lastly, this new fragment library was inputted to Rosetta for generation of another batch of all-atom models. This run was termed as a new run.

4.2.1. Benchmark data set and initial model generation

The benchmark test set consists of 30 globular proteins and their sizes range from 49 to 128 residues. These proteins were collected from different studies (Blum, et al., 2010) and (Tyka, et al., 2011) and few targets were taken from CASP8 and CASP9. Rosetta3.2 generated 120,000 full atom models using the amino acids sequence and fragment libraries (three-residue and nine-residue) only. Initial fragment libraries were generated from Robetta Server using the protein sequence only.

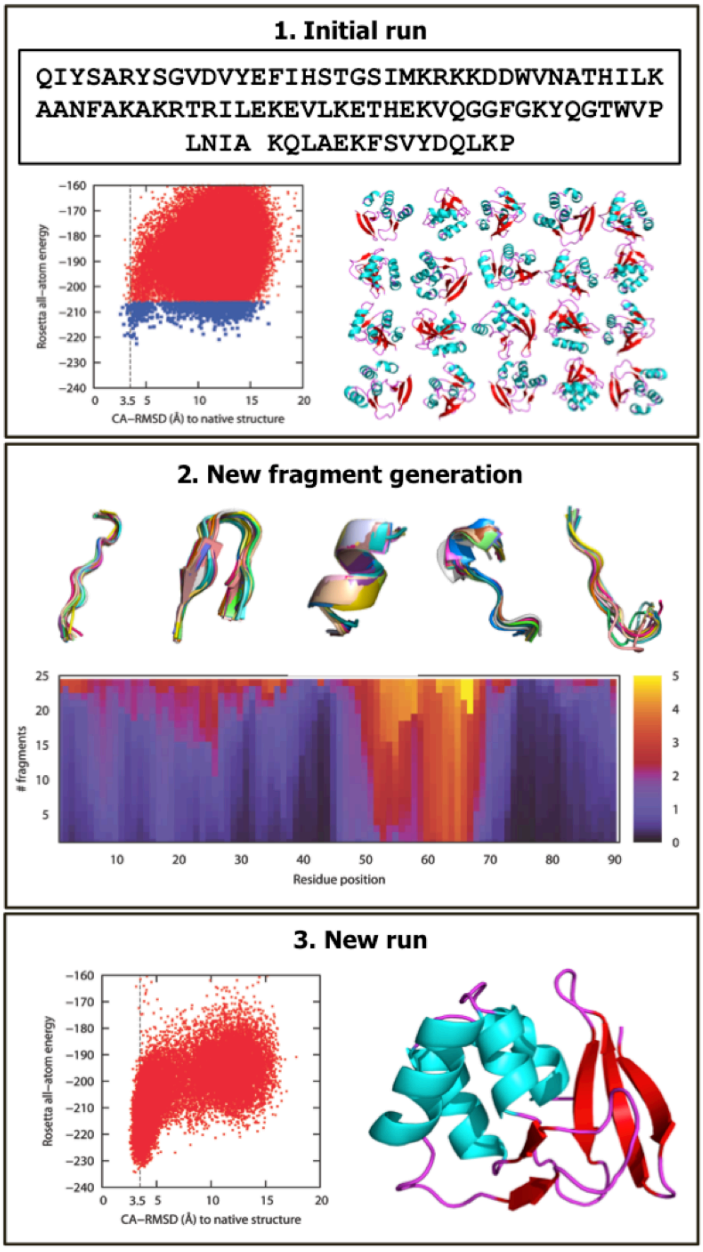

Figure 4.1 An overview of NEFILIM

Furthermore, homologous proteins of the target protein sequence were excluded in the fragment library generation. This model generation is termed as initial run.

4.2.2. Improved fragment library generation

As the representative models, one thousands lowest energy models from the pool of 1.2E+5 models were taken for each target. Each model was superimposed to other (999) models in order to calculate the APRDS in the residue level. This score was further normalized and processed to compute the average for each residue locally with two sliding windows (three and nine). The APRDS of three successive residues were averaged to get a score for the three-residue fragment. Same procedure was employed to obtain the average APRDS for nine-residue fragment. The APRDS was used to remove the fragments that are distant from the majority of fragments before clustering. The APRDS 0.30 was used as cutoff. This cutoff was determined by testing on a few proteins (2BC5, 1IG5, 1CTF, and 1BM8).

Three-residue and nine-residue fragments were generated from the selected models for each residue position. These two types of fragment were independently clustered using the algorithm implemented in Durandal (Berenger, et al., 2011). The cluster radius was determined for three-residue and nine-residue fragments in order to carry out the clustering of each target of benchmark data. The proteins (2BC5, 1IG5, 1CTF, and 1BM8), which were used to determine the optimal clustering threshold, showed the clustering radii of 1.00 Å and 0.20 Å suitable for nine-residue and three-residue fragments for the experiment. Twenty-five fragments were randomly picked from the top five clusters. The number of fragments in a particular cluster and their proportion in the 25 fragments decided the representation of each cluster in the fragment library. The rationale for choosing the top five clusters is to include a diverse set of fragments. The concept of choosing the fragments from top five clusters is also similar to selecting representative models at the center of clusters from the largest clusters in the structure prediction (Moult, 2005). The selection process was limited to the maximum number of available clusters when the number of clusters was less than five. In the reverse, more than five clusters were allowed for fragment selection when a number of fragments were sparsely distributed in the clusters. Clusters that contained less than two members were excluded in the selection process in this experiment.

4.2.3. Resampling with new fragments

Resampling is final step of this method. In this resampling step, a new run was carried out using Rosetta3.2 with the target sequence and the new fragment library as the input. The conformation sampling algorithm and energy function used were the same as in the initial run. In the resampling process, the simulation was again started with generation of coarse-grained models to the all-atom models. Conformational search mostly occurred near the conformational space that is responsible to yield the lowest energy models in the new run because the new fragments were generated from the lowest energy models from the preceding simulation. Therefore, 30,000 full atom models were generated for the new run. This experiment was stopped after two iterations. This experiment generated 1.5E+05 models in total for each protein sequence.

4.3. Results

4.3.1. New fragments from the *de novo* models

The accuracy of models can also be improved by providing better quality fragments in *de novo* structure prediction. This work studies the better fragments generation from the representative *de novo* models to improve the prediction

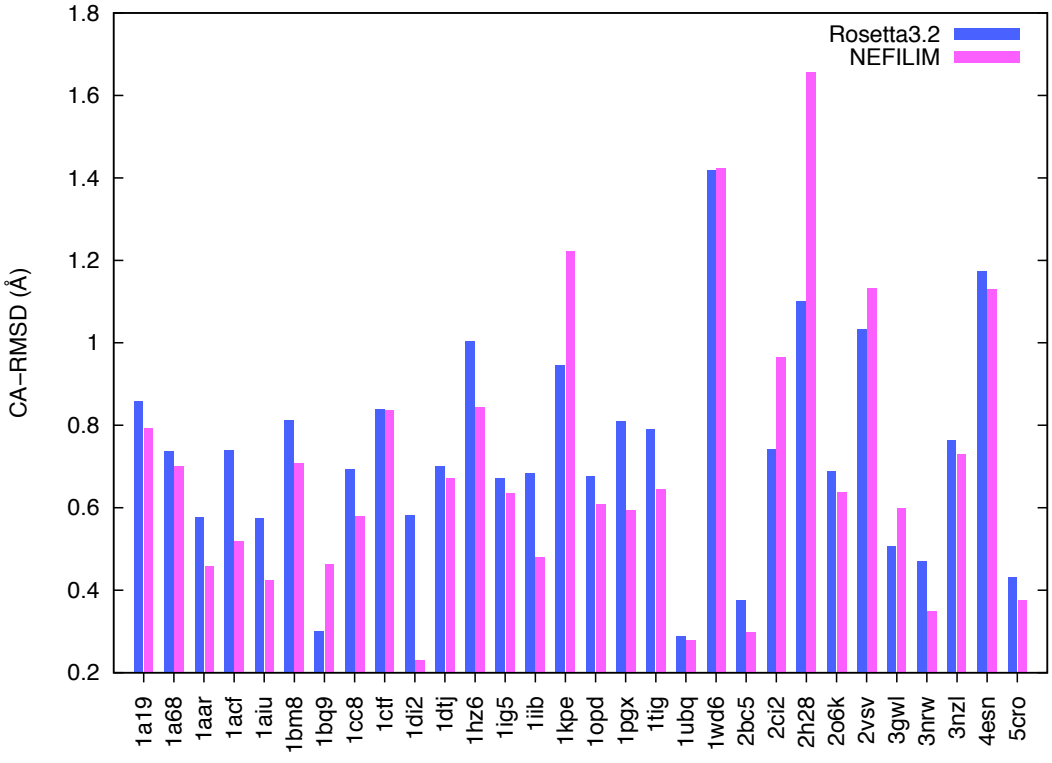

Figure 4.2 Quality of best fragment in structure-derived and sequence-derived fragment library

accuracy. The accuracy of selected fragments from the lowest energy models was examined. In order to evaluate the accuracy of fragments, each fragment was superimposed to the corresponding fragment in the native structure and their difference was measured in terms of CA-RMSD. The new structure-derived fragment library contained more accurate fragments than sequence-derived fragment library. The accurate fragments were observed for most residues of the benchmark proteins. The improved accuracy in the fragment sets was measured in two ways – improvement in the best fragment and enrichment of better fragments in the new fragment library.

The most accurate fragments used in the initial run were further improved in the new fragment library. Nine-residue fragments achieved the improvement in CA-RMSD for 23 proteins from 30 tested proteins (Figure 4.2). A small improvement from 0.73 Å to 0.70 Å was observed for all 30 proteins on average. This improvement was observed when the most accurate fragments of the two libraries were compared. The most significant improvement (0.35 Å CA-RMSD) was seen in protein 1DI2 followed by other proteins (1ACF, 1AIU, 1IIB, 1HZ6) where the improvement was more than 0.10 Å. The accurate fragments were also enriched in the structure-derived library (average CA-RMSD 1.24 Å) compared to the sequence-derived library

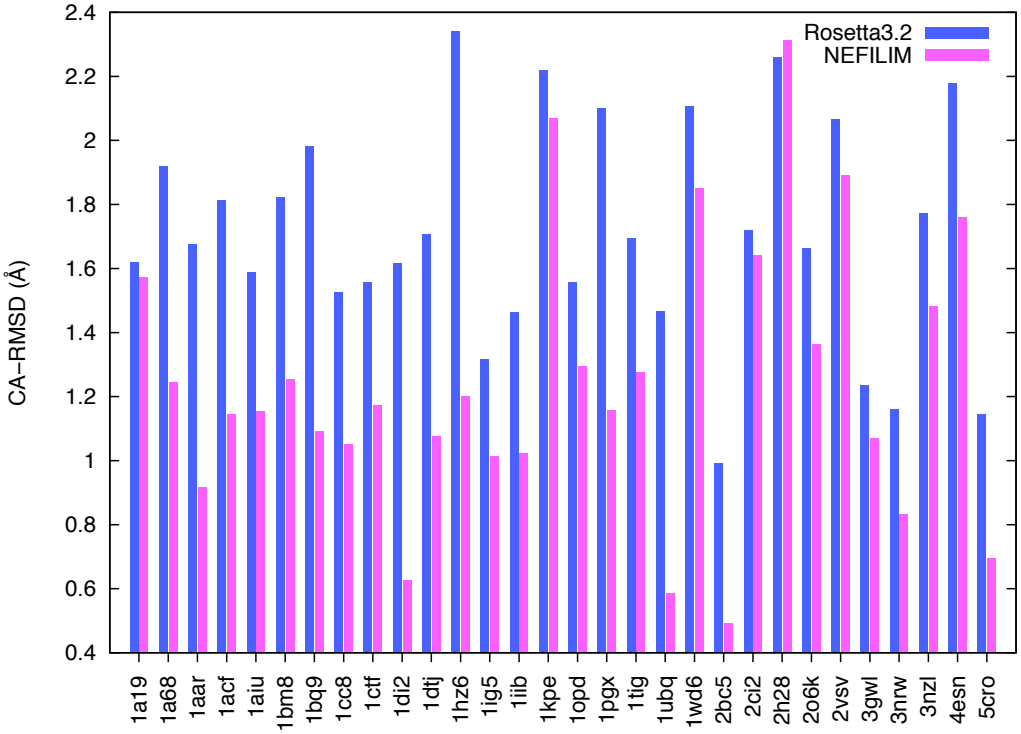

Figure 4.3 Enrichment of good quality in sequence-derived and structure-derived fragments

(average CA-RMSD 1.71 Å) for nine-residue fragments (Figure 4.3). Because the proportion of accurate fragments has increased in the nine-residue fragments, the average CA-RMSD of fragments was dropped for all the targets (Figure 4.3). The most significant improvement was observed in 1HZ6 where the improvement was 1.15 Å.

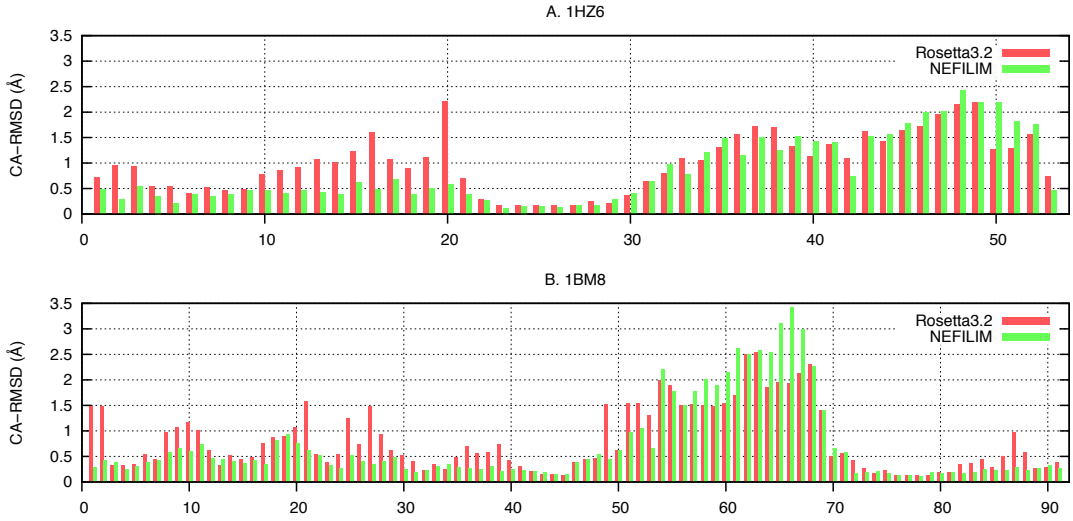

Figure 4.4 Best fragment for each residue position (nine-residue)

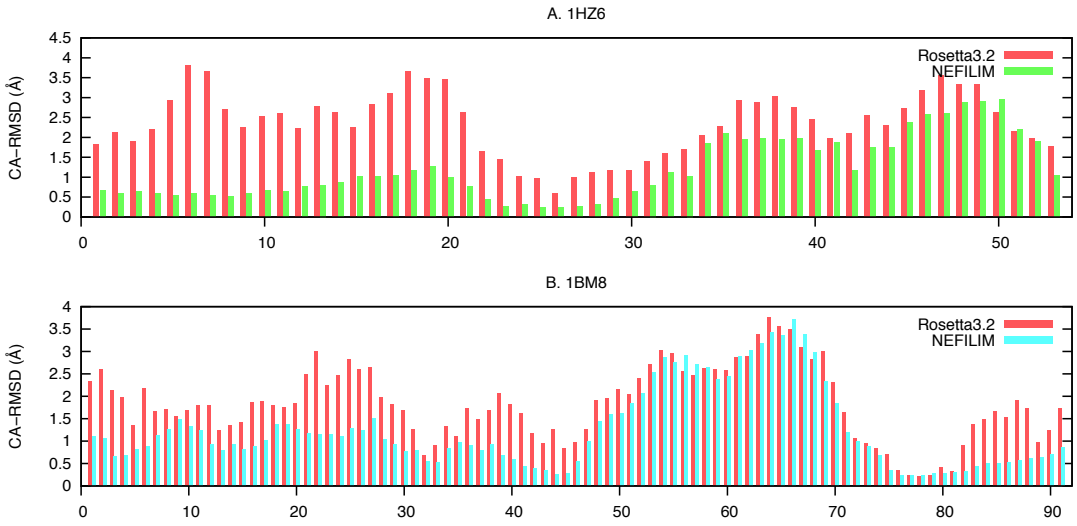

Figure 4.5 Average accuracy of fragments at each residue position (nine-residue)

