De novo protein backbone modeling with low-resolution density maps

Inaugural dissertation

for the attainment of the title of doctor

in the Faculty of Mathematics and Natural Sciences

at the Heinrich Heine University Düsseldorf

presented by

Zhe Wang

from Harbin

Jülich, Nov 2014

from the institute ICS-6

at the Forschungszentrum Jülich

Published by permission of the

Faculty of Mathematics and Natural Sciences at

Heinrich Heine University Düsseldorf

Supervisior: Jun. Prof. Dr. Gunnar Schröder

Co-supervisor: Prof. Dr. Dieter Willbold

Date of the oral examination: 03/02/2015

Declaration of Authorship

I, Zhe Wang, declare that this thesis titled, 'De novo protein backbone modeling with low-resolution density maps' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

De novo protein backbone modeling with low-resolution density maps

by Zhe Wang

The field of structural biology and in particular the method of 3-D electron cryomicroscopy (cryoEM) is developing quickly. More and more macromolecules or complexes of macromolecules are investigated and their structure determined by cryoEM. In recent years, a large number of these large structures were determined to subnanometer resolutions (3.5-10 Å). Since the number of subnanometer resolution density maps obtained by cryo-EM is steadily increasing, methods for the interpretation of these data are highly demanded.

To address this, we have developed a method that can build up the protein backbone structure from an intermediate resolution density map without any prior information from crystal structures or homology models. The method is based on a combinatorial optimization algorithm, the Lin-Kernighan heuristic, which is used for solving the Euclidean traveling salesman problem. Meanwhile, the search of models with the Lin-Kernighan heuristic is biased with additional restraints including secondary structure, statistical potential restraints and density based restraints. With this we generate ensembles of backbone traces, from which most likely traces, which fit best into the density map, are extracted and can be used to build up full atomic models.

In this work, the method was first extensively tested with an exemplified structure of the protein calmodulin. Later on, six structures were selected from three classes defined by the CATH classification for further tests.

The success of the method depends on the resolution of the density map. For the all-α-helix class, a resolution of 8 Å is sufficient to determine the correct backbone topology. For the all-β-sheet class, a resolution of 4.5 Å is necessary. For the mixed $\alpha - \beta$ class, a resolution of 6 Å was sufficient for our test cases to obtain the correct protein topology. The final backbone traces were then obtained after structure refinement with secondary structure distance restraints using the real-space refinement program DireX.

De-novo Proteinstrukturmodellierung mit Hilfe von Dichtekarten

by Zhe Wang

In der Strukturbiologie werden im Laufe der Zeit immer mehr Makromoleküle und Proteinkomplexe untersucht und deren Strukturen bestimmt. Zur Ermittlung der Strukturen sehr grosser Komplexe konnte bereits in vielen Fällen die 3D Cryo-Elektronenmikroskopie (Cryo-EM) erfolgreich eingesetzt werden. In den letzten Jahren konnten viele große Strukturen mit Auflösungen im Subnanometer-Bereich (3.5-10 Å) bestimmt werden. Da die Anzahl an Cryo-EM-Dichtekarten mit Subnanometer-Auflösungen stetig wächst, sind Methoden für deren Interpretation stark gefragt.

Wir haben eine Methode entwickelt, die, mit Hilfe einer niedrig-aufgelösten Dichtekarte, die Struktur des Proteinrückgrates ohne die Verwendung von Kristallstrukturen oder Homologiemodellen aufbauen kann. Die Methode basiert auf der Lin-Kernighan Heuristik, einem kombinatorischen Optimierungsalgorithmus welcher verwendet wird, um das Problem des Handlungsreisenden (traveling salesman problem) zu lösen. Die Suche mit der Lin-Kernighan-Heuristik wird zusätzlich durch Terme basierend auf den Sekundärstrukturelementen, statistischen Potentialen und der Dichte beeinflusst. Mit diesem Algorithmus wird ein Ensemble von Proteinrückgraten erstellt. Das Rückgrat, welches am besten in die Dichtekarte passt, kann anschließend als Grundlage zur Generierung von atomaren Modellen verwendet werden.

Unsere Methode wurde zunächst an einer beispielhaften Struktur des Proteins Calmodulin getestet. Anschließend wurde unser Test-Datensatz durch sechs Strukturen aus den drei verschiedenen CATH Klassen erweitert. Für jede der Strukturen wurde eine künstliche Dichtekarte generiert. Der Erfolg der Methode ist abhängig von der Auflösung der Dichte. Für α-helikale Strukturen reicht eine Auflösung von 8 Å, für β-Faltblatt Strukturen ist eine Auflösung von 4.5 Å notwendig, und für Strukturen mit gemischten Sekundärstrukturelementen liefert unsere Methode auch bei einer Auflösung von 6 Å die korrekte Topologie. Die endgültigen Proteinrückgrate werden anschließend mit dem Programm DireX bezüglich ihrer Sekundärstrukturgeometrie verfeinert.

Acknowledgements

I would like to express my special appreciation to my advisor, Gunnar Schröder, who accepted me under his wing as a PhD student. It is a pleasure to work with him in these years. I am also thankful for his guidance, sparking ideas and patience. His attitudes as a research scientist have greatly inspired me.

In addition, I would like to thank all the members in computational structural biology (CSB) group, in particular Andrè Wildberg, Benno Falkner, Kumaran Baskaran, Dennis Della Corte, Amudha Duraisamy, Michaela Spiegel and Tatjana Braun for giving me such a nice circumstance. Thanks for all the motivating discussions and the happy conversations.

Most of all I would like to thank my wife Su. Her love, encouragement and understanding gave me great strength. Being like a sun, shining into my heart all the time.

Contents

List of Figures

List of Tables

Abbreviations

PDB	**P**rotein **D**ata **B**ank
NMR	**N**clear **M**agnetic **R**esonance
CryoEM	**C**ryo **E**lectron **M**icroscopy
XRC	**X** **R**ay **C**rystallography
CCD	**C**harged **C**oupled **D**evice
SNR	**S**ignal to **N**oise **R**atiod
EMDB	**E**lectron **M**icroscopy **D**ata **B**ank
FSC	**F**ourier **S**hell **C**orrelation
MD	**M**olecular **D**ynamic
MDFF	**M**olecular **D**ynamic **F**lexible **F**itting
NMA	**N**ormanl **M**odels **A**nalysis
ENM	**E**lastic **N**etwork **M**odels
CATH	**C**olecular **A**ynamic **T**lexible **H**itting
DEN	**D**eformable **E**lastic **N**etwork
MC	**M**onto **C**arlo
RMSD	**R**oot **M**ean **S**quare **D**eviation
LKH	**L**in **K**ernighan **H**euristic

SA	**S**imulated **A**nnealing
TSP	**T**ravelling **S**alesman **P**roblem
NP	**N**on deterministic **P**olynomial
MLK	**M**odified **L**in **K**ernighan
DSSP	**D**oot **S**ean **S**quare **P**eviation
SSE	**S**econdary **S**tructure **E**lement
LSQ	**L**east **S**quares **Q**uadratic

CHAPTER 1

Introduction

1.1 Structural Biology of Proteins

Structural biology is a branch of biology that focuses on the structure determination of macromolecules. The aim of structural biology is to get an exhaustive understanding of the three-dimensional information of the macromolecules so as to extend this knowledge to understand the functions of different macromolecules and the mechanism of their reactions in the cell. These objects that structural biologist are interested in range from proteins and molecular complexes to organelles.

To determine the structures of those different objects, a variety of techniques are being used for different length scales. These methods generally measure a large number of identical molecules simultaneously. Researchers try to use them to study the native states of these objects and investigate their quantity, location and dynamical information.

Among all the objects that are associated with structural biology, the protein is one of the most attractive and fundamental objects for structural biologists. As proteins are the most essential fragments that are involved in the processes that happen within the cells. Some have catalytic ability during the metabolic processes and some are in charge of maintaining the structural or mechanical function, at the

same time cell signaling, immune responses are all closely connected with protein functions.