The improvement in the fragment quality on each residue was examined and the result for two proteins targets was provided – 1HZ6 and 1BM8. The most accurate nine-residue fragments were significantly improved for the N-terminal residues of 1HZ6 (Figure 4.4) in structure-derived library. Similarly, most residues for 1BM8 also achieved the more accurate fragments than its sequence-derived library (Figure

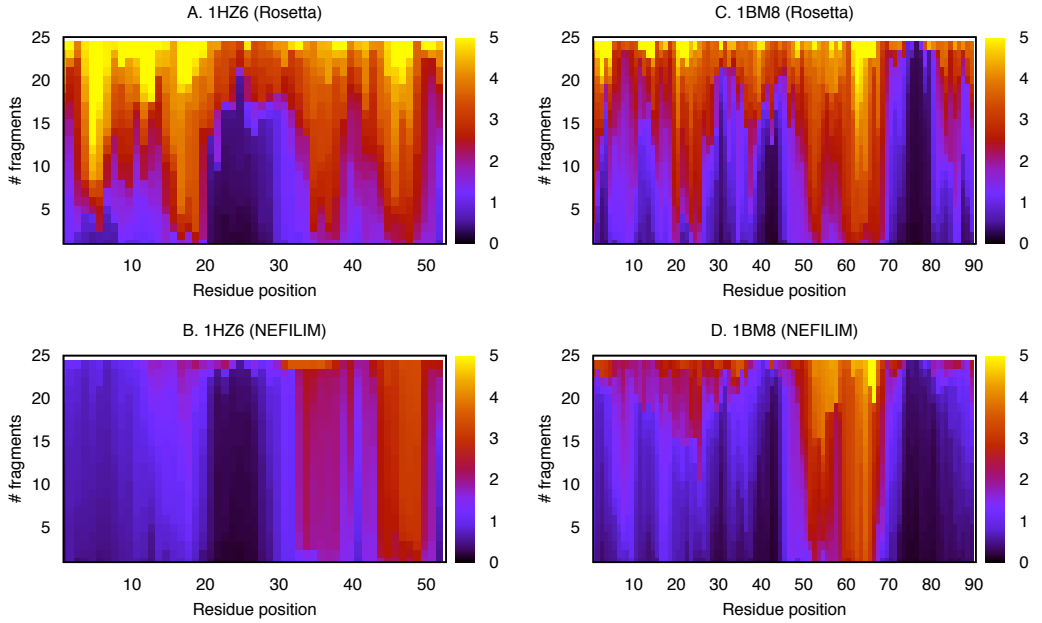

Figure 4.6 CA-RMSD of twenty-five fragments at each residue position (nine-residue)

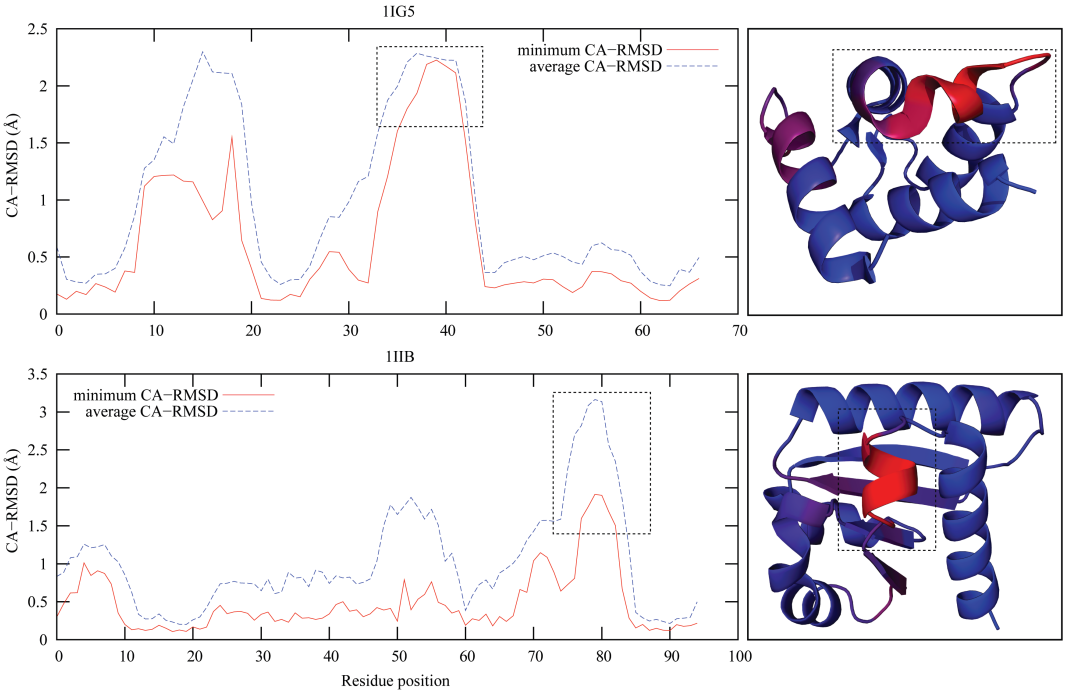

Figure 4.7 Fragment quality and unusual secondary structure

4.4). Structure-derived fragments libraries consist with better fragments for 1HZ6 and 1BM8 (Figure 4.5). Many better fragments were accumulated and for some residues (1HZ6 15G-20A and 1HZ6 20A-27E), the fragment quality was significantly improved although sequence-derived library did not initially have such a good fragments (Figure 4.6). This might be due to conformational sampling.

The fragments for some residues are difficult to improve because these residues belong to irregular secondary regions, such as residues 36P-41K in 1IG5, and residues 80D to 84Y in 1IIB (Figure 4.7). Furthermore, fragments for residues for some other proteins (2H28, 1KPE and 2VSV) were also hard for improvement because better fragments were not also present in the sequence-derived fragment library.

4.3.2. Model accuracy improvement

The accuracy of energy or scoring function is the key step in the model selection to assess the prediction quality in the absence of the native structure. In this study, the best prediction was evaluated in two ways mimicking a blind prediction. First, the lowest energy model was chosen as the best prediction. Secondly, the best model from the top five lowest energy models was selected as the best prediction. It was termed as the "best in five" model. For the comparative analysis, the results of initial run with new run were combined and the number of total models became

Table 4.1 Prediction performance by Rosetta and NEFILIM based on lowest energy models

PDB	Sequence length	Rosetta (Lowest energy model)				NEFILIM (Lowest energy model)			
		Rosetta all-atom energy	CA_RMSD	TM Score	GDT-TS	Rosetta all-atom energy	CA_RMSD	TM Score	GDT-TS
1A19	89	-192.5	7.49	0.41	44.9	-192.7	6.77	0.46	47.5
1A68	87	-197.7	6.70	0.44	45.1	-202.7	9.03	0.45	45.4
1AAR	76	-165.6	4.77	0.52	57.2	-170.2	2.35	0.84	85.2
1ACF	125	-261.5	5.68	0.58	52.6	-272.5	3.17	0.69	62.8
1AIU	105	-224.1	2.01	0.82	78.6	-225.4	1.46	0.91	90.0
1BM8	99	-233.4	3.93	0.67	65.4	-234.3	3.02	0.78	78.8
1BQ9	53	-103.1	1.99	0.74	82.1	-107.0	1.78	0.74	83.5
1CC8	72	-147.5	4.37	0.49	58.7	-147.5	4.37	0.49	58.7
1CTF	68	-154.9	3.58	0.55	62.9	-155.3	3.28	0.56	66.5
1DI2	69	-152.0	1.95	0.80	84.8	-154.9	0.89	0.92	95.3
1DTJ	74	-160.4	3.27	0.81	83.1	-162.1	2.24	0.83	86.2
1HZ6	61	-136.2	3.57	0.51	63.9	-137.3	3.15	0.60	72.5
1IG5	75	-173.6	3.35	0.63	68.3	-176.3	2.19	0.73	77.0
1IIB	103	-216.6	11.84	0.54	51.2	-221.3	7.87	0.63	61.9
1KPE	113	-220.9	10.66	0.39	36.5	-223.8	8.25	0.35	32.1
1OPD	85	-182.0	5.06	0.42	47.4	-183.3	4.63	0.49	52.4
1PGX	70	-152.4	5.42	0.69	72.9	-155.4	5.33	0.79	81.1
1TIG	88	-189.4	5.39	0.47	53.4	-194.2	3.05	0.67	67.9
1UBQ	76	-174.2	2.07	0.90	92.8	-174.2	2.07	0.90	92.8
1WD6	86	-198.0	12.70	0.27	33.1	-198.0	12.70	0.27	33.1
2BC5	106	-252.5	1.93	0.85	81.4	-257.2	1.55	0.89	86.1
2CI2	65	-135.7	7.70	0.39	43.9	-142.5	7.57	0.42	44.6
2H28	109	-212.3	14.83	0.31	28.2	-221.6	9.66	0.43	40.4
2O6K	72	-161.4	5.14	0.52	58.7	-168.4	10.00	0.45	46.9
2VSV	81	-163.5	6.72	0.48	51.5	-164.6	9.44	0.40	41.7
3GWL	106	-232.0	9.16	0.49	47.4	-238.2	9.70	0.39	39.2
3NRW	104	-230.6	7.16	0.37	34.4	-241.1	1.93	0.83	82.5
3NZL	73	-151.4	11.81	0.29	33.2	-154.7	5.10	0.44	52.1
4ESN	78	-155.9	6.15	0.28	32.4	-164.2	5.03	0.33	38.1
5CRO	60	-131.0	3.26	0.81	87.1	-133.5	3.39	0.83	88.3
Mean		-182.1	5.99	0.55	57.8	-185.8	5.03	0.62	64.3

1.5E+05. This is termed as a NEFILIM run.

NEFILIM run improved the accuracy for 22 proteins compared to Rosetta when the lowest energy model was considered as the best prediction. The prediction accuracy was assessed using three different measurement tools – CA-RMSD, template modeling score (TM-score), and global distance test – total score (GDT-TS). The CA-RMSD was improved to 5.03 Å (NEFILIM) from 5.99 Å (Rosetta) on average for all protein targets of benchmark data. Both, TM-score and GDT-TS, were also improved by 7% in NEFILIM than Rosetta (Table 4.1). The improvement was further tested using the Student's t-test package in R version 3.0.2 software (Team, 2008) for the statistical significance. The improved results were also statistically significant in all the three cases with 95% confidence interval. The p-value was less than 0.05 in the paired t-test for all three methods.

The further analysis was performed using only CA-RMSD as its simplicity and generality. Models having CA-RMSD less than 3.5 Å to native structure were closely inspected because these models are practically useful such as in solving crystallographic phase problem by MR (Blow and Rossmann, 1961; Qian, et al.,

Table 4.2 Accuracy of best in top five models generated by Rosetta and NEFILIM

PDB	Sequence length	Control run (Best in top five)			NEFILIM (Best in top five)		
		CA_RMSD	TM Score	GDT-TS	CA_RMSD	TM Score	GDT-TS
1A19	89	2.30	0.74	72.8	2.51	0.71	72.8
1A68	87	6.70	0.44	45.1	6.42	0.45	47.7
1AAR	76	4.70	0.54	60.2	1.30	0.90	91.5
1ACF	125	4.87	0.58	52.6	3.00	0.73	66.2
1AIU	105	1.98	0.83	80.0	1.46	0.91	90.0
1BM8	99	3.19	0.74	70.2	3.00	0.78	78.8
1BQ9	53	1.82	0.83	90.6	1.78	0.74	83.5
1CC8	72	2.18	0.78	80.2	3.65	0.59	64.2
1CTF	68	3.14	0.61	68.4	3.28	0.55	66.5
1DI2	69	0.99	0.91	94.6	0.86	0.93	95.7
1DTJ	74	2.30	0.85	87.2	2.21	0.84	86.8
1HZ6	61	3.49	0.60	69.7	3.06	0.60	72.5
1IG5	75	2.39	0.76	78.7	2.14	0.73	77.0
1IIB	103	2.71	0.72	69.9	2.08	0.80	77.2
1KPE	113	7.30	0.48	42.0	8.25	0.35	32.1
1OPD	85	3.14	0.57	62.4	4.09	0.49	53.2
1PGX	70	5.42	0.70	72.9	5.24	0.79	81.1
1TIG	88	3.16	0.65	67.1	2.84	0.71	72.4
1UBQ	76	1.41	0.92	94.4	1.52	0.91	93.4
1WD6	86	3.79	0.50	50.9	3.79	0.50	50.9
2BC5	106	1.83	0.87	85.1	1.52	0.89	86.8
2CI2	65	6.50	0.44	48.1	6.39	0.48	51.9
2H28	109	9.61	0.47	44.5	9.66	0.43	40.4
2O6K	72	4.83	0.52	58.7	2.33	0.72	75.7
2VSV	81	6.72	0.48	51.5	6.72	0.48	51.5
3GWL	106	8.15	0.49	47.4	5.52	0.55	55.4
3NRW	104	3.33	0.73	67.3	1.73	0.83	82.5
3NZL	73	5.49	0.39	46.2	3.44	0.58	62.7
4ESN	78	5.15	0.33	38.8	4.32	0.36	40.7
5CRO	60	3.24	0.82	87.1	3.39	0.83	88.3
Mean		4.06	0.64	66.1	3.58	0.67	69.6

2007). Fifteen tested proteins showed CA-RMSD better than 3.5 Å when the lowest energy models were compared, while nine proteins achieved the accuracy better than CA-RMSD 3.0 Å. The average accuracy for the proteins, which have CA-RMSD less than 3.5 Å either in Rosetta or NEFILIM, was improved from 3.59 Å (Rosetta3.2) to 2.36 Å (NEFILIM).

The performance was also evaluated using the "best in five" models. NEFILIM predicted better models (CA-RMSD of 3.58 Å) on average than Rosetta (CA-RMSD of 4.06 Å) for 20 out of 30 proteins (Table 4.2). The improvement was also evaluated using TM-score and GDT-TS. TM-Score was, on average, improved from 0.64 (Rosetta) to 0.67 (NEFILIM). Furthermore, NEFILIM improved the GDT-TS by 4% than Rosetta. The improvement has a p-value of 0.01 (CA-RMSD), 0.06 (TM-score), and 0.04 (GDT-TS) using the paired t-test with 95% confidence interval (Table 4.2). Nineteen proteins showed the accuracy less than CA-RMSD 3.5Å in NEFILIM. Among these proteins, fourteen targets were predicted with accuracy better than 3.0 Å CA-RMSD. In the contrary, Rosetta showed 17 cases with less than 3.5 Å and 10 cases with less than 3.0 Å. Altogether, twenty two proteins showed CA-RMSD better than 3.5 Å with a mean of 2.98 Å (Rosetta) and 2.44 Å (NEFILIM) respectively.

NEFILIM was further compared with a similar method, EdafoldAA (Simoncini and Zhang, 2013). Due to the huge computational power requirement, the result reported in the article (Simoncini and Zhang, 2013) was used in order to compare the accuracy of NEFILIM with EdafoldAA. Fifteen targets were observed common in both experiments. This experiment showed an average CA-RMSD of 3.73 Å for the lowest energy models compared to 3.96 Å from EdafoldAA. Similarly, NEFILIM also performed better than EdafoldAA when the best in five models were compared. The average CA-RMSD was 2.78 Å (NEFILIM) and 3.21 Å (EdafoldAA) respectively.