Proteins contain a certain number of amino acids which are encoded in genes. The genes in DNA are firstly transcribed into messenger RNA (mRNA) and the mRNA is loaded onto the ribosome. Every three nucleotides on the mRNA will pair with the anticodon on transfer RNA (tRNA) which carry one amino acid each. A peptide bond is formed between every two amino acid by losing a water molecule. This synthesis process from mRNA to protein in the ribosome is called translation. At the time the protein is synthesised, it also folds into certain three dimensional structure and the protein will be transported to its designated location to function.

1.1.1 Protein Structure

The protein is the fundamental unit of all the biological process happened in the cell during the metabolic process. Understanding the protein structure is the key aspect to know the molecular mechanisms as the three dimensional information for the protein determines how they will interact with other molecules. The protein structure can be classified into four levels.

- **Primary structure:**

  The primary structure of protein refers to the polypeptide chain of amino acids. It can be seen as the one dimensional information of the protein. Expect giving the type of amino acid and the number of amino acid, it also shows two terminals of the chain: carboxyl terminus (C-terminus) and the amino terminus (N-terminus).

  As the amino acid form the protein sequence, they also build up the basic structural unit of the protein, i.e. the bond distances and the dihedral angles make up the shape of the protein as shown in Fig. 1.1 There are three basic dihedral angles called ϕ that control the distance between C′-C′, ψ which control the N-N distance and the ω controls the Cα-Cα distance.

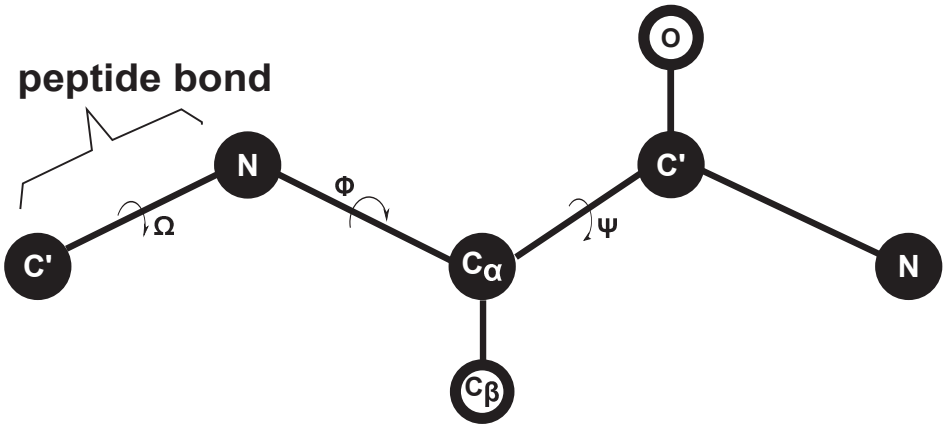

Figure 1.1: Backbone and dihedral angles of protein.

- **Secondary structure:** Protein can form some regular patterns locally which are called secondary structure. They are stabilized by hydrogen bonds between the amino and carbonyl groups and also the neighbouring unit. In general, there are two main types of secondary structures: α-helix and β-sheets. The α-helix is the most common helix shown as A in Fig. 1.2 The backbone in the helix follows a helical path. For each turn in the helix, there are 3.6 amino acids. Another common secondary structure is the β-sheet shown as B in Fig. 1.2 which is occur less often than the α-helix. Its amino acids consecutively connected and extend nearly straight to form a β-sheet. The inter and outer hydrogen bonds of the β-sheet form the β-sheet structure.

- **Tertiary structure:**

  This level of the structure always refers to the three dimensional structure of protein. The interaction of side chains within the protein contribute to the force that form the tertiary structure of the protein. It defines the three dimensional coordinate information. The tertiary structure information reveals out most of the protein functionality.

- **Quaternary structure:**

  The quaternary structure defines how several polypeptide chains form a multi-subunit three dimensional complex. This multiple subunits proteins

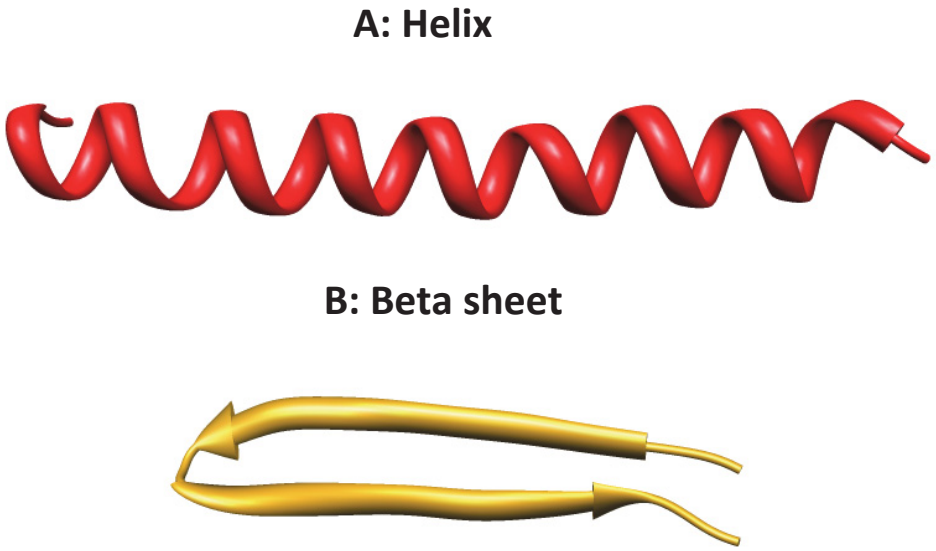

Figure 1.2: Secondary structure element.

have multiple functions like enzymes different parts may have different functions.

1.2 Protein Structure Determination

Nowadays, the number of sequences have been stored in the UniProt [1] is about 86 Mio, however, the number of structures that have been determined and saved in the Protein Data Bank [2] is 102158. The huge gap between these two numbers attracted a lot of attentions from researchers who dedicate their work to structure determination.

The history of protein structure research is accompanied by important discoveries, a new era of rapid progress began when the very first high resolution protein structures was determined: first the oxygen storage protein myoglobin (1958 by John Kendrew) and then the more complicated oxygen-transporting protein hemoglobin (1959 by Max Perutz). In contrast to DNA structures, the first determined protein structures show great irregularities; two structures having various different shapes and different properties. These enormous variations between protein structures show how the molecules can play many different roles in biology and also guided researchers to exploring new directions in the following decades.

In the 1960s, structure determination by X-ray diffraction was used to develop crystallographic electron microscopy that allowed scientists to solve more complex structures [3].

The Protein Data Bank (PDB), which acts as a global database to store 3-D protein structures, has been set up in 1971 as a consequence of the quickly growing number of determined protein structures. Around this time, nuclear magnetic resonance (NMR) spectroscopy was developed as an alternative method to determine protein structures [4, 5]. It uses proteins in solution rather than in a crystallized form and is therefore able to capture the structure in a closer to native state than in crystallography.

Furthermore, the development of clone gene technologies made the production of large amounts of proteins possible further speeding up the growth of structural

information in the PDB. In the beginning of 1980s, researchers started using electron cryomicroscopy (CryoEM or cryo-electron microscopy) to determine protein structures [3, 6, 7]. The 1990s welcomed the structural genomics era, during which thousands of structures were rapidly determined with high-throughput and synchrotron X-ray sources.

Besides experimental structure determination, computational methods such as ab initio methods like global energy optimization with a proper energy function, homology modeling and fold recognition are all widely used today. However, such methods still are not yet reliable enough and therefore far from replacing experimental structure determination and they are often used together with the experimental method to complement each other.

1.2.1 X-ray Crystallography

By far, X-ray crystallography (XRC) is the most dominant technique for solving protein structures. As it shows in the PDB, more than 90,000 structures are determined by XRC which is 88.6% of total structures stored in the database.