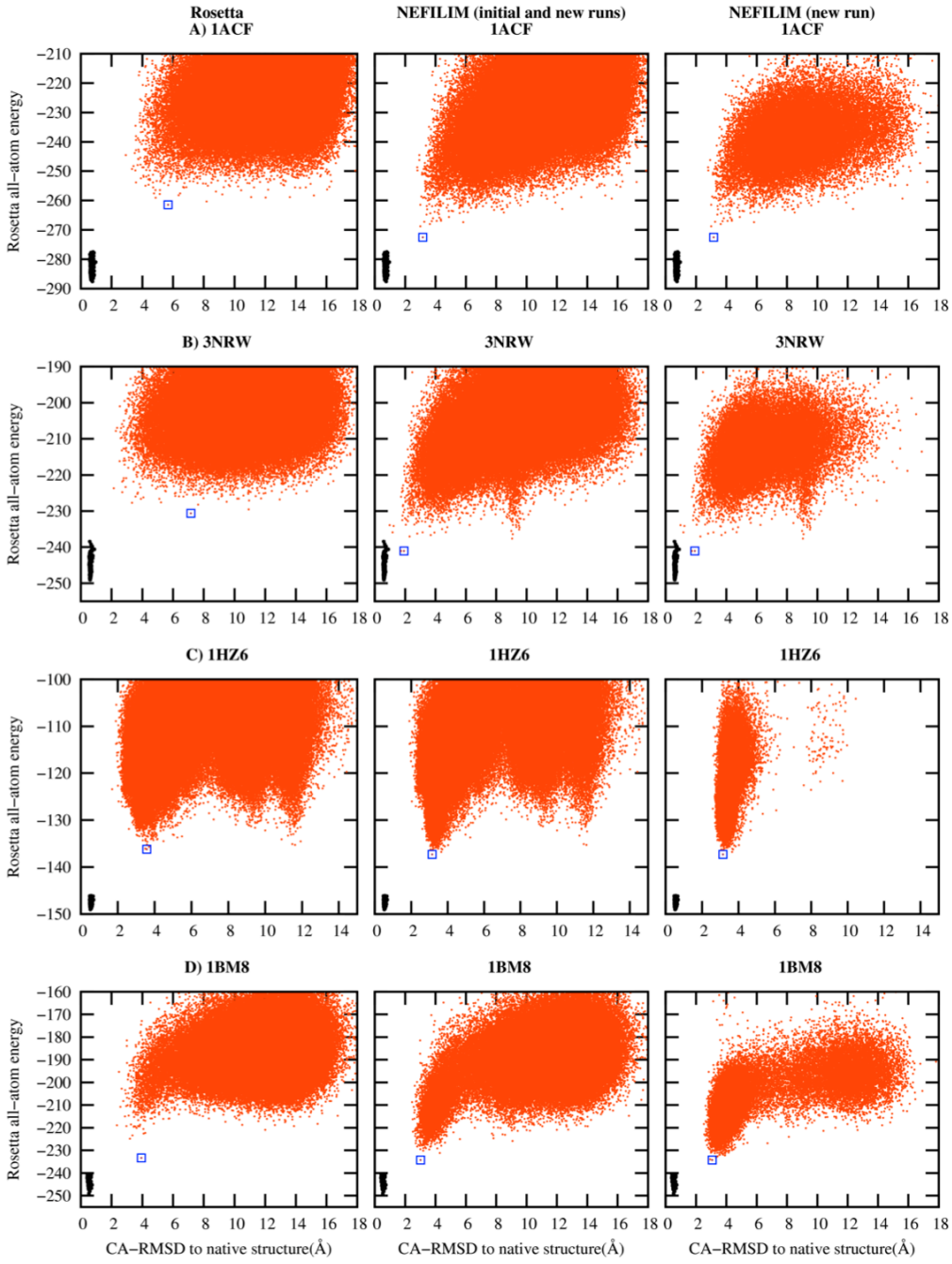

Figure 4.8 Scatter plot between energy and accuracy

Among the successful targets, four proteins (1ACF, 3NRW, 1BM8, and 1HZ6) were selected for the analysis of energy landscape verses prediction accuracy (Figure 4.8). Energy landscape contained less ruggedness in NEFILIM new run for three proteins (1ACF, 1BM8, and 1HZ6). In addition, the energy gap was reduced between the relaxed native models and the lowest energy models for 1ACF and 3NRW (Figure 4.8). NEFILIM new run predicted the accurate models with lower energies than Rosetta run for these two targets. Rosetta generated energy landscape

Table 4.3 Comparison of average energy and their accuracy

Targets	Sequence length	Rosetta (average 200 models)		NEFILIM (average 200 low energy models)	
		Average Rosetta energy	Average CA-RMSD	Average Rosetta energy	Average CA-RMSD
1A19	89	-184.3	10.3	-186.54	9.49
1A68	87	-186.6	10.8	-190.83	10.77
1AAR	76	-160.0	5.7	-163.81	4.19
1ACF	125	-252.9	9.5	-260.66	6.31
1AIU	105	-207.1	9.2	-209.35	9.37
1BM8	99	-214.3	8.9	-225.04	3.67
1BQ9	53	-95.2	4.3	-101.51	2.54
1CC8	72	-141.7	5.8	-143.2	6.84
1CTF	68	-147.7	5.2	-153.08	3.42
1DI2	69	-146.1	2.5	-150.43	1.58
1DTJ	74	-150.6	4.3	-154.22	3.74
1HZ6	61	-132.1	4.7	-134.53	3.4
1IG5	75	-166.9	4.3	-173.51	4.63
1IIB	103	-208.4	10.3	-215.34	10.44
1KPE	113	-210.2	11.8	-215.35	10.91
1OPD	85	-171.0	8.4	-171.57	8.23
1PGX	70	-145.8	6.7	-150.85	6.19
1TIG	88	-182.5	6.1	-183.73	4.88
1UBQ	76	-168.1	2.5	-171.1	2.66
1WD6	86	-183.5	9.0	-183.22	9.49
2BC5	106	-248.6	2.3	-252.58	2.02
2CI2	65	-127.5	8.0	-136.97	8.46
2H28	109	-204.3	12.5	-214.19	11.24
2O6K	72	-153.5	8.6	-160.12	8.36
2VSV	81	-154.0	11.0	-155.88	10.55
3GWL	106	-225.2	11.8	-232.27	9.82
3NRW	104	-223.9	9.0	-231.49	5.54
3NZL	73	-146.1	9.6	-149.13	6.29
4ESN	78	-150.8	7.6	-156.07	5.53
5CRO	60	-126.4	4.3	-129.93	4.1
Mean		-173.8	7.5	-178.6	6.5

for 1HZ6 that had three regions where the lowest energy models were located however NEFILIM new run sampled only on the region near to the native structure. It was explored how the multiple peaks vanished in the energy landscape of NEFILIM new run. In order to understand, NEFILIM was run with the fragments obtained without filtering using APRDS and the energy landscape again contained the same three peaks. This indicates that APRDS can be useful to vanish the distant fragments. The lowest energy model achieved the improvement of CA-RMSD 0.50 Å in NEFILIM for 1HZ6 but a large energy barrier still existed between the relaxed native structures and predicted models.

An analysis showed that lacking of better fragments for C-terminal residues could be the cause for this energy barrier (Figure 4.8). NEFILIM improved accuracy of the lowest energy model by 0.91 Å and energy landscape was more funnel-shaped for protein 1BM8. However, the lowest energy predicted models were not predicted closer to the relaxed native structures (Figure 4.8). Therefore, an energy gap was clearly seen between the relaxed native models and the lowest energy models.

The energy function has gained more discriminative power in quality assessment. NEFILIM predicted more accurate models with lower Rosetta all-atom energies. The lowest energy was improved to -185.82 (NEFILIM) from -182.07 (Rosetta run) on average for 30 targets. NEFILIM significantly improved Rosetta all-atom energies for 1ACF and 3NRW by about -10.00 with accuracy. The scatter plot of Rosetta energy versus CA-RMSD to native structure for 1ACF and 3NRW clearly showed the discrimination between native and non-native structures (Figure 4.8). More accurate models are enriched in the lowest energy regions in NEFILIM experiment than Rosetta experiment. In order to demonstrate the enrichment of good models with lower energies, two hundred lowest energy models were selected from the pool and averaged their energy and accuracy. Rosetta showed the average Rosetta energy and accuracy -173.8 and 7.5 Å respectively for 30 targets (Table 4.3). These scores were improved in the NEFILIM run to -178.3 (Rosetta energy) and 6.5 Å (CA-RMSD) respectively. This result suggests more accurate fragments increased the sampling ability near to the native region with lower energies.

Table 4.4 Comparison of best models predicted in NEFILIM initial and new runs

PDB	NEFILIM initial run (Lowest energy models)				NEFILIM new run (Lowest energy models)			
	Rosetta energy	CA-RMSD	TM Score	GDT-TS	Rosetta energy	CA-RMSD	TM Score	GDT-TS
1A19	-191.1	2.5	0.71	0.73	-192.7	6.8	0.46	0.47
1A68	-197.7	6.7	0.44	0.45	-202.7	9.0	0.45	0.45
1AAR	-164.9	5.2	0.45	0.52	-170.2	2.3	0.84	0.85
1ACF	-261.5	5.7	0.58	0.53	-272.5	3.2	0.69	0.63
1AIU	-217.5	2.9	0.80	0.80	-225.4	1.5	0.91	0.90
1BM8	-222.6	4.1	0.61	0.60	-234.3	3.0	0.78	0.79
1BQ9	-103.1	2.0	0.74	0.82	-107.0	1.8	0.74	0.83
1CC8	-147.5	4.4	0.49	0.59	-147.2	3.6	0.50	0.61
1CTF	-154.9	3.6	0.55	0.63	-155.3	3.3	0.56	0.67
1DI2	-152.0	1.9	0.80	0.85	-155.6	0.9	0.92	0.95
1DTJ	-159.2	2.7	0.78	0.81	-162.1	2.2	0.83	0.86
1HZ6	-136.1	3.5	0.54	0.66	-137.3	3.1	0.60	0.73
1IG5	-173.6	3.3	0.63	0.68	-176.3	2.2	0.73	0.77
1IIB	-216.6	11.8	0.54	0.51	-221.3	7.9	0.63	0.62
1KPE	-218.2	7.3	0.48	0.42	-223.8	8.2	0.35	0.32
1OPD	-176.8	5.4	0.50	0.53	-183.3	4.6	0.49	0.52
1PGX	-152.4	5.4	0.69	0.73	-155.4	5.3	0.79	0.81
1TIG	-189.4	5.4	0.47	0.53	-194.2	3.0	0.67	0.68
1UBQ	-174.2	2.1	0.90	0.93	-173.7	1.5	0.88	0.90
1WD6	-198.0	12.7	0.27	0.33	-188.8	9.5	0.27	0.31
2BC5	-252.5	1.9	0.85	0.81	-257.2	1.6	0.89	0.86
2CI2	-135.7	7.7	0.39	0.44	-142.5	7.6	0.42	0.45
2H28	-212.3	14.8	0.31	0.28	-221.6	9.7	0.43	0.40
2O6K	-161.4	5.1	0.52	0.59	-168.4	10.0	0.45	0.47
2VSV	-163.5	6.7	0.48	0.52	-164.6	9.4	0.40	0.42
3GWL	-232.0	9.2	0.49	0.47	-238.2	9.7	0.39	0.39
3NRW	-227.6	5.2	0.73	0.67	-241.1	1.9	0.83	0.82
3NZL	-150.5	7.1	0.39	0.46	-154.7	5.1	0.44	0.52
4ESN	-155.9	6.1	0.28	0.32	-164.2	5.0	0.33	0.38
5CRO	-130.0	3.2	0.82	0.87	-133.5	3.4	0.83	0.88
Mean	**-180.9**	**5.5**	**0.58**	**0.60**	**-185.51**	**4.9**	**0.62**	**0.64**

4.3.3. Improved performance in resampling

In the NEFILIM initial run, the sequence-derived fragments were used to generate $1.2X10^5$ models. In this run, the accuracy was produced on average 5.57 Å in CA-RMSD on the benchmark of 30 proteins from the native structures when lowest energy model was taken as best prediction. This CA-RMSD was improved to 4.88 Å in the NEFILIM new run (Table 4.4). NEFILIM new run performed better by 4% when TM-score and GDT-TS assessed the model quality (Table 4.4).

Another criteria were also employed for comparisons in which the best in five lowest energy models were selected. Their average CA-RMSD was improved from 4.13 Å (NEFILIM initial run) to 3.95 Å (NEFILIM new run) (Table 4.5). However, the performance was dropped to 2% increment from 4% when TM-Score and GDT-TS assessed the best in top five lowest models (Table 4.5). Still, the performance measured using GDT-TS and TM-score was better than Rosetta run.

Sixteen proteins showed accuracy better than 3.5 Å in CA-RMSD when the lowest energy models were selected. Their average CA-RMSD was improved from

Table 4.5 Comparison of best in top five models generated in NEFILIM initial and new runs

PDB	NEFILIM initial run (Best in top five)			NEFILIM new run (Best in top five)		
	CA_RMSD	TM Score	GDT-TS	CA_RMSD	TM Score	GDT-TS
1A19	2.3	0.74	72.8	6.2	0.46	50.0
1A68	6.7	0.49	51.4	6.4	0.45	47.7
1AAR	4.7	0.54	60.2	1.3	0.90	91.5
1ACF	5.1	0.66	62.2	3.0	0.73	66.2
1AIU	2.0	0.83	80.0	1.5	0.91	90.0
1BM8	3.5	0.68	66.7	3.0	0.78	78.8
1BQ9	1.8	0.78	85.4	1.8	0.75	82.1
1CC8	3.1	0.68	72.6	3.6	0.59	64.2
1CTF	2.7	0.64	71.3	3.3	0.56	64.3
1DI2	1.3	0.86	90.2	0.9	0.93	95.7
1DTJ	2.7	0.85	87.2	2.2	0.84	86.8
1HZ6	3.5	0.54	66.0	3.1	0.60	72.5
1IG5	2.4	0.76	78.7	2.2	0.73	77.0
1IIB	2.8	0.75	73.5	2.1	0.80	77.2
1KPE	7.3	0.48	42.0	8.2	0.35	32.1
1OPD	4.3	0.50	53.2	4.1	0.50	55.3
1PGX	3.3	0.79	82.1	5.2	0.79	81.1
1TIG	3.2	0.65	67.1	2.8	0.71	72.4
1UBQ	1.4	0.93	95.1	1.5	0.90	93.1
1WD6	3.8	0.50	50.9	8.6	0.28	31.4
2BC5	1.3	0.91	88.9	1.5	0.89	86.8
2CI2	6.5	0.44	48.1	6.4	0.48	51.9
2H28	10.1	0.40	40.4	9.7	0.43	40.4
2O6K	5.1	0.52	58.7	2.3	0.72	75.7
2VSV	6.7	0.48	51.5	9.1	0.44	46.3
3GWL	8.1	0.49	47.4	5.5	0.55	55.4
3NRW	3.9	0.73	67.3	1.7	0.84	82.5
3NZL	7.1	0.41	49.7	3.4	0.58	62.7
4ESN	5.1	0.31	38.8	4.3	0.36	40.7
5CRO	3.2	0.82	87.1	3.4	0.83	88.3
Mean	4.2	0.64	66.2	0.7	0.66	68.0

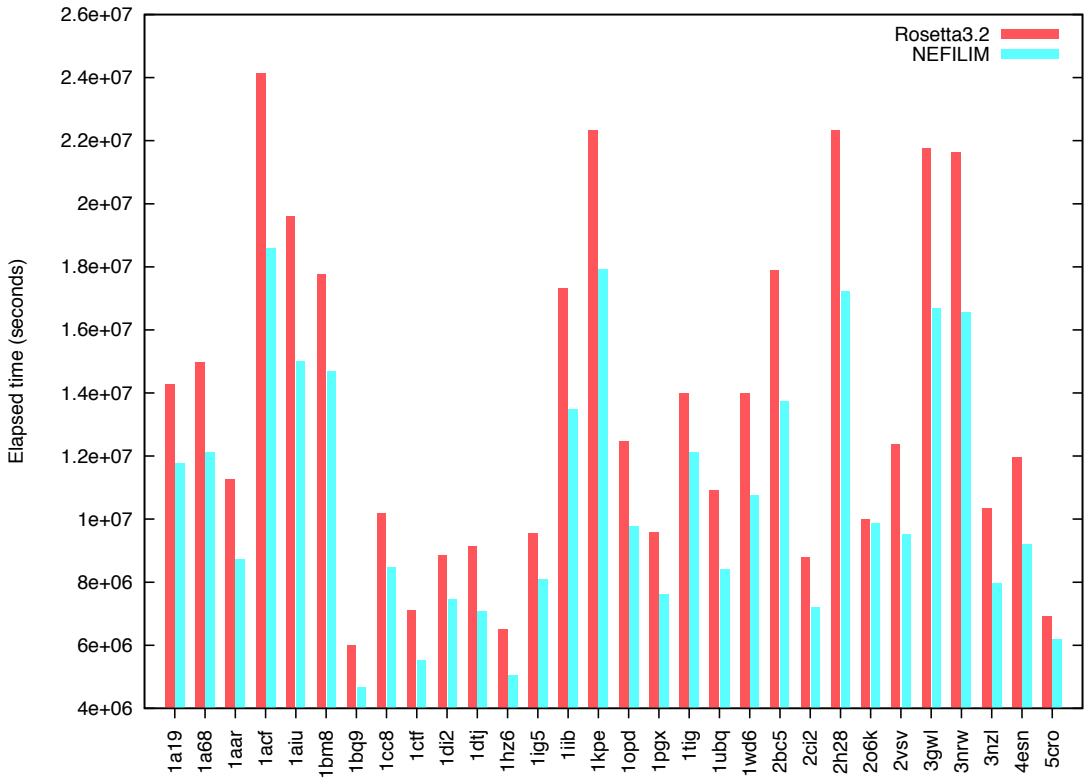

Figure 4.9 Total time spent by Rosetta and NEFILIM for each target

3.5 Å (NEFILIM initial run) to 2.6 Å (NEFILIM new run). Furthermore, the improvement in CA-RMSD to the native structure for 7 proteins was more than 3.5 Å (Table 4.1). Similarly, when the best in five models are used to evaluate the prediction performance, twenty-one proteins have CA-RMSD lower than 3.5 Å (15 proteins have CA-RMSD better than 3.0 Å). These twenty-one proteins showed the accuracy 3.2 Å CA-RMSD (initial run) and 2.7 Å (new run) on average (Table 4.4). The accuracy performance significantly dropped for two proteins 1A19 and 1PGX in the new run (Table 4.4). For protein 1A19, the CA-RMSD degraded from 2.5 Å to 6.8 Å when lowest energy model was taken and from 2.3 Å to 6.2 Å when the best in five models was considered for performance evaluation.

The total elapsed time spent for both simulations was also computed. This is considered as another criterion of performance evaluation. NEFILIM spent the time to generate fragments and the models in the initial and new runs, which is the total elapsed time required in this strategy for each protein. NEFILIM generated 1.2E+05 models in the initial run and 3E+04 models in the new run. The fragment generation time includes elapsed time spent in clustering and APRDS computation. I considered the time required to generate 2E+05 models as total elapsed time in a Rosetta run. The

total elapsed time spent by both methods was computed and compared. NEFILIM spent less elapsed time 8.93E+04 CPU core hours compared to Rosetta (1.12E+05 CPU core hours). The time taken by each target for NEFILIM and Rosetta is shown (Figure 4.9).

4.4. Discussion

De novo protein structure prediction using fragment assembly reduces the search space by using fragments from experimentally determined structures. The information provided in the fragments, therefore, influences the conformational search. This work focuses the study about conformation sampling near the native region using better quality fragments. The new fragments are generated from the lowest energy models predicted in the initial run in order to search the conformational space near the native region. Many models of the most tested proteins appeared closer to the native structure with lower energies using the new fragment library.

Unbiased trajectories in *de novo* structure prediction can be used to converge the models towards the global minima and to avoid them being trapped in the multiple local minima. Usage of sequence-derived fragments in structure prediction can search a larger conformational space. Despite the diversity of fragments in the sequence-derived fragments, the fragments, which guide the search procedure to converge toward deep local minima in conformational space, are often selected in the sampling using simulated annealing. These fragments, which appeared in the lowest energy models, are presumably native-like fragments. The obtained information from the models was exploited to guide conformation sampling near the global minimum. The information underneath in these models was converted into the fragments for next-round of model generation. Therefore, this method bears the similarity with other resampling methods that use the information achieved from the previous iteration in the subsequent iterations of structure prediction. However, this method has unique features that distinguished it from other methods. Most of the other resampling methods seek to generate the fragments efficiently from the sequence-derived fragment library using statistics on feature space (Blum, et al., 2010), function model (Brunette and Brock, 2005), fragment distribution (Simoncini, et al., 2012), or torsion angle distribution (Li, et al., 2008). However, NEFILIM generated fragments were not only enriched with better fragment from sequence-derived library but also contained more accurate fragments from *de novo* models.

Clustering was used in identifying the accurate fragments from representative *de novo* models generated in the initial run. The fragments in large clusters represent the several distinct instances of local conformations adopted in the lowest energy models. One of the clusters may capture the instance of the native conformation. However, in this procedure, top five clusters were chosen in the fragment selection in order to maintain the diversity in the new fragment library. The cluster radius might have significant impact on the composition of the new fragments in the library. As observed in the experiment, the number of clusters is mostly dependent on the secondary structure element. Helical fragments are often clustered densely and fragments with loops are sparsely distributed among the clusters. This suggests that the cluster radius can be chosen based on secondary structure element but this is complex because each target needs the multiple cluster radii. Moreover, secondary structure elements in the predicted models were varied. Instead, the cluster radius was determined by training the fragments of global proteins. This fragment generation procedure was carried out using a single clustering radius.

Two cluster radii were set based on fragment sizes (0.20 Å for three-residue fragments and 1.0 Å for nine-residue fragments) for this experiment. In order to determine the precise clustering radii, the models that are the densely and sparsely distributed near the native region were chosen. Indeed, the densely clustered models show the better prediction accuracy when a smaller clustering threshold is used. For example, smaller clustering threshold performed better for targets such as 2BC5 where many lowest energy models appeared near to the native structures. This is because smaller threshold finely put the representative fragments in top five clusters and one of these clusters contains the candidate fragments as in the native structures. However, this selection procedure does not work when the lowest energy models are distributed in the large conformation space even forming the multiple peaks in the energy landscape. A larger cluster radius might be appropriate for sparsely sampled lowest energy models. In the training set, the protein targets 1CTF and 1BM8 showed sparse distribution between selected lowest energy models and the accuracy. In these cases, a large clustering threshold could gather better fragments, which are distributed sparsely, in top five clusters. This type selection helps to select the rarely appeared better fragments in the new fragment library.

Those in majority dominated the candidate fragments in new fragment libraries because clustering was used for new fragment generation. Many fragments that adopted frequently in the lowest energy models are presumed native-like fragments. When the sequence-derived fragments libraries contain the low quality fragments, the new fragment library have also small probability to contain the better fragments unless conformation sampling significantly improved the accuracy of predicted models and their fragments. This was seen for few targets. The new fragment library, sometimes, contains the poor quality fragments such as in 1A19 (1K-10I and 24L-29Y) and produced the inaccurate models than in the initial run.

The energy versus CA-RMSD scatter plot became smoother in the new run than the initial run. Many near-native models were sampled with low energy near the native structures (Figure 4.8). Sampling with structure-derived fragments removed the multiple peaks observed in the energy landscape for a few proteins. Structure-derived fragments also produced more accurate lowest energy models than sequence-derived fragments but the accuracy has not reached the level of relaxed native models. Therefore, energy discrepancies exist between the predicted models and relaxed native models for those cases where energy function accurately guides the conformation sampling. Absence of native-like fragments for some residues created the energy barrier between the lowest energy models and the relaxed native models such as in 1BM8 and 1CTF.

NEFILIM generated inaccurate models (more than 3.5 Å CA-RMSD) for 1KPE, 2H28, 1EW4, and 2VSV because the selected lowest energy models in the initial run did not contain good models (< 3.5 Å) in the majority. Fragment generation misses the rarely sampled better fragments and includes the densely sampled worst fragments. These fragments did not focus sampling towards the native conformational space because torsion angles of selected fragments did not match with that of the target structures. This approach, by design, concentrates the sampling near the conformational space where the lowest energy models are sampled in the initial run. Therefore, torsion angle information provided by the selected lowest energy models plays pivotal role to forecast the results in the new run. This method also misguided the sampling when many lowest energy models were predicted inaccurately in the initial run, such as in 2CI2, 1AIU, 1IIB, 1A19 and 1CC8. In these cases, although many residues positions achieved the fragment quality improvement, subsequently

prediction quality was degraded due to loss of better fragments for a few residue positions. These better fragments rarely appeared in the lowest energy models and subsequently vanished.

Structure-derived fragments improved the model quality assessment using Rosetta energy in most tested targets. However, the best prediction was not selected from the prediction pool for some targets using the Rosetta energy. The inaccuracy in the energy function might be the cause (Das, 2011). Structure-derived fragments generated accurate models (better than 2.0 Å) in the pool for 13 out 30 proteins but for 9 cases, these accurate models did not adopt the lowest Rosetta energy. Therefore, Rosetta energy was not able to identify these models as the best predicted models in these targets. Out of these 9 targets, Rosetta energies of the relaxed native structures of 7 targets (1IG5, 1DI2, 1DTJ, 1PGX, 2O6K, 2BC5, and 5CRO) overlapped with or became worse than that of predicted models. In these targets, accurate model prediction cannot be possible by providing better quality fragments and by increasing the number of iterations for conformational sampling. This was one of the major reasons of stopping the simulation after the second iteration. Here is another reason for stopping the simulation after the second iteration. The information (torsion angles) propagated from the initial prediction to structure-derived fragments was inaccurate for some residues for better prediction. Fragments for these residues that do not have correct torsion angles like in the native structures after the initial run are difficult to improve. In this case, improving the fragments only for certain residues does not substantially assist to predict near-native models. This was observed for protein targets 1CTF, 1IIB, 1HZ6, and 1OPD respectively.

4.5. Conclusion

This study improves the accuracy of fragments using initially predicted lowest energy models and then uses these new set of fragments for the next-round of prediction in order to generate the better models. The accuracy of nine-residue fragments is improved for 77% cases. The fragments, which were adopted in the lowest energy models, were gathered using clustering algorithm and the most frequently occurring fragments were selected as new fragments. Therefore, new fragment libraries (new three-residue and nine-residue) were enriched with better fragments. Fragment quality was improved by 0.47 Å for nine-residue fragments. The experimental result shows that these improved fragments predict the lowest energy

models closer to the native structure. The accuracy performance shows the success in better model prediction on average for a benchmark of 30 targets from 5.99 Å to 5.03 Å for the lowest energy model and 4.06 Å to 3.58 Å for the best in five models as compared to Rosetta. The accuracy better than 3.5 Å CA-RMSD was further analyzed. This was observed in 50% tested targets and their average accuracy was improved from 3.59 Å to 2.37 Å when the lowest energy models were considered as the best prediction. Furthermore, the success of new run over initial run was also measured and their CA-RMSD of the lowest energy model and the best in five predictions was improved, on average, by 0.68Å and 0.19Å respectively.

Chapter 5. FRAP – *ab initio* phasing with *de novo* fragments for difficult targets

5.1. Objective

The most widely used computational tool for phasing of diffraction pattern of protein crystal is MR. It requires the search model that should have structural similarity with the target structure to locate the placement in the unit cell. Advancement in bioinformatics for sequence alignment (Altschul, et al., 1997) and development of comparative structure modeling provides the robust tools to identify the suitable search model for MR (Marti-Renom, et al., 2000). The utility of MR in solving phase problem is also due to the ever-increasing number of protein structures deposited in PDB (Berman, et al., 2002). Despite all these successes, there are numerous sequences that do not have homologous structures. MR cannot be used to determine the protein structure in these cases. However, these problems can be tackled using the computationally predicted models as searched models. Indeed, accuracy achieved by recently developed methods for protein structure prediction using only sequence information increases the utility of the MR.

Rosetta, one of the principle methods, demonstrated to solve the crystallographic phase problem using the *de novo* models (Qian, et al., 2007). The predicted *de novo* model for natural protein using Rosetta achieved the success in MR experiment (Qian, et al., 2007). These models for large dataset were further extensively tested for phasing and the model accuracy appeared as major constraint in successful phasing (Das and Baker, 2009). In addition, computational power substantially increased the success rate in MR (Das and Baker, 2009). However computational power was efficiently managed by incorporating phasing program in the course of *de novo* structure prediction refinement for large cluster without compromising on model accuracy (Shrestha, et al., 2011). The accuracy of search model has been improved by trimming wrongly predicted regions (Bibby, et al., 2014; Rigden, et al., 2008), or resampling more on error-prone residues (Shrestha, et al., 2012). The maximum likelihood target functions introduced in Phaser has increased the sensitivity of MR searches (McCoy, et al., 2007). The increased sensitivity in MR searches can correctly identify the location of fragments in the unit cell (Rodriguez, et al., 2009; Sammito, et al., 2013).

Phasing using MR generally requires closer search model to the target structure either for homologous protein or *de novo* model. Although the quality of search model required for MR significantly relied on the targets, it should be generally within 3.0 Å CA-RMSD from the target structure. The methods (conformational sampling and energy function) used in *de novo* modeling have successful made the suitable search model for MR from distant homologous proteins, NMR models (Qian, et al., 2007). Furthermore, electron density map guided energy-optimization has improved the quality of search model for difficult targets and obtained the final models (DiMaio, et al., 2011). However, *de novo* modeling is still practically challenging to produce the accurate search model independently from amino-acids sequence due to conformational sampling and energy function. Therefore, the utility of MR with *de novo* models is still far away from routine. When *de novo* models were predicted with accuracy low-quality (3.0–4.0 Å CA-RMSD or beyond), the models have unlikely to be used in MR for phasing. In the absence of accurate search models, the concept of using idealized alpha helical fragments (Rodriguez, et al., 2009) was further explored and implemented for *de novo* models. Therefore, instead of improving the accuracy of full-length models, these models were broken into many smaller fragments for phasing.

In this study, a new method was introduced for protein structure determination using MR with template from low-quality *de novo* models. This approach uses the fragments from low-quality *de novo* models. The best predicted *de novo* models for these targets were low-quality and insufficient of MR. These selected targets were previously unsuccessful and considered as difficult for *de novo* modeling. These difficult targets include the alpha, beta, and mixed alpha-beta proteins. This method breaks the *de novo* models into constant-length overlapping fragments, clusters fragments, selects the representative fragments, uses these fragments independently for phasing, phases the each fragments to identify the correct places and assembles the phased fragments using crystallographic operators to obtain the final models. In this study, the method was tested with ten difficult targets and best-predicted full-length *de novo* models were unable to provide the phases. The results from our method showed phasing using MR with challenging targets solve for 80% of cases

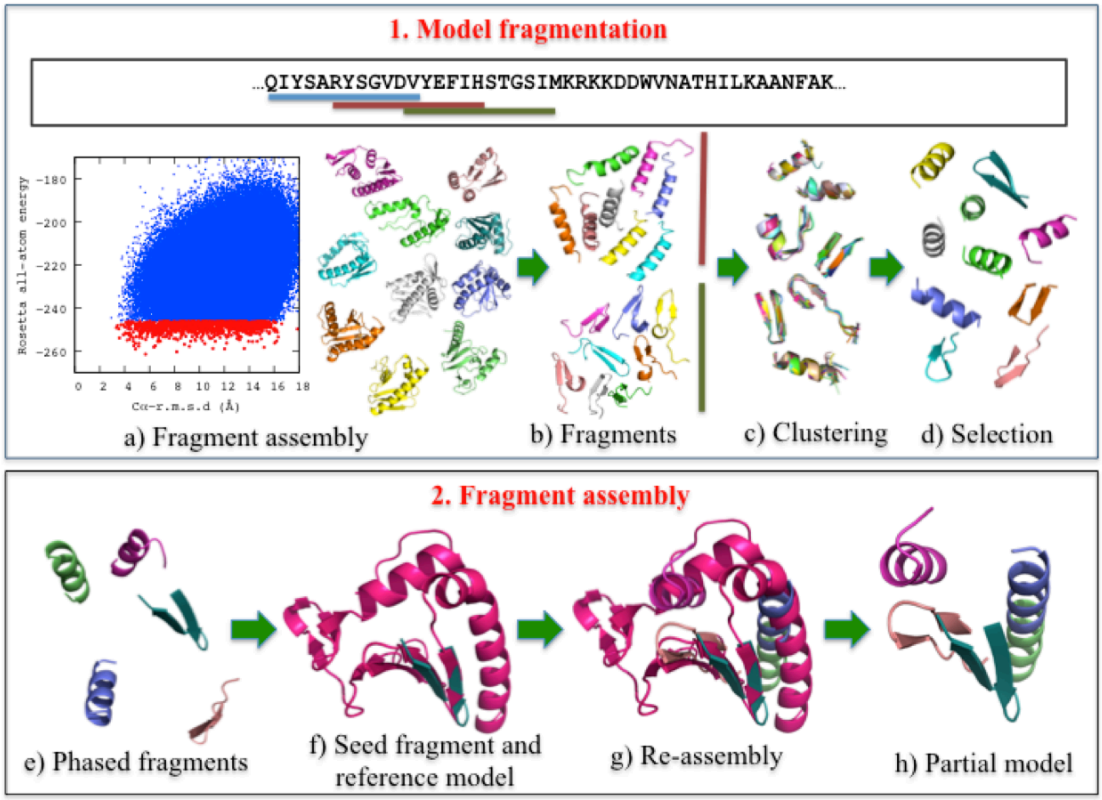

Figure 5.1 Schematic diagram of FRAP

5.2. Methods

This method mostly includes the *de novo* model generation, substructure generation, independently phased the substructures, and assembled substructures to obtained the final models respectively (Figure 5.1). The substructure was called as the fragment in this work. The overlapping constant-length fragments were constructed from selected low-energy *de novo* models generated using Rosetta. *De novo* models were generated that were broken into fragments for each residue position. The representative fragments were selected after clustering of fragments and then phased independently using Phaser. The fragments selected after MR were assembled together in order to place in the unit cell. The partial models after fragment placement in the unit cell were significantly closer to the native structure than best predicted *de novo* models. Therefore, the phase angles from partial model were sufficient to determine final structure using automated model building program. The algorithm was implemented using C++ programming language and called as FRAP that stand **FR**agment **A**ssembly **P**hasing.