To pursue the atomic resolution of protein structure, X-ray crystallography is always recognized as the most efficient method. In XRC, the purified and crystallized protein is exposed to the intense X-ray beam. The crystallized protein diffracts the X-ray beam into different directions which is captured by the screen as several characteristic pattern of spots. After collecting a large amount of patterns by rotating the crystal in the beam, a Fourier-transform method is applied over the patterns to complete the real-space electron density map.

Normally, the recorded pattern only contains the intensity information of the diffracted radiation without any phase information. This is referred to as "Phase-problem" in crystallography. To complement this problem, some experiments (e.g., like isomorphous replacement [8] or anomalous dispersion [9]) can be used to determine the missing phase information. In some cases where some similar

crystal structures have been solved previously, phase information can be derived by molecular replacement [10].

Although XRC is a well matured technique to determine protein structures, there are still limitations. As the final quality of the result highly depends on the condition of the crystals, the crystals is the pivot for the final result. The crystallization process depends on lots of conditions like temperature, pH and buffer concentrations which may result in proteins in non-native states.

1.2.2 Electron Microscopy

Electron microscopy (EM) is the third mainstream technique for protein structure determination. There are 0.8% in the PDB come from this category, however it recently became a more attractive and powerful tool in structural biology area as it yields structures with increasing resolution for macromolecules.

The electron beam as a illumination source will go through the biological sample to obtain images of the sample. Different views of the images are taken to achieve the 3D density map.

In contrast to XRC and NMR that can produce atomic level data, EM data does not allow yet to reach the atomic information.

1.3 Cryo Electron Microscopy

Cryo electron microscopy (CryoEM) is a type of electron microscopy in which a biological sample can be easily frozen by either liquid nitrogen or ethane to study their near-native state. CryoEM is becoming more and more popular in structural biology not only for its flexibility to stain or fix the sample but also for its big span in resolution and size. That covers the range provided by XRC and NMR. Therefore CryoEM is not only a complementary tool to study macromolecules which have difficulty in crystallizing or beyond the size limitation of NMR spectroscopy.

In the beginning, the homogeneous sample which contained in the buffer is spread over the EM grid that is covered by a holy carbon. This carbon film is quickly frozen by plunging into the liquid ethane which suddenly form a layer of vitreous ice that is embedding the particles. These particles were assumed to be randomly distributed of orientations. This step prevents the water from crystallizing, that may cause damage to the sample. The EM grid is then inserted into the microscope and imaged with electrons at liquid nitrogen temperature. All 2D images are taken in the low dose electron conditions which is $20e^- - 25e^-/\text{Å}^2$, so that without causing significant damage to the sample, the signal-to-noise ratio (SNR) is kept in a low state. The 2D images are recorded on film or digital camera such as a charged-coupled device (CCD) or CMOS detector. By recording the 2D images, CryoEM does not have to face the phase problem as XRC.

1.3.1 Single Particle Analysis

The 2D images record a large collection of projections of the same molecule corresponding to different projection orientations. In order to reconstruct a 3D molecule, it is necessary to have an informative data set which typically ranges from 10^4 to 10^5 projections which are selected from several micrographs. They are centered and normalized.

Once particles are picked, they need to be aligned and classified for the reconstruction. The alignment is done via placing the particle into similar orientation. By doing this, the relationship between particles, i.e. the orientation of particles can be used later when making a three dimensional model. Then similar particles are grouped that also separate images that are different. The aligned particles are averaged to achieve higher SNR. After the alignment, the angular information from it can be used for the backprojection [11] procedure that construct 2D images into 3D shape.

The alignment, backprojection and refinement are combined and implemented iteratively for a certain number of rounds until the 3D reconstruction converges to a unimprovable state. The 3D reconstruction is then used for further model building. The interpretation of the density map in this step is crucial for biological information to be concluded.

All the density maps that obtained from CryoEM experiments are stored in Electron Microscopy Data Bank (EMDB) [12].

1.3.2 Limitations of CryoEM

CryoEM has developed dramatically in the last decades especially as a tremendous upswing technology progress in microscope, imaging and computing. The quality of the structure determined by CryoEM is increasing fast. Though, limitations still exist.

The main limitation is that the sample can not be exposed under the beam for long time which may cause sample damage. Because of this, the resolution is limited in the end. Also, to avoid this sample damage, the intensity of the beam is also kept at low dose of electrons. At the same time, biological samples mostly comprised of carbon, hydrogen, oxygen and nitrogen which show weak electron scattering. This results in the low SNR of the images, so that the image contrast is also low and useful structural features may be blurred or difficult to identify.

As the grid will be frozen rapidly, the best case would be using thin sample so that it cools in a short time and stays in the native state. The movement of the sample which caused by either mechanical or electron beam induced effect also limits the resolution.

1.4 Low Resolution

1.4.1 Resolution

Definition of Resolution

In general, the density map resolution is a measurement to quantify the resolvability, i.e. at which level the structural information can be extracted. High resolution data can provide the information about atomic details while low-resolution data show structural details only at the secondary or even only tertiary structure level.

XRC and CryoEM use different definitions of resolution:

In XRC, from the diffraction pattern, which depends on the type of crystal structure, the furthest resolvable peak in it defines the resolution (i.e., the pattern in which the atoms are arranged).

In CryoEM, the SNR of the Fourier components defined the resolution. For this the 2D projections are split into odd and even data sets, for which reconstructions are calculated and compared at different frequency shells in Fourier space [13] independently. Then the resolution is calculated based on the Fourier shell correlation (FSC) [14] as shown in Eq. (1.1)

$$FSC(|s|) = \frac{\sum_i |F_1^i|\,|F_2^i|\,cos(\phi_1^i - \phi_2^i)}{\sqrt{\sum_i |F_1^i|^2 \sum_i |F_2^i|^2}} \tag{1.1}$$

in which i stands for the set of points found at spatial frequency s in the 3D Fourier transforms of the two reconstructions, F_1^i and F_2^i are the Fourier coefficients of the odd and even data set, ϕ_1^i and ϕ_2^i are their corresponds phases. The most popular usage of the FSC is a cutoff at 0.5 [15, 16]. This Fourier correlation coefficient corresponding to the resolution of the density map.

Limitation of Resolution

In XRC, to reach high resolution the molecules have to arrange in a highly ordered fashion in the crystal. Considering the difficulty of different size of molecules, small ones are much easier to crystallize than larger ones, which typically show a higher degree of internal heterogeneity and flexibility. The diffraction data collected in the end is actually an average of all the repeating units, minor differences in the unit cell lead to high resolution peaks in the diffraction pattern to become very weak.

Therefore, in that case, the final electron density map does not contain adequate structural features such as sidechain positions, loop structures or missing density of domains.

Different from XRC, CryoEM typically yields low-resolution (>3.5Å) data. The reason of low-resolution is also from the molecule heterogeneity and flexibility as in XRC. The reason of conformational heterogeneity in the assemblies is the fluctuations of the structure around the native state. Comparing with XRC, single particles from CryoEM are imaged independent of their flexibility and it has been proven helpful in determining macromolecular assembly structures at resolutions lower than 5 Å [17, 18].

During the processes of reconstruction, the assembly of particles in different conformations and functional states are averaged. The variations between particles will be averaged out which yields a blurred density and leads to a limited resolution. However, the variations contain a lot information about the conformational variability, but to extract is information is a challenge in computational work. To solve this problem, different sorting algorithms, either supervised or unsupervised, are used [19]. By classifying 2D images into various conformational states which are then reconstructed individually higher resolution density maps can be obtained. The algorithms designed in this active field of research still have much more space for improvement.

Visualizing the effect of resolution

Fig. 1.3 depicts the density maps of calmodulin at different resolutions Calmodulin is a calcium-binding messenger protein; it exists in most organisms and plays an important role in transducing calcium signals that functions in processes such as inflammation, metabolism and immune response.