5.2.1. Benchmark data selection

Ten proteins of different topologies (all alpha, all beta, and alpha + beta topologies) were selected as benchmark dataset to test the programs. Few of these proteins were collected from previous studies (Shrestha, et al., 2011; Shrestha, et al., 2012). Importantly, these targets were considered as the hard targets for phasing because best-predicted *de novo* models were unable to provide the accurate phases using MR. Therefore, the criterion set for selection of each target was whether full all-atom models provided the solution in phasing experiment using Phaser. In target selection process, Rosetta was used to generate 1.2E+05 all-atom models for each protein of benchmark data using amino-acids sequence and two types of fragment libraries. Three-residue and nine-residue fragment libraries were generated from Robetta server (Chivian, et al., 2003). As the representative models, 1000 low-energy models were selected using Rosetta energy function and then given to the phasing program. The solution of each of 1000 models was verified. If MR was unable to place the model in asymmetric unit of the unit cell, the target was selected to test the FRAP.

5.2.2. *De novo* fragments generation for molecular replacement

One thousands lowest energy models that cannot be used as search model for MR were selected based on Rosetta all-atom energy. These models were fragmented into many overlapping constant-length fragments. Three different types of constant-length fragments (thirteen-residue, seventeen-residue and twenty-one-residue) were independently generated. There could be many candidate fragments as search model in MR but few representative fragments were only selected due to the computational complexity. In order to select the representative fragments, fragments for each residue position were clustered and two hundred fragments were randomly picked from the top ten clusters. One of the fragments taken from the largest clusters can be instance fragment of the native structure that can be suitable search model for MR. This fragment selection procedure for MR was same as in the previous study (Shrestha and Zhang, 2014). The fragments selected using Rosetta energy were subjected to phase with MR program Phaser (McCoy, et al., 2007). For each fragment, the entire protein sequence was used and as the model error in the Phaser, CA-RMSD 0.50 Å was set for each fragment.

5.2.3. Fragment assembly after molecular replacement

The placement of fragments together in the asymmetric unit in the absence of native structure is challenging and important. The independently placed fragments after MR cannot be used in structure determination due to the ambiguities in the permissible origins. Therefore, it is necessary to make same permissible origins to obtain phase. This can be done using either real-space approach or reciprocal-space approach. A real-space approach was used to assemble the phased fragments and these are the steps. One fragment from the pool of phased fragments was selected that was termed as seed fragment. The fragment that has high TFZ or LLG score can be candidate for seed fragment. In addition, sometimes, secondary structure of the fragment was another key in the seed fragment selection. Afterward, full-length *de novo* model of the target sequence was also randomly selected from the prediction pool. This model is often low-energy *de novo* models and called as reference model. The reference model was superimposed to the seed fragment using rigid body transformation (Kabsch, 1976) imposing the residues that covered the seed fragments.

If seed fragment and reference model were close to each other, the other fragments were placed based on the seed fragment and reference model. The fragments were rotated and translated using provided crystallographic operators. All the crystallographic operators provided by permissible origin, symmetry operators allowed for space group including unit cell translation were employed on each fragment. For each operation, the *Euclidean* distance of candidate fragment to the seed fragment and specific portion of the *de novo* model were computed. After computing all the combinations, the minimum distance were kept as the correct solution for given fragments. While identifying correct location for the given fragment, the numbers of clash to the already kept fragments were also used to filter out wrong solutions and overlapping fragments. The partial models were obtained after fragment assembly but the phases obtained was sufficient to construct the final models using automated model building method using Phenix (phenix.autobuild) without manual intervention.

Fragment assembly after phasing of fragments using MR for polar space group was complex since permissible origins were not specified for these space groups. Their origins are related by translation along the polar axes. In order to solve

the origin conflict problem in polar space group, fast translation function implemented in Phaser was executed.

5.3. Result and Discussion

5.3.1. Seed fragment and reference model

Successful fragments in MR could also be placed at different unit cells that were related by crystallographic operators. Without the reference, all the fragments are impossible to assemble together at the same unit cell because of symmetry, permissible origins, and unit cell. Therefore, the reference point was at first selected. Seed fragment and reference models were the starting point of fragment assembly in this experiment. Seed fragment and reference model determine the location of other fragments. The selection of the seed fragments is challenging because seed fragment determines the location of other fragments. Importantly, if the seed fragment is placed at incorrect location, other fragments cannot be placed at correct position though these fragments might be correctly placed. The fragments with high TFZ and LLG scores were taken as the seed fragments (Figure 5.2). Seed fragments often showed the TFZ scores more than 7.0 except for the molecules with the polar space group and their LLG scores were also high (Figure 5.2). Seed fragments were often observed either α-

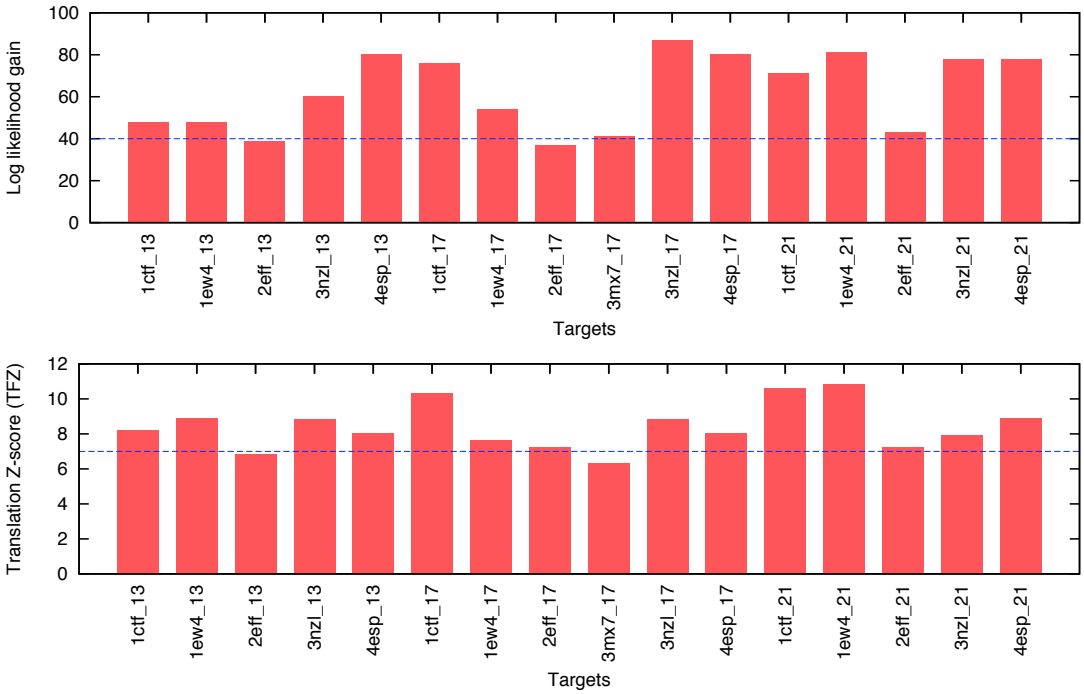

Figure 5.2 Quality of seed fragments measured by LLG and TFZ scores

Table 5.1 List of benchmark dataset and their MR result

SN	Targets	Resolution	Space group	Sequence length	SCOP Classification	CA-RMSD/MR-CA-RMSD
1	1OPD	1.50	P1	85	α + β	2.78/19.42
2	1CM3	1.60	$P2_1$	85	α + β	2.72/14.79
3	1EW4	1.40	$P3_221$	106	α + β	4.98/8.57
4	2EFF	1.80	$P3_221$	106	α + β	4.99/21.44
5	3O55	1.90	$C222_1$	119	α	6.18/19.17
6	3NZL	1.20	$P2_12_12_1$	73	α	3.48/20.91
7	1CTF	1.70	$P4_32_12$	68	α + β	3.16/10.44
8	1MB1	2.10	$P4_12_12$	98	α + β	2.43/18.31
9	4ESP	1.10	$P4_12_12$	130	α + β	5.58/16.59
10	3MX7	1.76	$P3_121$	90	β	3.40/18.48

helices or anti-parallel β-strands. For seventeen-residue fragments, seed fragments were α-helical for five proteins (1CTF, 1CM3, 1OPD, 3NZL, and 4ESP) and anti-parallel β-strands for three proteins (1EW4, 3EFF, and 3MX7).

Because polar space group does not have predefined permissible origins and provides the ambiguities of translation vectors, the seed fragments cannot be directly used to assemble the fragments. Therefore, the fast translation function was used in order to solve the ambiguities in permissible origin translation. Afterward, seed fragments were selected for polar space groups after fast translation function (FTF). This experiment contains two polar space groups (P1 and $P2_1$).

5.3.2. *De novo* fragments and molecular replacement

Poor quality *de novo* models cannot be considered as suitable search template for MR (Table 5.1). Many methods have been developed to improve the accuracy of models in order to make suitable search template. The quality of model was harder to improve when majority of predicted models exceed the CA-RMSD beyond limit of correct fold. One of the *de novo* modeling programs, Rosetta, predicts local substructures accurately although global structures may not have the correct fold. Therefore, identifying and using the fragments as starting template for MR could

Table 5.2 Phasing result with different fragment size

Targets	Thirteen-residue residues)			Seventeen-residue fragment			Twenty-one-residue fragments		
	CA-RMSD	R-/R-free factor	CA-RMSD	CA-RMSD	R-/R-free factor	CA-RMSD	CA-RMSD	R-/R-free factor	CA-RMSD
1OPD	1.03	0.26/0.31	0.68	0.99	0.26/0.27	0.76	1.06	0.28/0.33	0.83
1CM3	1.03	0.30/0.36	1.93	1.45	0.29/0.34	1.42	-	-	-
1EW4	0.68	0.34/0.37	0.88	1.98	0.31/0.34	1.20	1.73	0.29/0.33	1.34
2EFF	2.31	0.38/0.42	2.41	1.4	0.34/0.37	1.15	1.34	0.32/0.36	1.21
3O55	-	-	-	-	-	-	-	-	-
3NZL	1.45	0.31/0.35	1.15	1.76	0.35/0.39	1.35	2.51	0.36/0.39	1.85
1CTF	1.44	0.31/0.34	0.09	1.66	0.29/0.34	1.46	1.86	0.27/0.31	1.41
1MB1	-	-	-	-	-	-	-	-	-
4ESP	1.38	0.35/0.33	0.28	1.77	0.34/0.34	1.22	1.92	0.33/0.34	0.49
3MX7	-	-	-	1.55	0.35/0.39	1.43	-	-	-

be the promising approach to solve the phase problem of poor quality *de novo* model. These fragments generated from predicted low-quality models may provide the useful information for phasing. However, the critical issue is how the suitable fragment can be generated for search models. For simplicity, the constant-length overlapping fragments from lowest energy *de novo* models - thirteen-residue, seventeen-residue, and twenty-one-residue were constructed. Three types of fragment of ten difficult targets were tested independently (Table 5.2).

Seventeen-residue fragments succeeded in providing the final solution for eight from ten targets in MR experiment. Similarly, thirteen-residue and twenty-one-residues fragments succeeded in seven and six targets respectively. As the success rate decreased with less and more residues from seventeen-residue fragment, seventeen-residue fragment was most suitable search template for MR in our experiment. Six successful proteins were common in three types of fragment. Fragments from twenty-one-residue of proteins (PDB ID 1CM3 and 3MX7) failed to provide the accurate phases to construct the final model using automate model building. The fragments generated for 3MX7 using thirteen-residue was also inaccurate to obtain final models. MR experiment for proteins with PDB ID 3O55 and 1BM8, were unsuccessful using three types of fragments.

The fragments, which are α-helical and β-strands with short loops, obtain their location in unit cell in MR experiment with high LLG and TFZ. Although there are many examples, one example for PDB ID 1EW4 was provided. The crystal structure of 1EW4 contains five antiparallel beta strands and two α-helices. The most accurately predicted model showed the 4.98 Å CA-RMSD to the native structure. Full-length all-atom models failed in MR experiment so that the experiment was started with fragments. Anti-parallel beta strands (thirteen-residue fragment) with small loop from predicted *de novo* models showed high scores (TFZ=8.9 and LLG=48). Similarly, another N-terminal anti-parallel β-strand (thirteen-residue fragment) was also correctly located. Both fragments showed high scores after MR. Phaser enabled to put three strands (D29-N52) out of five β-strands. Similarly, Phaser identified the places of two fragments (N2-W14 and D11-D23); their LLG scores are 55 and 38 and TFZ scores are 5.9 and 6.2. These fragments matched with long α-helices (twenty-one residues) of native structures. Two fragments were placed at the location of another α-helix starts from 87 to 99 with loops at both ends. Using phases

obtained from short fragments, the final models with incomplete residues were constructed and this partial structure includes all secondary structure elements existed in the native structure.