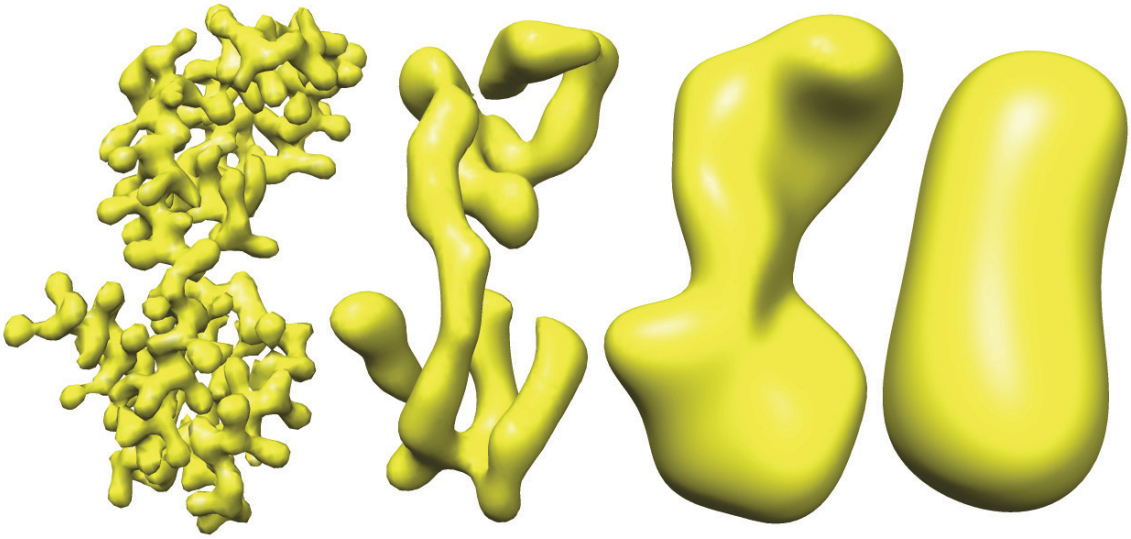

Figure 1.3: The structure of calmodulin shown at different resolutions. From left to right, 4 Å, 8 Å, 16 Å, 32 Å.

From the left, that has the highest resolution of 4 Å at which most of the residues can be visualized. The density becomes fuzzier with decreasing resolution. The rightmost structure (Fig. 1.3) only shows the global shape of the structure.

As the resolution in XRC and CryoEM are defined differently, it may be conclude that an XRC map contains more information at the same nominal resolution; however, a comparison study [20] showed that the 7.5 Å E.coli 50S subunit from CryoEM looks better than 7.8 Å Thermus thermophilus 70S ribosomal from XRC, which is likely due to the fact that CryoEM also provides phase information.

Low resolution data

Nowadays, at high resolution (<3.5 Å) useful atomic models can be built from XRC which means sometimes that atomic detail information can be obtained from the electron density with even sub-atomic precision (like the hydrogen bonds).

To investigate the detailed information of the protein, high resolution is needed, however to get high resolution data for all systems is not possible, especially as the complexity of the testing structures is increasing and thus the probability of generating low resolution structures (>3.5 Å) is also getting higher. This rapidly increasing amount of low resolution data giving an important source of biological information. In the early days, these low resolution data were often ignored as they are not that informative as higher resolution ones. However as structural biologist are trying to solve more complicated and larger complexes, low-resolution data are more and more considered as a important to provide useful biological informations. As a consequence of this, the amount of structures determined at low resolution has dramatically increased in recent years. This can been seen from the low resolution structures deposited in the PDB (shown in Fig. 1.4). 55.6% of all structures stay in the intermediate resolution which is between 3.5 to 10 Å. In recent years, the developments in both experimental and computational techniques profit the determination of low resolution structures. It makes the interpretation of low resolution data increasingly feasible.

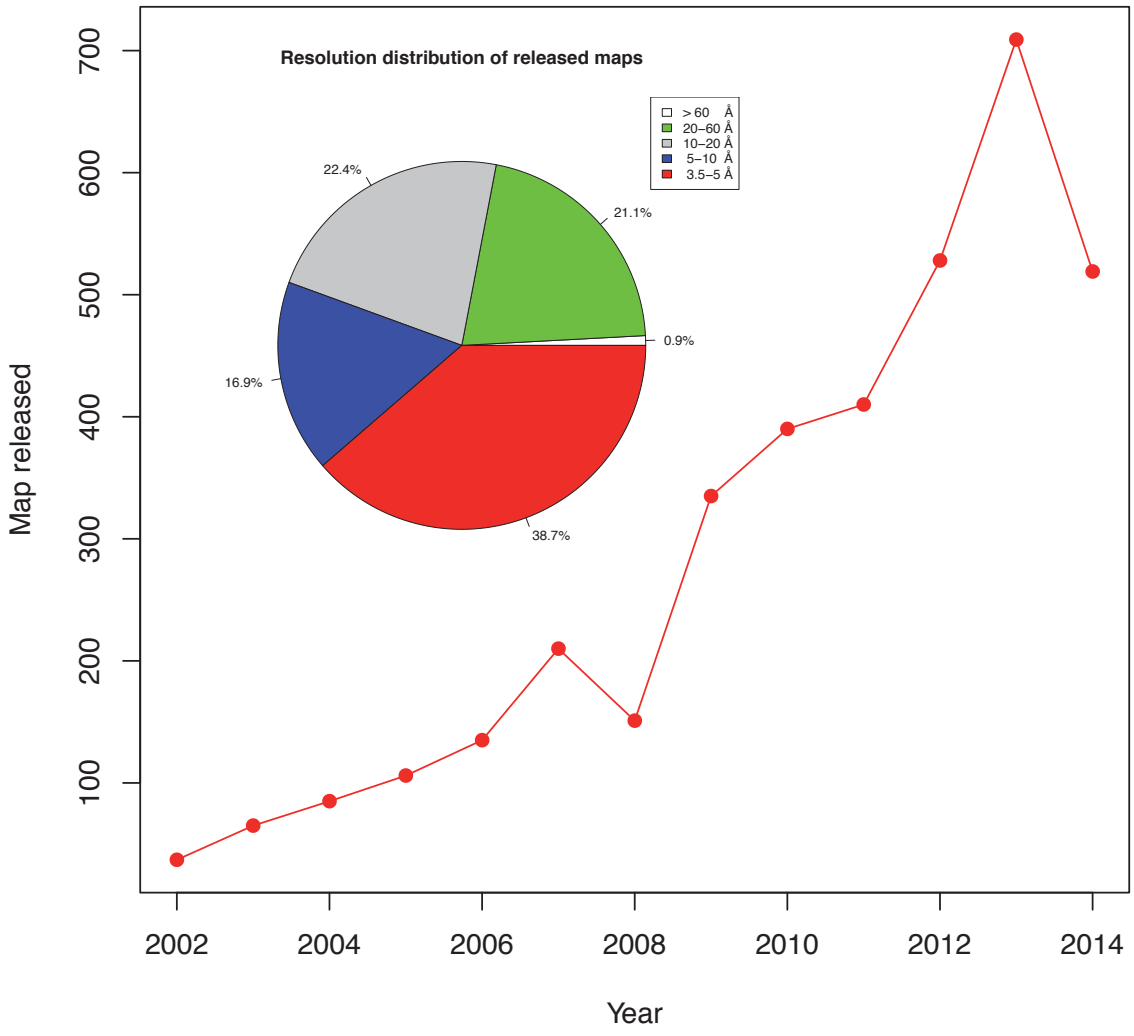

Figure 1.4: Low resolution statistic. The number of new structures deposited each year in the Protein Data Bank which have a resolution worse than 3.5 Å plot as the red curve. The pie chart illustrate the numerical proportion of different resolution groups.

1.5 Interpretation of Low Resolution Data

As the number of low resolution-data is increasing quickly, building models from low-resolution data gets much more attention than ever before. Different resolutions correspond to different levels of information content. For different amounts of structural information, people developed various methods to solve the model.

For low-resolution data, obtained from CryoEM, is typically in the range from about 6 to 20 Å. Normally, at such a low resolution, it is not possible to build the

atomic structures directly. In those cases, the data is normally interpreted by using high-resolution structures determined by either XRC or NMR [21]. Originally, the high-resolution structures are docked into the density map manually. As the method is developing, it comes to two classes of fitting techniques: one is rigid body fitting that places the high-resolution structure as a rigid template and search the best fit within the volume; the other one is flexible fitting where different levels of flexibility are applied to the high-resolution structures to fit into the density.

1.5.1 Rigid Body Fitting

In this method, people are trying to optimize the six dimensional parameters which are translation of three directions (x,y,z) and rotations of three angles. Two important factors in this method are the evaluating scoring function and the search space in the density map. EMfit [22, 23] is one of the earliest programs doing rigid body fitting. By analyzing the symmetry information, it searches in a predetermined interval to refine the positions. Situs [24, 25] uses a positional vector to represent the local density by which a fast density comparison can be made without superimposing the density maps explicitly. Gmfit uses a coarse-grained model to represent the density map which improves the fitting procedure [26]. The EM map and atomic structures are represented as a Gaussian mixture which shows the shape of a density map. MultiFit [27, 28] uses a divide-and-conquer approach for searching the space to fit multiple domains into the density. When fitting multiple proteins into the density, segmenting the map first is very helpful to identify the location of each individual component.

As the resolution of the map is improving, conformational change between the target and the known XRC structure become large, and a basic rigid body fitting is not sufficient to interpret the density map. Therefore, more flexibility is needed for the fitting.

1.5.2 Flexible Fitting

A protein itself is a flexible molecule. During its functioning in the reaction, it undergoes conformational changes. In this class of fitting methods, different levels of flexibilities are considered to sample different conformation states. Flexible fitting has been proven to be helpful in understanding conformational changes at a near atomic level of detail.

Several programs have been used to introduce the flexibility. The traditional molecular dynamic (MD) simulation is used to handle the protein flexibility. Molecular dynamics flexible fitting (MDFF) [29, 30] is a method based on MD simulation. Starting from the traditional MD force field, it added two new terms: one gives the force for moving the protein to the density map and the other term tries to maintain the secondary structure. The combination of real-space refinement and molecular dynamics simulations have been used to fit the structures into the density map (RSRef/X-PLOR:RMSD) [31, 32] Different from MD-based methods, Flex-EM [33, 34] uses coarse-grained optimization combined with the conjugate gradients minimization (CG) and simulated annealing molecular dynamics to achieve flexible fitting.

Normal Model Flexible Fitting (NMFF) [35, 36] and NORMA [37] have used low frequency normal modes (NMA) [38–40] analysis, particularly on Elastic Network Models (ENM) [41] to follow the dynamic components in the context of density map. NMA is used to identify flexible regions and the principal motion directions of atoms or residues. ENM can use a coarse-grained model and applies a harmonic potential on atomic distances to explore a large conformational space [42]. The combination of comparative modeling (based on alternative alignments and loop conformations) and structure refinement was used to improve the sequence alignment and obtain better homology models [43].

Except using dynamics calculations for protein flexibility, S-flexifit [44] derives the structure variations from the protein in the same superfamily according to the

CATH protein structure classification database and uses this information for the flexibility.

Besides using high-resolution structures as a starting point, homology models could also be used to fit into the density map. When the high-resolution structures could not be used directly, then the homology model could be used as a complementary way to refine a structure [45, 46]. Recently, the deformable elastic network (DEN) refinement [47] has been introduced to fit models into density maps, allowing large deformations in the atomic structure model [48]. The DEN defines springs between randomly chosen atom pairs that have a distance value that falls into an interval between 3 and 15 Å, using a reference model (i.e. a high resolution structures) as a template. During the refinement the target distances of these DEN restraints change according to two types of forces: 1) restoring forces which pull the restraints towards a reference model, and 2) trailing forces which follow the current structure as it is being pulled into the density map, which controls how the reference structures move by following the motion of the structures. The DEN restraints are therefore able to balance forces from the reference model and from the density map. The DEN approach was implemented into the program DireX [47], which combines robust real-space refinement with an efficient conformational sampling algorithm.

1.5.3 De Novo Model Building

The resolution of the map determines what kind of method can be used to interpret the density map. Table 1.1 gives an overview of the correspondence among resolution range, structural information content and the method that can be used for interpreting the density maps.

Resolution(Å)	Visible features in the density Map	Method for model building
>10	Shape of the tertiary structures	Rigid body fitting
10–7	Domains and α-helices	Flexible fitting
7–5	α-helices and β-sheets	Flexible fitting
5–4	Possible to trace main chain with prior information. Large stretches of α-helical region will increase the success rate. Small molecules will be visible and gross interpretations can be made in certain cases (e.g. ADP versus ATP). Some side chains may be visible.	Refinement is possible with secondary-structure restraints. Reasonable R/Rfree values and geometry can be achieved.
4-3.5	Main-chain trace more reliable. β-Sheets can be built with more confidence. Conformational flexibility can be assessed. More side chains visible. When two or more states are crystallized, conformational changes in side-chain positions may be visible.	Standard refinement techniques can be applied
3.5-0.5	Fold information,clear loop regions and sidechains	De novo method can be applied

Table 1.1: Effect of resolution on density map interpretation and modeling approaches

When the resolution of the map is less than 15 Å, then density map defines only the general shape of the protein without any details. Therefore, no information for the residue or atom level can be used, rigid body fitting would be recommend. As the resolution goes higher, more structural information can be derived from the map such as the domain, secondary structure and even the residue content. For those cases, flexible fitting can be applied.

When the resolution is high enough (<3.5 Å), it is not necessary to use a initial model or known structures, rather, a de novo method can be used to build an atomic model, i.e. based on the density map alone. There are some programs from XRC that works on de novo modeling such as ARP/wARP package [49–51] , SOLVE/RESOLVE [52] and TEXTAL [53]. These methods are mainly aim at the resolution range smaller than 3.5 Å. However, as the size of testing molecule is increasing, most of the resolution stay in the low-resolution range. From which, secondary structure can be seen and in some near atomic resolution (3.8-4.5 Å)

maps side-chains can barely be seen. For this resolution group, de novo modeling of protein structures is often combined with the analysis of secondary structure features from the density map such as SSEHunter [54], SSELearner [55], SSE-Tracer [56] and VolTrac [57]. It identifies the secondary structure segments according to the geometric shape of the map. By matching the sequence secondary structure information onto the density segments, a secondary structure based backbone can be build [58–60], which can be further optimized with other programs like Rosetta [61]. Gorgon [62] is a program using pattern matching and geometry processing algorithms to model protein structures. However, without an accurate SSE identification, a possibly wrong structure is built. This type of modeling can be time consuming as it requires intensive human interactive manual work. Pathwalker [63] is a program doing de novo backbone tracing program which is based on the algorithm for solving the Travelling Salesman Problem (TSP) and trying to find out the right backbone connection. However, this work quite depends on the TSP solver which only tries to find out the shortest path. Additionally, an assumption of known location of C and N terminus has to be made before hand. During the model building process, manual work is still greatly involved.

CHAPTER 2

Method

2.1 Density Maps

2.1.1 Density Maps Simulation

For testing purpose of our method, the calmodulin structure (PDBID:1S26) was selected for analysis. Only chain D was kept by deleting all the calcium atoms. The remaining model has 143 amino acids and contains both helices and sheets both. This atomic calmodulin structure is used as our target structure for testing purposes. With this structure, 5 groups of noisy density maps were simulated at 4, 5, 6, 7, 8 Å (1 Å/pixel) resolution respectively with Direx [47].