FRAP has taken the advantage of correctness of secondary structure elements in the *de novo* models and large structure different between models with target structure due to mis-alignment. Here are the examples (1CTF and 1EW4) to show how FRAP enables to solve the low-quality models in MR using fragments (Figure 5.3). The best *de novo* models contained error 3.16 Å (1CTF) and 4.98 Å (1EW4) to the native structure. The error was reduced 1.46 Å (1CTF) and 1.20 Å (1EW4) in the models using FRAP. The error was accumulated in *de novo* models because some secondary structure elements were arranged wrongly. Therefore, these secondary structure elements were individually compared with that of native structure. For 1CTF, α1- helix (A63-G77) differed by 17.8º rotation with best *de novo* models and this angular difference was reduced to 4.7º when partial model generated by FRAP was used (Figure 5.3). The improvement was also observed in α2- helix (G79-A90). The orientation of α2- helix was very close in assembled structure using FRAP (2.7º) than that in best *de novo* models (29.9º). The improvement was also

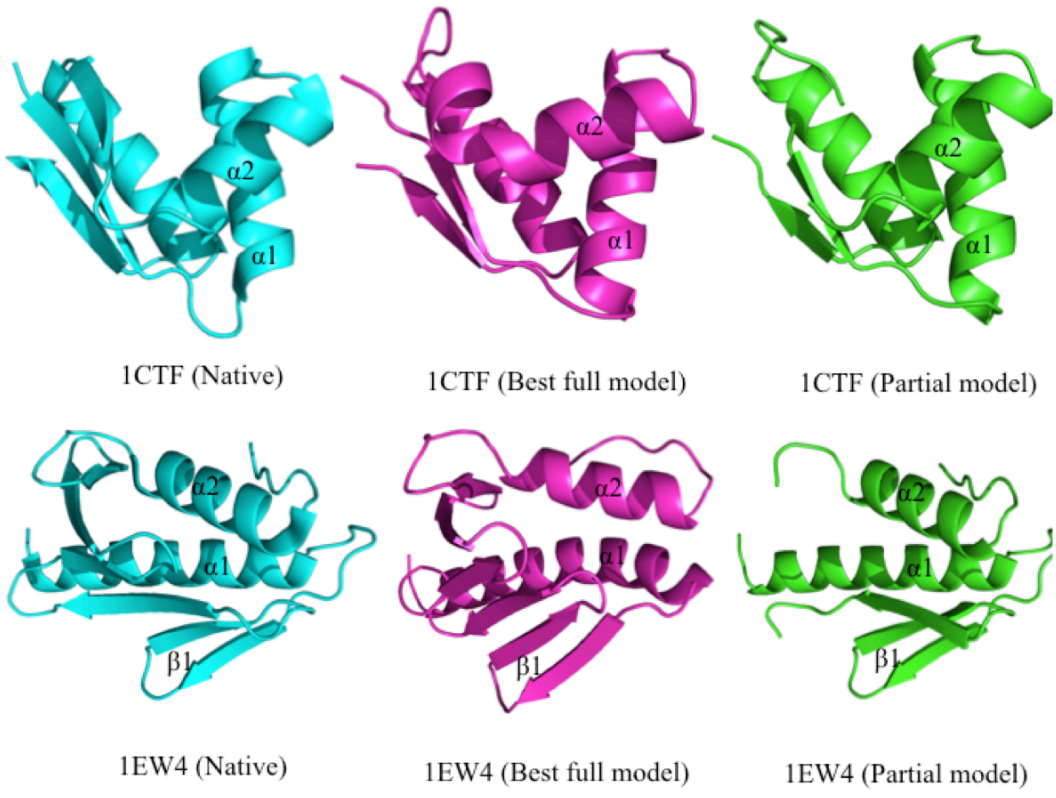

Figure 5.3 Orientation of best-predicted *de novo* model and model after FRAP

observed in another target 1EW4 (Figure 5.3). In this target, the orientation difference was measured for three different secondary structures (α1- helix, α2- helix, and β1-hairpin). The orientation of α1- helix (M1-W24) was differed by 23.2º and this was reduced to 0.22º using partial model. The improvement was continuously observed in α2-helix (T86-G100) from 42.4º to 0.24º, and in β1- β-hairpin (I30-F43) from 25.5º to 0.19º respectively. FRAP extracted sufficient information for structure determination from wrongly oriented regular secondary structure of *de novo* models. This cannot be solved using alternative approaches such as trimming of loop regions employed for homologous proteins (Stein, 2008), identifying conserved core regions (Bibby, et al., 2014) and others.

5.3.3. Fragment assembly

Although many fragments were independently located at correct places, these fragments were necessary to be at same unit cell to obtain the phases required for structure determination. Therefore, all correctly placed fragments were necessary to bring together in the unit cell. The fragment assembly phasing was started with seed fragment and reference model. The reference models were often low quality. FRAP placed more than 60% residues on average in correct orientation and position for the successful proteins in all three experiments of different residue size. Similarly, average CA-RMSD to the partial model to the native structures reached at atomic level accuracy (< 1.75 Å) (Table 5.2).

FRAP correctly located 65.10% residues on average for 8 proteins using seventeen-residue fragments. The average accuracy of these partial structures is 1.57 Å. Secondary structure elements (α-helical and anti-parallel beta strands connect by small loops) were placed correctly in asymmetric unit in most cases. Here, the example of protein 1EW4 is provided, which is α+β proteins containing two α-helices, six anti-parallel β-strands with long loops. First, FRAP started with the seed fragments of anti-parallel β-strand. One of low energy models that deviated 9.33 Å CA-RMSD from native structure was superimposed to the seed fragment. FRAP searched for correct position and orientation of other fragments. The degrees of freedom allowed were origins translation vectors, crystallographic symmetry operators and unit cell translation. FRAP selected fragments that are nearest to the aligned region of reference *de novo* model using crystallographic symmetry operators and permissible origins shift. Unit cell translation vector was computed with reference

to the seed fragment. FRAP enables to put 73.58% residues that were belonged to two α-helices, three anti-parallel β-strands, and few residues of loop. The phases provided using partial models was sufficient to construct the final model using Phenix.autobuild program. It was verified by the provided R- and R-free factors that are 0.31 and 0.34 respectively. However, the best predicted global models (CA-RMSD 4.98 Å to the native structure) including 1000 lowest energy models for this protein was unsuccessful in MR experiment. Similarly, FRAP achieved the success in placing more than 80% residues for proteins of PDB IDs 1CTF and 1OPD (Table 5.2). Both proteins contain α-helices and anti-parallel β-strands. FRAP assembles most part of secondary structure elements adopted in the proteins including the small loop residues. The partial models after assembly showed an accuracy of 0.99 Å in the case of 1OPD and 1.66 Å for 1CTF.

FRAP assembled the fragments after MR for polar space groups (1OPD in P1 and 1CM3 in $P2_1$) differently because their permissible origins are independent. In order to solve origin translation problem, the fast translation function (FTF) was independently run from Phaser on selected fragments from initial Phaser run. After FTF, FRAP computed available crystallographic operators and unit cell translation vector to the seed fragments to place the other fragments in the asymmetric unit. The experiment was repeated to run the FTF for next two fragments in the absence of common residues between fragments. FRAP succeed to obtained the partial models from fragments using the three different size fragments for PDB ID 1OPD. They achieved an accuracy of CA-RMSD of 1.03 Å, 0.99 Å, and 1.06 Å for thirteen-residue, seventeen-residue, and twenty-one-residue fragments respectively. These partial structures contained more than 75% residues. For 1CM3, thirteen-residue and seventeen-residue fragments were provided the adequate phases necessary to build the final models so that their CA-RMSDs to native structure were 1.03 Å and 1.45 Å respectively.

5.3.4. Model quality assessment after model building

The phase obtained from assembled partial models were used for final model building to assess the correctness of phases. The final models were constructed using automated software Phenix.autobuild using obtained phases. The result was assessed using three different criteria. R- and R- free factors (Table 5.2), number of dummy

atoms appeared in the final models, and CA-RMSD between final models and native structure were analyzed (Table 5.2).

The partial models built from thirteen-residue fragments provided accurate phases for seven proteins to complete more than 60% residues in the final models. The number of success was increased to eight proteins when seventeen-residue fragments were selected for phasing but the success was decreased to six proteins for twenty-one-residue fragments. R and R-free factors were, at first, monitored to evaluate the final model. The R-factor ranged from 26% to 38% for successful proteins in all three different size fragments (Table 5.2). Similarly, R-free factor also started from 27% for protein 1OPD and ended at 42% for 4ESP (Table 5.2). Protein 1OPD achieved the best R and R-free factors (26% and 27%) for seventeen-residue fragments. Same trend was followed for proteins 1CM3 and 1CTF in which R- / R-free factors were 29% and 34% respectively using same fragments size. Similarly, seventeen-residue fragments showed successful model building for proteins 3MX7 with R / R – free factors of 35% and 39% respectively which was unsuccessful using other two fragments. Furthermore, partial models from thirteen-residue and twenty-one-residue fragments showed best R- and R-free factors for protein 1OPD (26% and 31%) and 1CTF(27% and 31%) respectively.

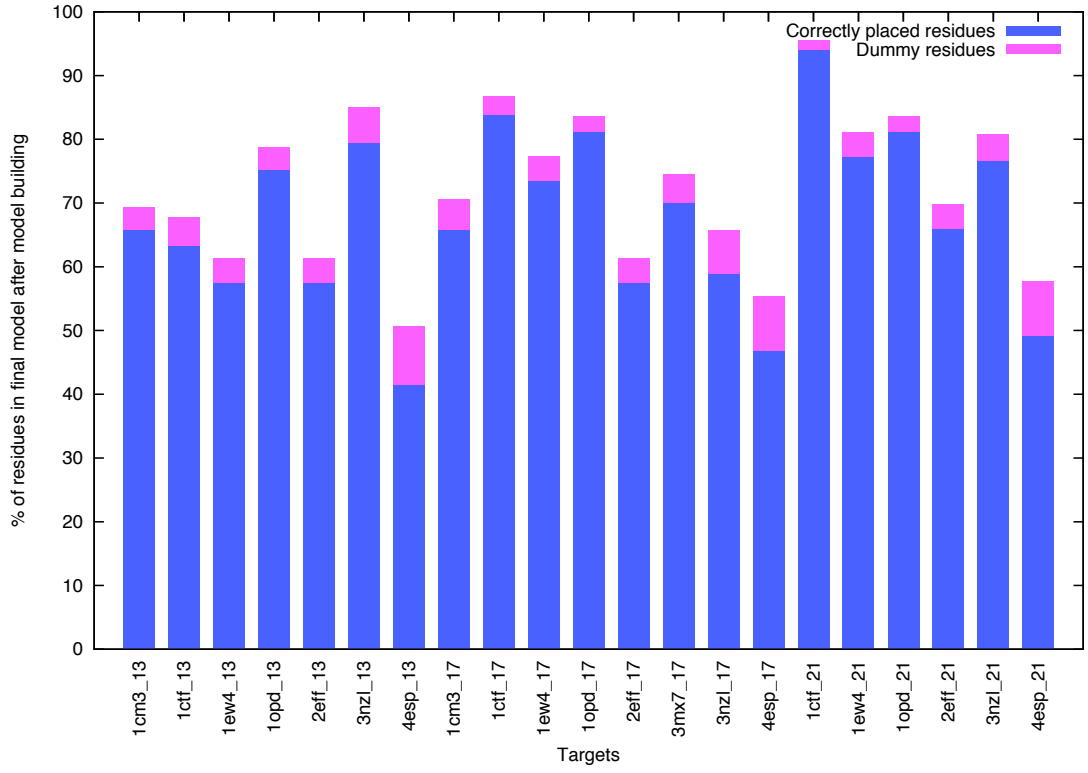

Figure 5.4 Proportion of dummy residues and correctly placed residues in final models

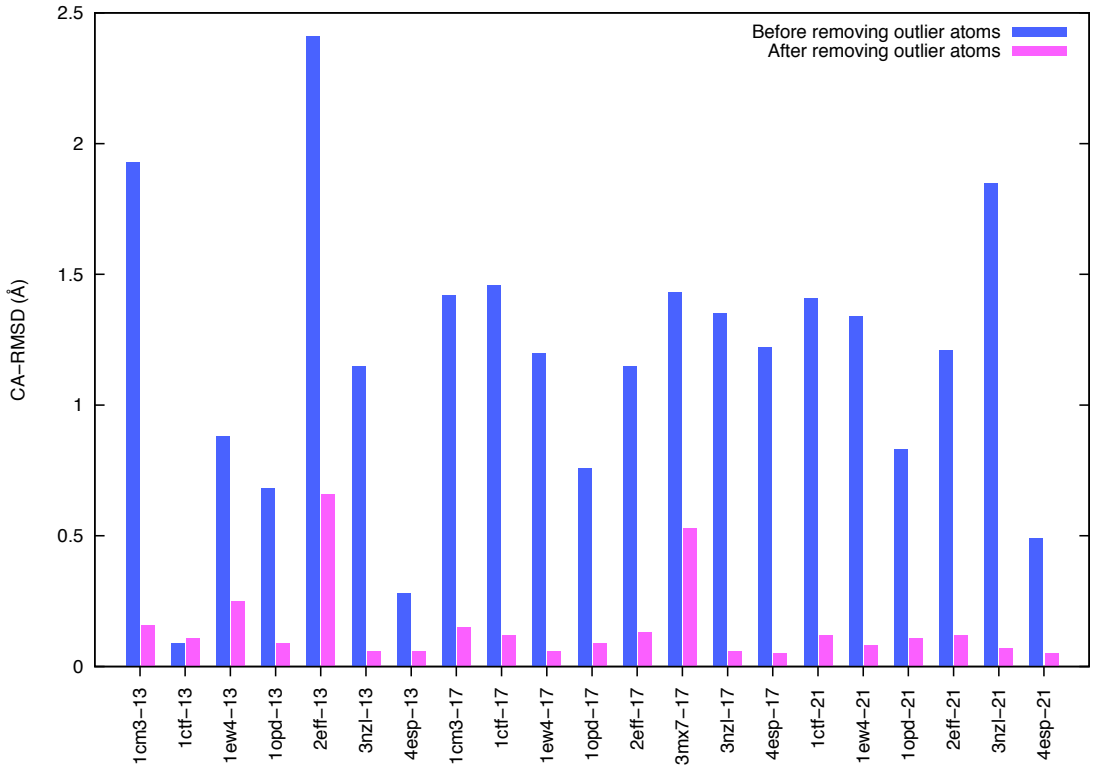

Figure 5.5 Accuracy of models before and after removing outlier atoms

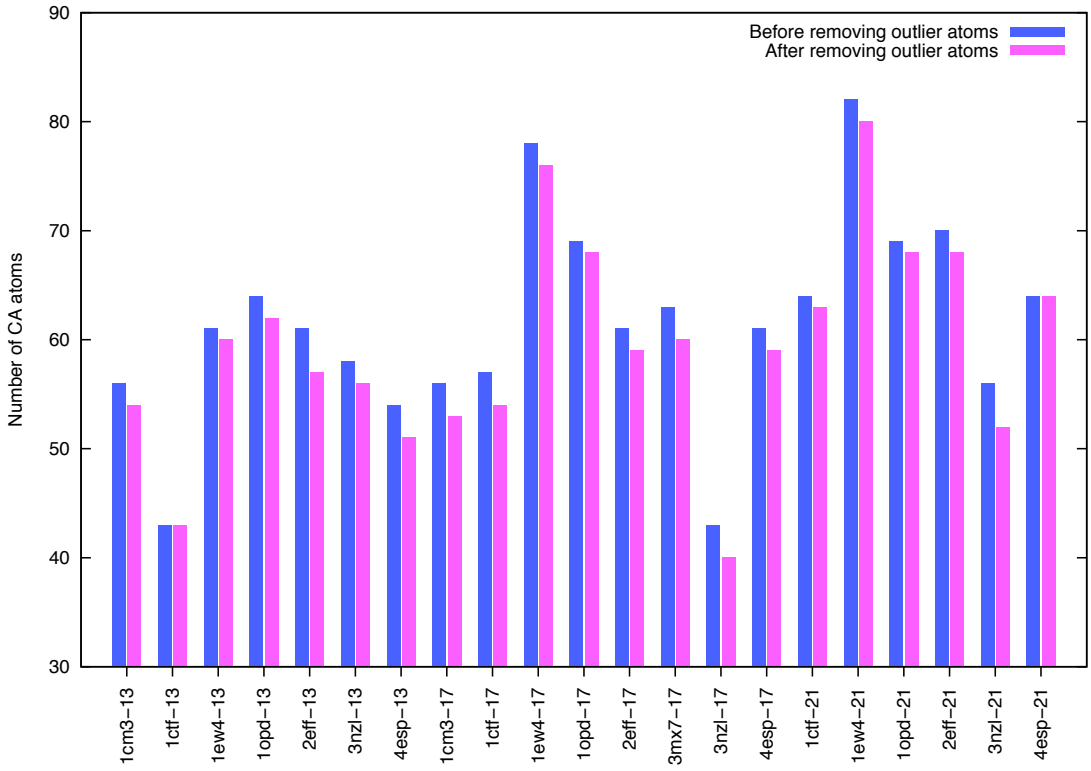

Figure 5.6 Number of outlier atoms in final models

The number of atoms appeared in auto-built models should be evaluated to evaluate the R- and R-free factors properly. The missing atoms in final models and the atoms of water molecules that were not appeared in the native structure and occupied the position of atoms in residues were compared. These atoms were considered as the dummy atoms and it is the water atoms in final models when it falls below the 2.5 Å to the any atoms of residue in native structure. The minimum (18 atoms) and maximum (215 atoms) were observed in the proteins 1CTF and 4ESP. Furthermore, the equivalent number of residues computed from the dummy atoms and this was compared with correctly placed residues in the final model, which is the partial model. The number of atoms is equivalent to 9 residues and 1 residue respectively (Figure 5.4). The R-/R-free factors are 0.35/0.34 for 4ESP and 0.27/0.31 for 1CTF respectively (Table 5.2). Furthermore, in most cases, the number of residues equivalent to dummy atoms appeared almost negligible compared to correctly placed residues (Figure 5.4). It can verify that observed R- / R-free factors were not mainly due to the presence of dummy atoms but with information from interpreted residues in electron density map.