2.1.2 Map Filtering

To each of these five maps, a filter was applied. The filtering was implemented in Fourier space by using a low-pass filter at a certain resolution. That is similar to use a Gaussian filter on each grid point in real space. The rationale behind this is the implementation in the Fourier space is much faster and how much the map is filtered can be easily and numerically determined. The aim of this approach

is to smoothen the density map, and at the same time by filtering, some high-frequency noise can be deleted. This would approximate a realistic experimental map showing the general shape of the protein but with some of the side chain information eliminated from the map. The smoothed map is helpful for placing our pseudo Cα atoms in the later step.

2.1.3 Map Normalization

To make all the input density maps comparable, we normalize the input density map to a uniform scale. The normalization was implemented according to

$$\rho_{norm} = \frac{\rho_{dens} - \mu}{\sigma} \tag{2.1}$$

in which ρ_{dens} is the density value, μ is the mean value of all the density and σ is the variance of the density map. After the normalization, different density maps with different value ranges were scaled so that they are comparable.

2.1.4 Calmodulin Test Result

For the calmodulin test, 5 groups of maps with resolutions ranging from 4 to 8 Å were generated. Fig. 2.1 A shows for example the 6 Å simulated density map. Noisy density maps were simulated with EMAN [64] based on the noise free density maps which were generated by Direx [47]. Fig. 2.1 B shows the 6 Å noisy density map.

Afterwards, we filtered the maps to a relatively lower resolution than the corresponding original maps. For these 5 groups of noisy maps, we filter them all to 8 Å. Fig. 2.1 C, shows how much the maps are different from each other with the same noise level but at different resolutions. Fig. 2.1 D shows all those maps after normalization and filtering. After filtering to 8 Å, they all look very similar to each other as they contain the same amount of the structure factor information.

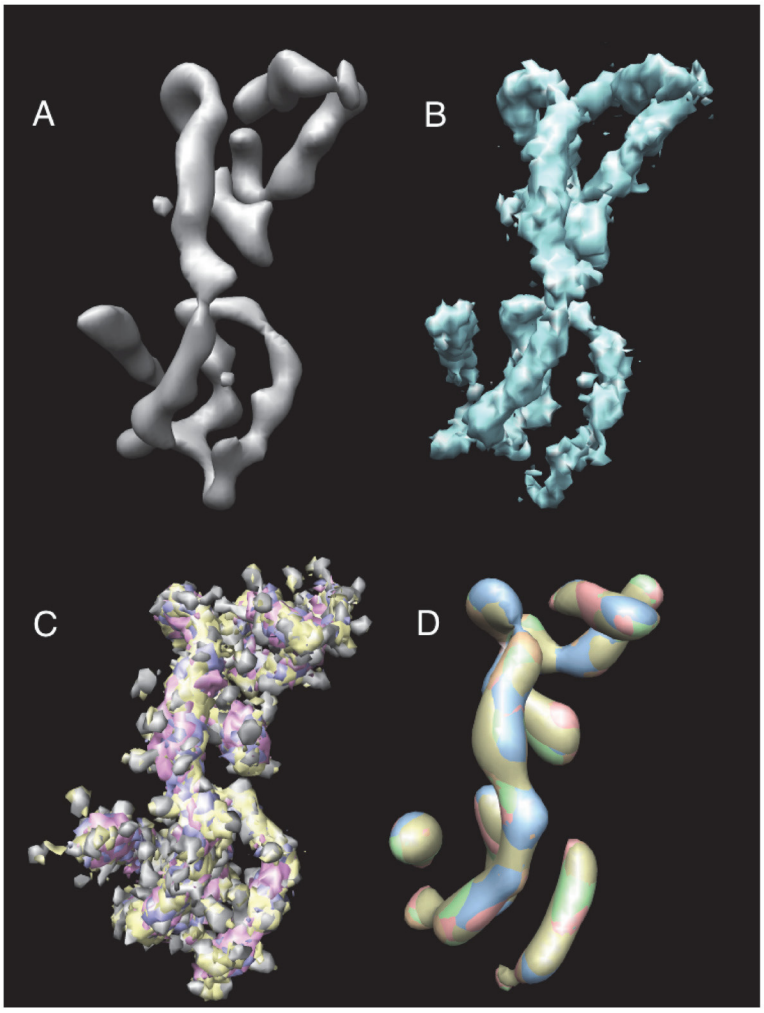

Figure 2.1: Synthesis and processed density maps. A: Synthesised 6 Å density map. B: Noise added on A. C: All noisy maps. D: Maps after normalization and filtering.

2.2 Optimization

With our simulated map and the corresponding bead model, we are still not able to figure out how the trace looks like and how these beads are connected. For this purpose, we have to find a smart way to describe the right connections which represents the Cα atom trace in the backbone. Here we generalized all the methods applied in the process as an optimization step. These methods include the Lin-Kernighan heuristic (LKH) [65, 66], Monto carlo method (MC) [67], simulated annealing (SA) [68], Hungarain algorithm [69] and also a combination of different restraints. These methods together build up the entire backbone tracing algorithm.

2.2.1 Travelling Salesman Problem and Lin-Kernighan Heuristic

If a salesman is given a list of cities $\{c1, c2, \cdots, c_N\}$ and the distances $d(c_i, c_j)$ between each pair $\{c_i, c_j\}$ of cities, how could he visit each city exactly once and return to the starting city with the shortest distance? This is the travelling salesman problem (TSP). Our goal is to find an ordering δ of the cities that minimizes the function

$$l(\delta) = \sum_{i=1}^{n-1} d_{\delta(i),\delta(i+1)} + d_{\delta(n),\delta(1)} \tag{2.2}$$

The final purpose is to find the shortest path with the best method δ. Nowadays, as a combinatorial optimization problem, TSP has become more attractive especially as the computing ability dramatically increased during the last decades. It becomes attractive not only because it is interesting but also for its availability for a large amount of real world problems. It is highly applied to a variety of aspects, besides working on searching for the shortest tour, it is also used on machine scheduling [70], data analysis in psychology [71] and X-ray crystallography [72]. In biology, Korostensky et al [73] used TSP to compute a near optimal multiple sequence alignment.

Although the problem is easy to state, coming to the solution is quit another level of difficulty. TSP is a non deterministic polynomial time hard (NP-hard) problem, it means the problem can not be solved in polynomial time. As the number of cities increased, it is unlikely to guaranteed any efficient exact algorithm can be used to find the optimal tours. These question make people put all efforts on finding out an approximation algorithm instead of a exact optimization algorithm. These algorithms normally compromised between running quickly and finding the optimal tours but not doing both at the same time. The 'heuristic algorithms' was designed to find the near optimal tours but relatively larger number of cities

than the exact solutions. Among many different heuristic algorithms, the Lin-Kernighan Heuristic algorithm [65] was one of the best which has a big influence over the later algorithms design for TSP.

Lin-Kernighan Heuristic

Lin and Kernighan developed a heuristic algorithm in the 70s. It is a well known method based on edge exchange procedures. In general case, k edges in a tour are exchanged with k edges that were not in the tour. This procedure tries to shorten the tour. The exchange procedures are referred to as $k-opt$ procedures where k is the number of edges exchanged at each step. Fig. 2.2 shows the case when $k = 2$. The edge x_1 and x_2 are removed and replaced by y_1 and y_2. This $2-opt$ move is the fundamental step of the whole Lin-Kernighan heuristic algorithm (LK). The LK algorithm is a dynamic procedure which performs a series of $2-opt$ moves. The $2-opt$ moves are executed in a successive order. The edge which connected from level i will be deleted in the next level $i+1$.

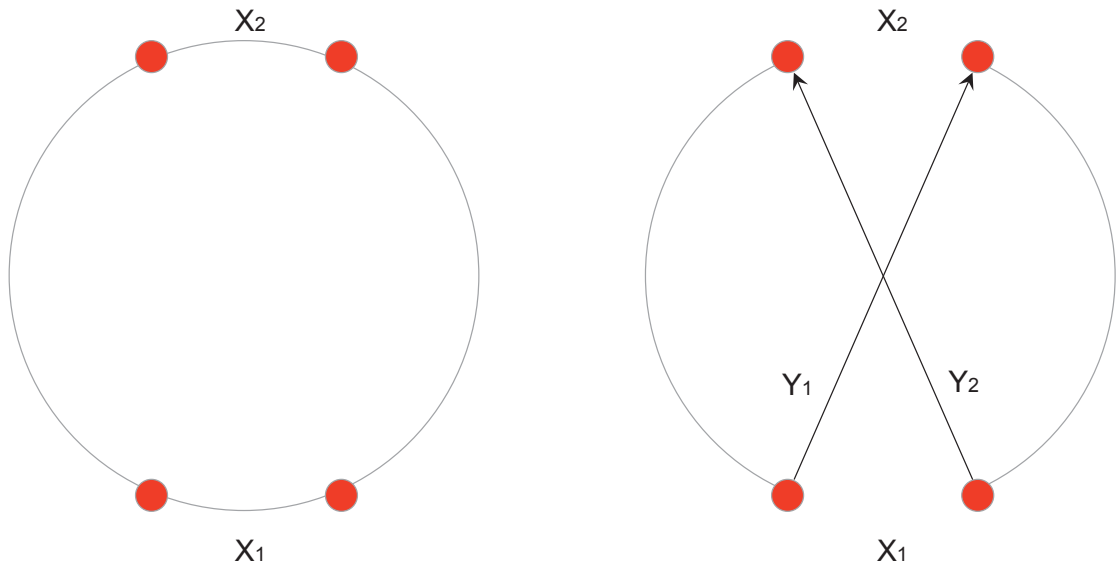

Figure 2.2: One step 2-opt move. The red circles represent the beads and the connection curve between them describe the edge in between. Arrows are pointing to the direction of the trace connection.

Here the basic classical Lin-Kernighan procedure is sketched in Fig. 2.3 and it works according to the following description:

1. For the visualization purpose, the initial random connection tour is shown as a closed circle, as the real connection picture is difficult to distinguish the ongoing changes. Panel (A) in Fig. 2.3 gives the exemplified circle connection. In this depiction, the lengths of the edges here are not indicating the real lengths between points.

2. A random point t_1 is selected as the starting position and x_1 (red edge in Figure 2.3) is one of the edges adjacent to t_1. t_2 is the other end of x_1. From t_2, we delete the connection x_1 and make another new edge y_1 (blue edge in Figure 2.3) which does not exist in the current tour and connected to t_3. And it also satisfies the equation $g_1 = (x_1 - y_1) < 0$. The tour was shortened by this exchange.

3. However, the tour right now is not a closed path. To make a reasonable new tour, the adjacent point t_4 of t_3 is selected which connected by edge x_2. There are two choices for t_4, but to close up the tour only one choice can build a full path, that is by connecting t_4 and t_1, which gives edge E_{last}. The other choice would split the tour into two sections that would destroy the connectivity of the entire path as shown in C of Fig. 2.3. At this point, by substituting x_2 with E_{last} (dotted line in panel B) of Fig. 2.3, the new cost of the exchange is

$$G_{last} = x_2 - D_{last} \tag{2.3}$$

   By now, the whole process shows how the $2-opt$ work as in Fig. 2.2. The one step $2-opt$ cost is calculated by

$$G_2^* = g_1 + G_{last} \tag{2.4}$$

   No matter how we do the $2-opt$ exchanges, we have to make sure the G^* is always positive.

4. $2-opt$ move is the basic step of the algorithm. Without stopping by two exchanges, the algorithm goes on with further moves. Instead of connecting

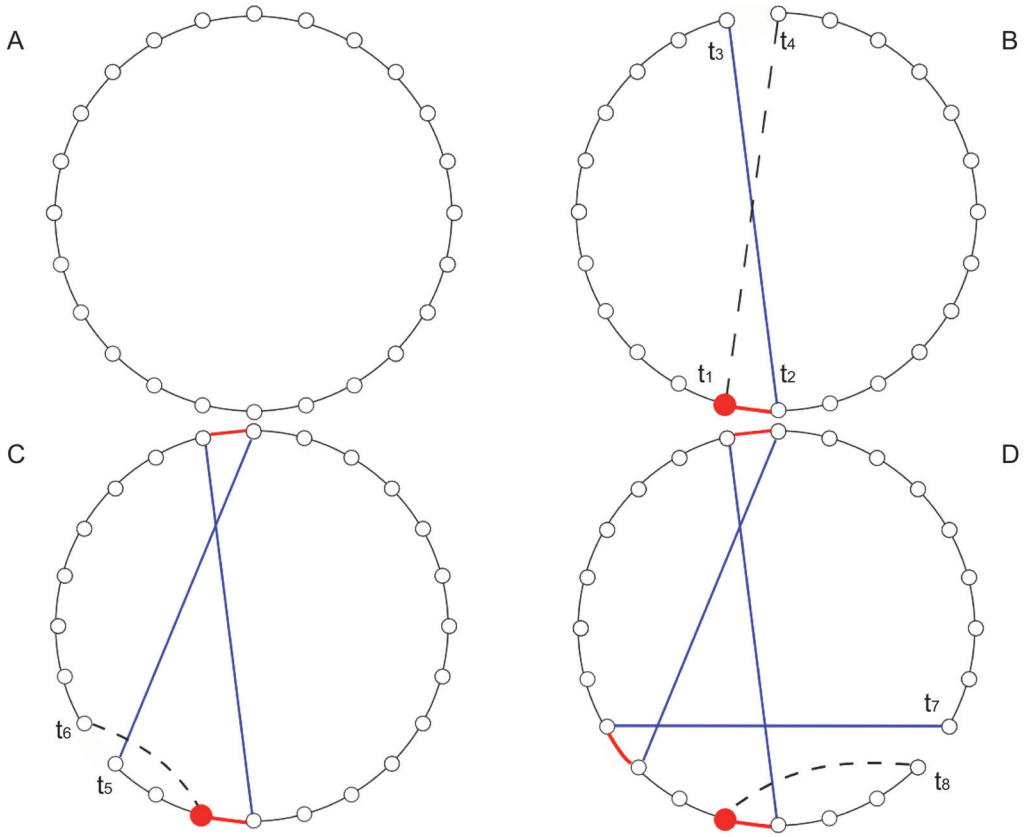

Figure 2.3: LKH. The four pictures shows the basic steps of LKH. Fig. A shows the simplified three dimensonal random connected points as a 2-D circle. Fig.B shows the 2-*opt* in which connections t_1t_2 and t_3t_4 are replaced by t_2t_3 and t_4t_1. Fig. C shows the 3-*opt*. In case that 2-*opt* does not improve the path, instead of closing t_4t_1, LKH try to do a further step of exchange by connect t_4t_5. Afterwards, finish by closing t_6t_1. Fig. D show the 4-*opt* which is similar to the 3-*opt* procedures.

E_{last} in step 2, we try to get another point t_5 on the tour. It connects to t_4 by edge y_2 (the new blue line in panel C of Fig. 2.3). The edge x_3 from t_5 reaches another adjacent point t_6. Similar as in step 2, the connection of x_3 is removed and substituted by the connection from t_6 to t_1, which is the new E_{last}.

5. Now by completing the new second exchange ($x_2 \rightarrow y_2$) and the closing up exchange ($x_3 \rightarrow E_{last}$), the one step $3-opt$ has been done. The cost of the second exchange can be calculated by this equation:

$$g_2 = x_2 - y_2 \tag{2.5}$$

The cost of the closing tour exchange is calculated by:

$$G_{last} = x_3 - D_{last} \tag{2.6}$$

The overall cost of one step $3-opt$ is calculated from:

$$G_3^* = G_2 + G_{last} \tag{2.7}$$

, in which $G_2 = g_1 + g_2$

As long as G_3^* is greater than zero, the algorithm implements one step $3-opt$ move accordingly.