As another assessment strategy, coordinate errors in final models was measured. The CA-RMSD of the final models to their native structure varies from 0.09 Å (1CTF for thirteen-residue fragments) to 2.41 Å (2EFF using thirteen-residue fragments). The average CA-RMSD of final models for seven proteins deviated 1.06 Å from native structures, which was the better than CA-RMSD observed using seventeen-residue (1.25 Å) and twenty-one-residues (1.19 Å). Although CA-RMSD to native structure reached the atomic level accuracy, few atoms have had higher CA-RMSD that made overall CA-RMSD worst (Figure 5.5). These atoms were called as the outlier and appeared at termini of the fragments in the final models. A noticeable example is 2EFF with thirteen-residue fragment that showed CA-RMSD accuracy 2.41 Å to the native structure. After removing the outlier atoms, the CA-RMSD computed using only remaining atoms were improved for successful proteins and their average CA-RMSD were significantly reduced for all test cases in our experiment (Figure 5.5). The average CA-RMSD without outlier atoms was 0.45 Å, 0.50 Å, and 0.60 Å for final models from thirteen, seventeen, and twenty-one residues fragments. Superpose program (Krissinel and Henrick, 2004) was used to remove the outlier atoms. The number of outlier atoms for each target was observed very few,

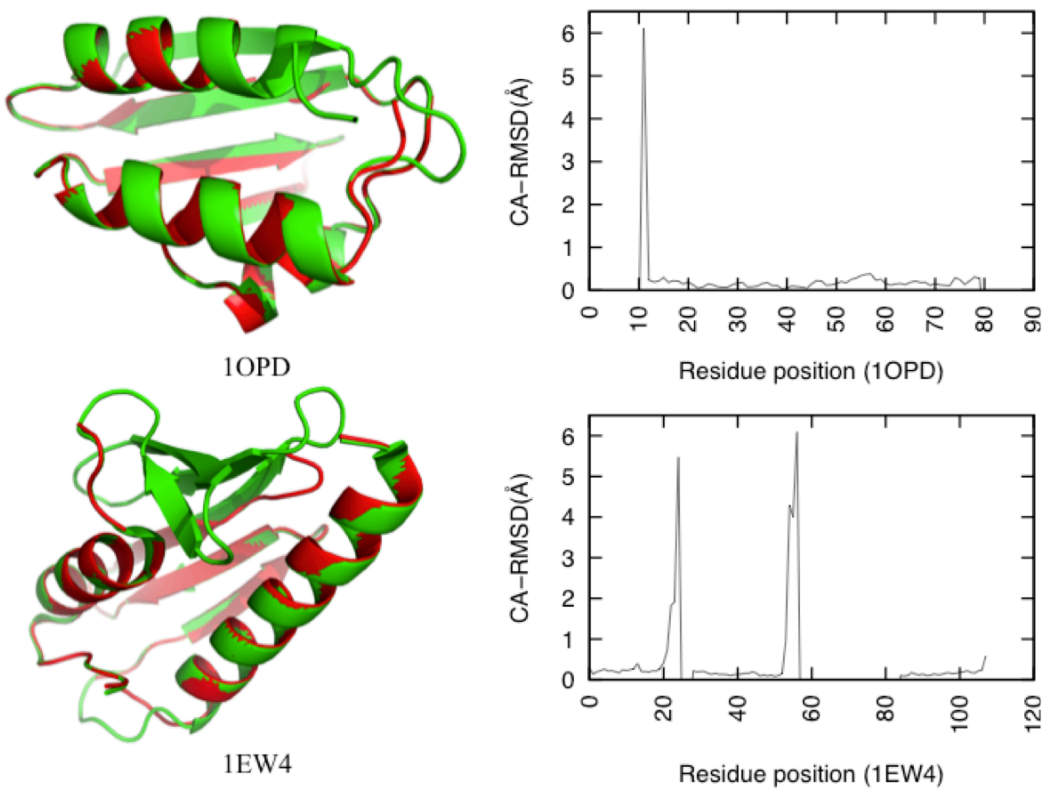

Figure 5.7 Final models superposed with native structure and CA-RMSD for each residue

when number of CA-atoms in models with and without was compared (Figure 5.6). Finally, final models after model building were superposed to native structure for two cases 1OPD and 1EW4 (Figure 5.7). The histogram of CA-RMSD showed very high CA-RMSD for few residue position and their CA-RMSD made overall CA-RMSD worst. In the case of CA-RMSD distribution for 1OPD, only one atoms were deviated by 6 Å in N-terminal and that is main cause for having CA-RMSD high. Similary, this pattern was also observed in 1EW4 where outlier was oberved at two places, C-terminal segments of the fragments.

5.4. Conclusion

In this study, phasing of difficult targets was solved using the fragments from low-quality *de novo* models with MR. The method was tested using three different size fragments and seventeen-residue fragment showed maximum number of successful cases compared with others. On the benchmark test cases of ten targets, this method succeeded in phasing for 80% targets using seventeen-residues fragments. The *de novo* models for these targets were very far away from target structure and subsequently unable to provide the required phases for structure determination. The successful final models from seventeen-residue fragments contained majority of

residues (67.23%). These incomplete models achieved the accuracy of 1.25 Å on average.

Chapter 6. Summary

My research in PhD study has focused in protein structure modeling. Since protein structures have played pivotal role in biology, it provides crucial information required to understand various biochemical mechanism in atomic level. Therefore, proteins with atomic level accuracy are highly demanding. However, protein structure determination is difficult as well as costly using biophysical methods. Therefore, it is necessary to design the efficient computational methods to generate the protein structure from its sequence. I have focused in the development of novel methods for protein structure modeling in PhD study. Three different programs, MORPHEUS, NEFILIM, and FRAP, were developed during the period of my PhD study. These methods focused for protein structure prediction as well as its utility in solving crystallographic phase problems in the absence of homologous protein. In MORPHEUS and NEFILIM programs, the algorithms were developed for improvement of the accuracy of *de novo* models. The accuracy of the predicted models using MORPHEUS was adequate for solving the crystallographic phase problem. Both methods have focused on improvement in conformation sampling to predict better quality models using biased conformation space searching and providing better quality fragments. In FRAP, I focused specially on solving phase problem when the *de novo* models were poor quality and unable to provide phases. FRAP used the fragments generated from *de novo* models for solving phase problem when full-length *de novo* model was not adequate for successful phasing.

These methods solved the protein structures that were unable to solve using existing methods. Despite all these success, *ab initio* phasing using *de novo* model is still challenging to become practical method as well as is far away from routine. There exists the major challenge of predicting high-quality models in protein structure modeling using computer program for practical problems. Therefore, improvement in sampling algorithm, devising accurate energy function, and providing better quality fragments are fundamentally necessary. Although more than half century was spent in development of conformation sampling algorithms and energy functions, problem has still remained unsolved so that future efforts are necessary to make handy tool for biologist. The methods that were developed during my PhD study can be further improved to tackle the underlying challenges. Therefore, I am also interested to extend the concepts implemented in MORPHEUS, NEFILIM and FRAP in the future.

The major goal of future work is to focus on development of novel methods to tackle real-life problems faced by biologist. There are few keys that can be considered as the future work in MORPHEUS and NEFILIM. These programs focused on improving the conformational sampling strategy by exploiting the knowledge from the predicted structures. The concept can be further furnished by incorporating more knowledge to improve the conformational sampling. Before incorporating the knowledge, it is necessary to measure the information available in already predicted models. Afterward, the information can be further incorporated from structures deposited in PDB if possible. In addition, weak information obtained from biophysical experiments can be also provided to enhance the conformation sampling. *Ab initio* approaches can be utilized to enhance the conformation sampling however this approach is more challenging and time consuming. NEFILIM uses the fragments generated from its first run in the resampling approach. I am also interested to generate fragments from the known structure available in PDB given the query fragment and use these fragments for resampling. I am also interested to identify the wrong regions in predicted models and variable-length fragments will be generated to sample focusing on these regions instead of using constant-length fragments. FRAP has many utilities as future work. It identified the position and orientation of partial models in the asymmetric unit and atomic positions of many residues are still unknown in the unit cell. I will continue to develop a method that can identify the position of missing residue in asymmetric unit using the information obtained from known partial structures and diffraction data. In addition, I plan to extend the utility of this concept by solving the challenging cases such as the structure of large domains and structures that have multi copies in asymmetric unit, which are major problems among the crystallographers with the help of currently available conformational sampling algorithm. Furthermore, natural extension of FRAP can be to use fragments identified directly from PDB based on sequence similarity or secondary structure elements. An ensemble of NMR models also can be used in FRAP for crystallographic structure determination. FRAP can be also applied with fragments from distant homologous. Protein structures and its dynamics provide inter-atomic and intra-atomic interactions in atomic level. These interactions can provide the crucial information about mechanism happened in molecule and then cell. Therefore, it is necessary to develop an efficient method for structure modeling to understand the biological mechanism in cellular level and further.

Chapter 7. Reference

Adams, P.D., Afonine, P.V., Grosse-Kunstleve, R.W., Read, R.J., Richardson, J.S., Richardson, D.C. and Terwilliger, T.C. (2009) Recent developments in phasing and structure refinement for macromolecular crystallography, *Curr Opin Struct Biol*, **19**, 566-572.

Adams, P.D., Baker, D., Brunger, A.T., Das, R., DiMaio, F., Read, R.J., Richardson, D.C., Richardson, J.S. and Terwilliger, T.C. (2013) Advances, Interactions, and Future Developments in the CNS, Phenix, and Rosetta Structural Biology Software Systems, *Annu Rev Biophys*, **42**, 265-287.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H. and Adams, P.D. (2012) Towards automated crystallographic structure refinement with phenix.refine, *Acta Crystallogr D*, **68**, 352-367.

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

Anfinsen, C.B. (1972) The formation and stabilization of protein structure, *The Biochemical journal*, **128**, 737-749.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics, *Science*, **294**, 93-96.

Barford, D., Das, A.K. and Egloff, M.P. (1998) The structure and mechanism of protein phosphatases: Insights into catalysis and regulation, *Annu Rev Bioph Biom*, **27**, 133-164.

Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J.L. and Levy, Y. (2009) Assessment of CASP8 structure predictions for template free targets, *Proteins-Structure Function and Bioinformatics*, **77**, 50-65.

Berenger, F., Zhou, Y., Shrestha, R. and Zhang, K.Y. (2011) Entropy-accelerated exact clustering of protein decoys, *Bioinformatics*, **27**, 939-945.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C. (2002) The Protein Data Bank, *Acta crystallographica. Section D, Biological crystallography*, **58**, 899-907.

Bibby, J., Keegan, R.M., Mayans, O., Winn, M.D. and Rigden, D.J. (2014) AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models. Corrigendum, *Acta crystallographica. Section D, Biological crystallography*, **70**, 1174.

Blow, D.M. and Rossmann, M.G. (1961) The single isomorphous replacement method, *Acta Crystallographica*, **14**, 1195-1202.

Blum, B., Jordan, M.I. and Baker, D. (2010) Feature space resampling for protein conformational search, *Proteins-Structure Function and Bioinformatics*, **78**, 1583-1593.

Blundell, T.L., Jhoti, H. and Abell, C. (2002) High-throughput crystallography for lead discovery in drug design, *Nat Rev Drug Discov*, **1**, 45-54.

Blundell, T.L. and Patel, S. (2004) High-throughput X-ray crystallography for drug discovery, *Current opinion in pharmacology*, **4**, 490-496.

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

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules, *Nature*, **326**, 347-352.

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

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

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

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.

Brunette, T.J. and Brock, O. (2005) Improving protein structure prediction with model-based search, *Bioinformatics*, **21**, 66-74.

Brunette, T.J. and Brock, O. (2005) Improving protein structure prediction with model-based search, *Bioinformatics*, **21**, I66-I74.

Canutescu, A.A. and Dunbrack, R.L., Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure, *Protein Sci*, **12**, 963-972.

Carvalho, A.L., Trincao, J. and Romao, M.J. (2009) X-ray crystallography in drug discovery, *Methods Mol Biol*, **572**, 31-56.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta crystallographica. Section D, Biological crystallography*, **66**, 12-21.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E.M., Bonneau, R., Rohl, C.A. and Baker, D. (2003) Automated prediction of CASP-5 structures using the Robetta server, *Proteins-Structure Function and Bioinformatics*, **53**, 524-533.

Chothia, C. (1984) Principles that determine the structure of proteins., *Annu Rev Biochem.*, **53**, 537-572.

Claude, J.B., Suhre, K., Notredame, C., Claverie, J.M. and Abergel, C. (2004) CaspR: a web server for automated molecular replacement using homology modelling, *Nucleic Acids Research*, **32**, W606-W609.

Collins, F.S., Morgan, M. and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology, *Science*, **300**, 286-290.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998-2003, *Science*, **282**, 682-689.

Correia, B.E., Bates, J.T., Loomis, R.J., Baneyx, G., Carrico, C., Jardine, J.G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M.J., Stevens, E., Schroeter, A., Chen, M., MacPherson, S., Serra, A.M., Adachi, Y., Holmes, M.A., Li, Y.X., Klevit, R.E., Graham, B.S., Wyatt, R.T., Baker, D., Strong, R.K., Crowe, J.E., Johnson, P.R. and Schief, W.R. (2014) Proof of principle for epitope-focused vaccine design, *Nature*, **507**, 201-206.

Das, R. (2011) Four Small Puzzles That Rosetta Doesn't Solve, *PLoS One*, **6**, e20044.

Das, R. and Baker, D. (2008) Macromolecular modeling with rosetta, *Annu Rev Biochem*, **77**, 363-382.

Das, R. and Baker, D. (2009) Prospects for de novo phasing with de novo protein models, *Acta Crystallogr D Biol Crystallogr*, **65**, 169-175.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M.D., Bhat, D., Chivian, D., Kim, D.E., Sheffler, W.H., Malmstrom, L., Wollacott, A.M., Wang, C., Andre, I. and Baker, D. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home, *Proteins*, **69 Suppl 8**, 118-128.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M.D., Bhat, D., Chivian, D., Kim, D.E., Sheffler, W.H., Malmstrom, L., Wollacott, A.M., Wang, C., Andre, I. and Baker, D. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home, *Proteins*, **69**, 118-128.

delaFortelle, E. and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods, *Method Enzymol*, **276**, 472-494.

Desnick, R.J. and Schuchman, E.H. (2002) Enzyme replacement and enhancement therapies: lessons from lysosomal disorders, *Nature reviews. Genetics*, **3**, 954-966.

Dill, K.A. (1990) Dominant forces in protein folding, *Biochemistry*, **29**, 7133-7155.

Dill, K.A. and MacCallum, J.L. (2012) The protein-folding problem, 50 years on, *Science*, **338**, 1042-1046.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L., Das, D., Vorobiev, S.M., Iwai, H., Pokkuluri, P.R. and Baker, D. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization, *Nature*, **473**, 540-543.

Eisenberg, D. (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins, *Proc Natl Acad Sci U S A*, **100**, 11207-11210.

Fleishman, S.J. and Baker, D. (2012) Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution, *Cell*, **149**, 262-273.

Fleishman, S.J. and Baker, D. (2012) Role of the biomolecular energy gap in protein design, structure, and evolution, *Cell*, **149**, 262-273.

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

Geffeney, S., Brodie, E.D., Jr., Ruben, P.C. and Brodie, E.D., 3rd (2002) Mechanisms of adaptation in a predator-prey arms race: TTX-resistant sodium channels, *Science*, **297**, 1336-1339.

Giorgetti, A., Raimondo, D., Miele, A.E. and Tramontano, A. (2005) Evaluating the usefulness of protein structure models for molecular replacement, *Bioinformatics*, **21 Suppl 2**, ii72-76.

Glotzer, M., Murray, A.W. and Kirschner, M.W. (1991) Cyclin Is Degraded by the Ubiquitin Pathway, *Nature*, **349**, 132-138.

Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E.M. and Baker, D. (2011) Generalized Fragment Picking in Rosetta: Design, Protocols and Applications, *PLoS One*, **6**, e23294.

Hamelryck, T., Kent, J.T. and Krogh, A. (2006) Sampling realistic protein conformations using local structural bias, *PLoS computational biology*, **2**, e131.

Han, K.F. and Baker, D. (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins, *Proc Natl Acad Sci U S A*, **93**, 5814-5818.

Hegler, J.A., Latzer, J., Shehu, A., Clementi, C. and Wolynes, P.G. (2009) Restriction versus guidance in protein structure prediction, *Proc Natl Acad Sci U S A*, **106**, 15302-15307.

Hendrickson, W.A., Horton, J.R. and Lemaster, D.M. (1990) Selenomethionyl Proteins Produced for Analysis by Multiwavelength Anomalous Diffraction (Mad) - a Vehicle for Direct Determination of 3-Dimensional Structure, *Embo J*, **9**, 1665-1672.

Joachimiak, A. (2009) High-throughput crystallography for structural genomics, *Curr Opin Struct Biol*, **19**, 573-584.

Jones, D.T. and McGuffin, L.J. (2003) Assembling novel protein folds from super-secondary structural fragments, *Proteins*, **53 Suppl 6**, 480-485.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition, *Nature*, **358**, 86-89.

Jones, S. and Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc Natl Acad Sci U S A*, **93**, 13-20.

Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors, *Acta Crystallographica Section A*, **32**, 922-923.

Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, **22**, 2577-2637.

Kalev, I. and Habeck, M. (2011) HHfrag: HMM-based fragment detection using HHpred, *Bioinformatics*, **27**, 3110-3116.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. and Hughey, R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction, *Proteins*, **53 Suppl 6**, 491-496.

Keegan, R.M. and Winn, M.D. (2007) Automated search-model discovery and preparation for structure solution by molecular replacement, *Acta Crystallogr D*, **63**, 447-457.

Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution, *Nature*, **185**, 422-427.

Khatib, F., DiMaio, F., Foldit Contenders, G., Foldit Void Crushers, G., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popovic, Z., Jaskolski, M. and Baker, D. (2011) Crystal structure of a monomeric retroviral protease solved by protein folding game players, *Nature structural & molecular biology*, **18**, 1175-1177.

Kim, D.E., Blum, B., Bradley, P. and Baker, D. (2009) Sampling bottlenecks in de novo protein structure prediction, *J Mol Biol*, **393**, 249-260.

Kissinger, C.R., Gehlhaar, D.K. and Fogel, D.B. (1999) Rapid automated molecular replacement by evolutionary search, *Acta crystallographica. Section D, Biological crystallography*, **55**, 484-491.

Kornberg, R.D. (1974) Chromatin structure: a repeating unit of histones and DNA, *Science*, **184**, 868-871.

Kratzner, R., Debreczeni, J.E., Pape, T., Schneider, T.R., Wentzel, A., Kolmar, H., Sheldrick, G.M. and Uson, I. (2005) Structure of Ecballium elaterium trypsin inhibitor II (EETI-II): a rigid molecular scaffold, *Acta crystallographica. Section D, Biological crystallography*, **61**, 1255-1262.

Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta crystallographica. Section D, Biological crystallography*, **60**, 2256-2268.

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

Langer, G., Cohen, S.X., Lamzin, V.S. and Perrakis, A. (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7, *Nature Protocols*, **3**, 1171-1179.

Lazaridis, T., Archontis, G. and Karplus, M. (1995) Enthalpic Contribution to Protein Stability: Insights from Atom-Based Calculations and Statistical Mechanics, **47**, 231-306.

Lazaridis, T. and Karplus, M. (2000) Effective energy functions for protein structure prediction, *Curr Opin Struct Biol*, **10**, 139-145.

Leahy, D.J., Hendrickson, W.A., Aukhil, I. and Erickson, H.P. (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein, *Science*, **258**, 987-991.

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

Levintha.C (1968) Are There Pathways for Protein Folding, *J Chim Phys Pcb*, **65**, 44-&.

Levitt, M. (1992) Accurate Modeling of Protein Conformation by Automatic Segment Matching, *Journal of Molecular Biology*, **226**, 507-533.

Li, S.C., Bu, D.B., Xu, J.B. and Li, M. (2008) Fragment-HMM: A new approach to protein structure prediction, *Protein Science*, **17**, 1925-1934.

Li, Z. and Scheraga, H.A. (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding, *Proc Natl Acad Sci U S A*, **84**, 6611-6615.

Liwo, A., Czaplewski, C., Oldziej, S. and Scheraga, H.A. (2008) Computational techniques for efficient conformational sampling of proteins, *Curr Opin Struct Biol*, **18**, 134-139.

Long, F., Vagin, A.A., Young, P. and Murshudov, G.N. (2008) BALBES: a molecular-replacement pipeline, *Acta Crystallogr D*, **64**, 125-132.

M. F. Perutz, F.R.S., M. G. Rossmann, ANN F. Cullis, Hilary Muirhad, George Will (1960) Structure of Haemoglobin a three-dimensional fourier synthesis at 5.5 angstrom resolution, obtained by X-ray analysis, *Nature* **185**, 416-422.

MacCallum, J.L., Perez, A., Schnieders, M.J., Hua, L., Jacobson, M.P. and Dill, K.A. (2011) Assessment of protein structure refinement in CASP9, *Proteins*, **79 Suppl 10**, 74-90.

Mandell, D.J., Coutsias, E.A. and Kortemme, T. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling, *Nature methods*, **6**, 551-552.

Mao, B., Guan, R. and Montelione, G.T. (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement, *Structure*, **19**, 757-766.

Mao, B.C., Tejero, R., Baker, D. and Montelione, G.T. (2014) Protein NMR Structures Refined with Rosetta Have Higher Accuracy Relative to Corresponding X-ray Crystal Structures, *J Am Chem Soc*, **136**, 1893-1906.

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

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software, *Journal of applied crystallography*, **40**, 658-674.

McCoy, A.J., Storoni, L.C. and Read, R.J. (2004) Simple algorithm for a maximum-likelihood SAD function, *Acta Crystallogr D*, **60**, 1220-1228.

Mirsky, A.E. and Pauling, L. (1936) On the Structure of Native, Denatured, and Coagulated Proteins, *Proc Natl Acad Sci U S A*, **22**, 439-447.

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

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Tramontano, A. (2014) Critical assessment of methods of protein structure prediction (CASP)--round x, *Proteins*, **82 Suppl 2**, 1-6.

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method, *Acta Crystallogr D*, **53**, 240-255.

Navaza, J. (2001) Implementation of molecular replacement in AMoRe, *Acta crystallographica. Section D, Biological crystallography*, **57**, 1367-1372.

Pauling, L. and Corey, R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains, *Proc Natl Acad Sci U S A*, **37**, 251-256.

Pauling, L., Corey, R.B. and Branson, H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain, *Proc Natl Acad Sci U S A*, **37**, 205-211.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G. and North, A.C. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis, *Nature*, **185**, 416-422.

Pozharski, E. (2010) Percentile-based spread: a more accurate way to compare crystallographic models, *Acta Crystallogr D*, **66**, 970-978.

Procko, E., Berguig, G.Y., Shen, B.W., Song, Y., Frayo, S., Convertine, A.J., Margineantu, D., Booth, G., Correia, B.E., Cheng, Y., Schief, W.R., Hockenbery, D.M., Press, O.W., Stoddard, B.L., Stayton, P.S. and Baker, D. (2014) A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells, *Cell*, **157**, 1644-1656.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J. and Baker, D. (2007) High-resolution structure prediction and the crystallographic phase problem, *Nature*, **450**, 259-U257.

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

Rajagopalan, S., Wang, C., Yu, K., Kuzin, A.P., Richter, F., Lew, S., Miklos, A.E., Matthews, M.L., Seetharaman, J., Su, M., Hunt, J.F., Cravatt, B.F. and Baker, D. (2014) Design of activated serine-containing catalytic triads with atomic-level accuracy, *Nature chemical biology*, **10**, 386-391.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations, *J Mol Biol*, **7**, 95-99.

Ramelot, T.A., Raman, S., Kuzin, A.P., Xiao, R., Ma, L.C., Acton, T.B., Hunt, J.F., Montelione, G.T., Baker, D. and Kennedy, M.A. (2009) Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study, *Proteins*, **75**, 147-167.

Rigden, D.J., Keegan, R.M. and Winn, M.D. (2008) Molecular replacement using ab initio polyalanine models generated with ROSETTA, *Acta Crystallographica Section D*, **64**, 1288-1291.

Rodriguez, D.D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I.M., Becker, S., Sheldrick, G.M. and Uson, I. (2009) Crystallographic ab initio protein structure solution below atomic resolution, *Nature methods*, **6**, 651-653.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. and Baker, D. (2004) Protein structure prediction using rosetta, *Numerical Computer Methods, Pt D*, **383**, 66-+.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. and Baker, D. (2004) Protein structure prediction using rosetta, *Methods in Enzymology*, **383**, 66-93.

Rose, G.D., Gierasch, L.M. and Smith, J.A. (1985) Turns in peptides and proteins, *Advances in protein chemistry*, **37**, 1-109.

Rossmann, M.G. and Blow, D.M. (1962) The detection of sub-units within the crystallographic asymmetric unit, *Acta Crystallogr D*, **15**, 24-31.

Rowland, R.S. (2002) Using X-ray crystallography in drug discovery, *Current opinion in drug discovery & development*, **5**, 613-619.

Roy, A., Kucukural, A. and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction, *Nat Protoc*, **5**, 725-738.

Sali, A. and Blundell, T.L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints, *Journal of Molecular Biology*, **234**, 779-815.

Sammito, M., Millan, C., Rodriguez, D.D., de Ilarduya, I.M., Meindl, K., De Marino, I., Petrillo, G., Buey, R.M., de Pereda, J.M., Zeth, K., Sheldrick, G.M. and Uson, I. (2013) Exploiting tertiary structure through local folds for crystallographic phasing, *Nature methods*, **10**, 1099-1101.

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

Schroder, G.F., Levitt, M. and Brunger, A.T. (2010) Super-resolution biomolecular crystallography with low-resolution data, *Nature*, **464**, 1218-1222.

Sela, M. and Lifson, S. (1959) The reformation of disulfide bridges in proteins, *Biochimica et biophysica acta*, **36**, 471-478.

Shortle, D., Simons, K.T. and Baker, D. (1998) Clustering of low-energy conformations near the native structures of small proteins, *Proc Natl Acad Sci U S A*, **95**, 11158-11162.

Shrestha, R., Berenger, F. and Zhang, K.Y. (2011) Accelerating ab initio phasing with de novo models, *Acta Crystallogr D Biol Crystallogr*, **67**, 804-812.

Shrestha, R., Simoncini, D. and Zhang, K.Y. (2012) Error-estimation-guided rebuilding of de novo models increases the success rate of ab initio phasing, *Acta Crystallogr D Biol Crystallogr*, **68**, 1522-1534.

Shrestha, R. and Zhang, K.Y. (2014) Improving fragment quality for de novo structure prediction, *Proteins*.

Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality, *Bioinformatics*, **16**, 776-785.

Simoncini, D., Berenger, F., Shrestha, R. and Zhang, K.Y.J. (2012) A Probabilistic Fragment-Based Protein Structure Prediction Algorithm, *Plos One*, **7**.

Simoncini, D. and Zhang, K.Y. (2013) Efficient sampling in fragment-based protein structure prediction using an estimation of distribution algorithm, *PLoS One*, **8**, e68954.

Simoncini, D. and Zhang, K.Y.J. (2013) Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm, *PLoS One*, **8**, e68954.

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

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

Soding, J., Biegert, A. and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction, *Nucleic Acids Res*, **33**, W244-248.

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

Song, Y., DiMaio, F., Wang, R.Y., Kim, D., Miles, C., Brunette, T., Thompson, J. and Baker, D. (2013) High-resolution comparative modeling with RosettaCM, *Structure*, **21**, 1735-1742.

Stein, N. (2008) CHAINSAW: a program for mutating pdb files used as templates in molecular replacement, *Journal of applied crystallography*, **41**, 641-643.

Strop, P., Brzustowicz, M.R. and Brunger, A.T. (2007) Ab initio molecular-replacement phasing for symmetric helical membrane proteins, *Acta crystallographica. Section D, Biological crystallography*, **63**, 188-196.

Strynadka, N.C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N. (1996) Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase, *Nature structural biology*, **3**, 233-239.

Szep, S., Wang, J. and Moore, P.B. (2003) The crystal structure of a 26-nucleotide RNA containing a hook-turn, *Rna*, **9**, 44-51.

Team, R.D.C. (2008) R: A language and environment for statistical computing. URL http://www.R-project.org., *R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0*. R Foundation for Statistical Computing.

Terwilliger, T.C. and Berendzen, J. (1999) Automated MAD and MIR structure solution, *Acta Crystallogr D*, **55**, 849-861.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, P.H., Hung, L.W., Read, R.J. and Adams, P.D. (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard, *Acta crystallographica. Section D, Biological crystallography*, **64**, 61-69.

Thompson, J.M., Sgourakis, N.G., Liu, G.H., Rossi, P., Tang, Y.F., Mills, J.L., Szyperski, T., Montelione, G.T. and Baker, D. (2012) Accurate protein structure modeling using sparse NMR data and homologous structure information, *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 9875-9880.

Trievel, R.C., Beach, B.M., Dirk, L.M.A., Houtz, R.L. and Hurley, J.H. (2002) Structure and catalytic mechanism of a SET domain protein methyltransferase, *Cell*, **111**, 91-103.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A. and Baker, D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction, *Proteins*, **53**, 76-87.

Tyka, M.D., Keedy, D.A., Andre, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S. and Baker, D. (2011) Alternate states of proteins revealed by detailed energy landscape mapping, *J Mol Biol*, **405**, 607-618.

Ueno, G., Kanda, H., Hirose, R., Ida, K., Kumasaka, T. and Yamamoto, M. (2006) RIKEN structural genomics beamlines at the SPring-8; high throughput protein crystallography with automated beamline operation, *J Struct Funct Genomics*, **7**, 15-22.

van Gent, D.C., Hoeijmakers, J.H. and Kanaar, R. (2001) Chromosomal stability and the DNA double-stranded break connection, *Nature reviews. Genetics*, **2**, 196-206.

Wallner, B. and Elofsson, A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information, *Protein Science*, **15**, 900-913.

Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature*, **171**, 737-738.

Wayne A. Hendrickson, M.M.T. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur, *Nature* **290**, 107-113.

Weiss, M.S. and Hilgenfeld, R. (1999) A method to detect nonproline cis peptide bonds in proteins, *Biopolymers*, **50**, 536-544.

Xu, D., Zhang, J., Roy, A. and Zhang, Y. (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement, *Proteins*, **79 Suppl 10**, 147-160.

Xu, D. and Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, *Proteins: Structure, Function, and Bioinformatics*, **80**, 20.

Xu, D. and Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, *Proteins*, **80**, 1715-1735.

Xu, D. and Zhang, Y. (2013) Toward optimal fragment generations for ab initio protein structure assembly, *Proteins*, **81**, 229-239.

Xu, D. and Zhang, Y. (2013) Toward optimal fragment generations for ab initio protein structure assembly, *Proteins*, **81**, 229-239.

Yusupova, G. and Yusupov, M. (2014) High-Resolution Structure of the Eukaryotic 80S Ribosome, *Annu Rev Biochem*.

Zemla, A. (2003) LGA: A method for finding 3D similarities in protein structures, *Nucleic Acids Res*, **31**, 3370-3374.

Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. (1999) Processing and analysis of CASP3 protein structure predictions, *Proteins*, **Suppl 3**, 22-29.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A. and Honig, B. (2012)

Structure-based prediction of protein-protein interactions on a genome-wide scale, *Nature*, **490**, 556-560.

Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality, *Proteins*, **57**, 702-710.

Zhang, Y. and Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds, *Journal of computational chemistry*, **25**, 865-871.