6. The above procedures describe the $2-opt$ and $3-opt$ moves. Those are the two basic steps of the LK algorithm. However, from here on, we can keep on running the same moves further. That expands the whole algorithm to the $k-opt$ move procedure, which means the number of exchanges are not defined before but determined during the implementation of the algorithm. Deducing from the above procedure, the entire procedure has to satisfy the criterion

$$G_i^* = (G_{i-1} + G_{last}) > 0 \tag{2.8}$$

, in which $G_{i-1} = \sum_{n=1}^{i-1} g_i$, which also has to be positive. The rational idea behind this is that if a sequence of numbers has a positive sum, there is a cyclic permutation of these numbers so that every partial sum is also positive [66]. This positive gain criterion makes the algorithm know when to stop and work efficiently.

2.2.2 Modifications of LKH

In our backbone tracing case, we first build a bead model as described in Section 3.1 already. However, there is no information on the connection among those beads.

To find out the right trace, we have to figure out how these beads are connected to each other. If we think of each of those beads as a single city or point, then we can directly map the protein backbone tracing problem to the TSP. Based on this, we decide to use the LK algorithm to solve the tracing problem. As we are trying to solve a protein trace, which is not the same as the TSP, we did some modifications on the original LK algorithm that makes the algorithm fit to our purpose. In this paragraph, we introduce what modifications we made on the original LK algorithm.

1. **Modified Lin Kernighan (MLK)**

   Firstly, we reserve the basic $k - opt$, and the LK procedures as it states in the previous part. That controls the changes of the whole tracing algorithm during the optimization. However, as the k value is increasing ($k >= 3$), there are cases that the moves are infeasible, that means the move results in a disconnected tour. In this situation, the original algorithm temporarily allows the infeasible move, and then makes the reconnection in the later exchange. To consider those infeasible moves, all the non repetitive ways of connections have to be counted. If k edges are broken in a tour, there are $(k-1)!2^{k-1}$ ways to make the reconnections to form a valid tour [74], however not all of them are considered [75]. For example, if $k = 3$ which means break 3 edges, so in total 8 cases can be built but only 4 of them actually contain all new connections. Similarly for $4 - opt$ moves, among all the 48 different ways of reconnections, there are 25 tours that contain only new edges. As the k value is increasing, more infeasible moves have to be considered. Dealing with the infeasible moves dramatically complicates the original LK algorithm.

   To avoid those infeasible moves, we decided to use the modified LK method, which was introduced in 90s by King and Andrew [76]. The main idea is to try to avoid the infeasible moves by extension of the $2 - opt$ LK method. The following description shows the main procedures:

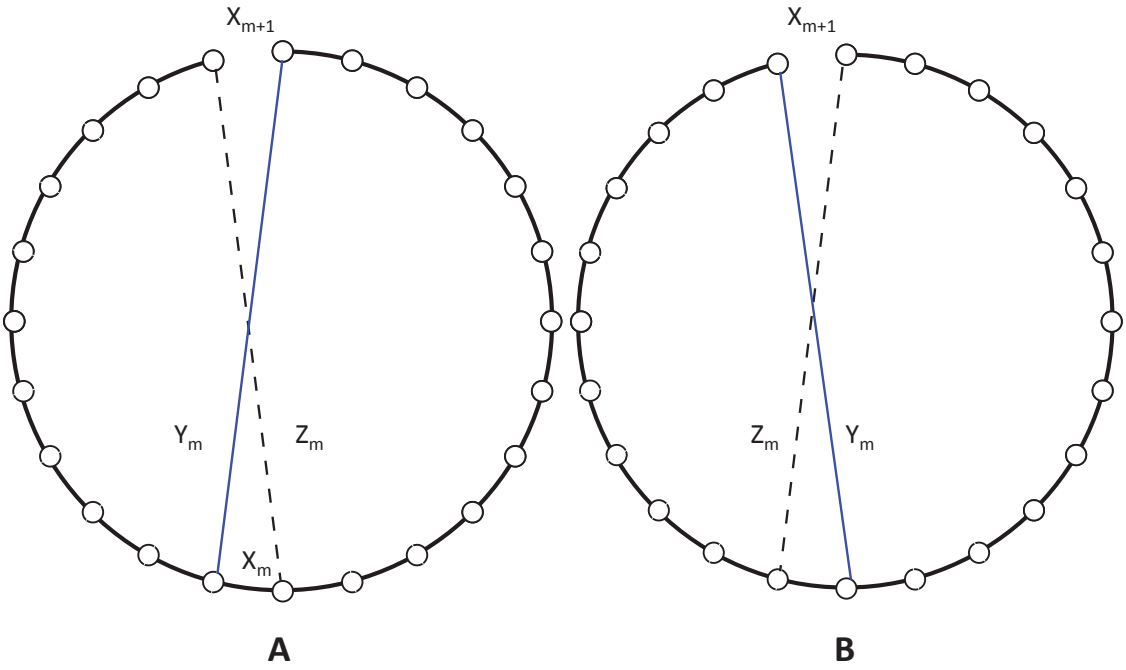

Figure 2.4: MLK. The MLK choose a better path from the two symmetric connections as shown in A and B. The accepted better connection is shown in blue and the dotted line shows the ignored connection which is used as the base for the next round of 2-*opt*.

- When we do the $2-opt$ moves, we need to choose two edges to break at the same time reconnect the other two edges as described in the Lin-Kernighan Heuristic of item 1 to 3. The newly built edges can be seen in a symmetric way as depicted in Fig. 2.4. We start from the first connection x_m which is recognized as the base edge. Then we can get the other two points and three edges y_m, z_m and x_{m+1}. The new named z_m is similar as the E_{last} in the original $2-opt$ method. From a of Fig. 2.4, one can clearly see the symmetry between y_m and z_m.
- Based on this symmetry property, we calculate which would make the most gainful exchange either between x_m and y_m or x_m and z_m. If there is no gain for either of them, we need to reselect other starting points. If there is one gain better than the other one, then we perform the basic

$2-opt$ move by deleting the base and connect y_m or z_m depending on which is more fruitful. At the same time, x_{m+1} is also added and the left one of y_m or z_m becomes the new base for the next iteration $m = m+1$.

2.4 B illustrated how the aforementioned procedures work out without generating any infeasible moves intermediately. On the left of that, A shows how it looks with the normal LK, how can we get that result with the infeasible move.

2. **Longest Edge Identification**

The original LK algorithm deals with the TSP. In the TSP, the salesman have to leave from one city and come back to the same city in the end, that means the path which is travelled by the salesman is a closed path. In our case, we are processing the protein backbone, which normally have two termini one is the N-terminus and the other is C-terminus. They are disconnected in the protein structure. Additionally, in the real Cα backbone of one protein, the longest connection always exists between these two termini. We implemented a step that is embedded in the LK procedure so that during the optimization the connection between the these two termini, that should not exist can be distinguished automatically.

In Section 3.1, while placing the beads, we simulated one more bead than the right number of the amino acid in the protein sequence. This bead is called a phantom bead, because we do not really give a real position for that bead in the space. It exists as a non-real atom. We only assign connections to it. In the starting step, we define both lengths of the edges that connect to the phantom bead as 0. The purpose of doing this is that in the optimization procedure the longest edge will tend to be connected with the phantom bead, as when this happens both connections to the phantom will be considered as 0 and it shortens the path most comparing with other phantom connections. After the optimization, we can delete the edges that are connected to the phantom bead and get the shortest path of the rest without returning back to the starting bead. By this step, the final result just matches the Cα

backbone structure which without the longest connection between C and N terminus.

2.2.3 Monte Carlo Optimization

The Monte Carlo method is defined as a branch of computational algorithms that is based on using random numbers, for the solution of numerical results. In general, the random numbers are independent random variables uniformly distributed over the unit interval $[0, 1]$. The MC method taking into account the uncertainty of models and determine the probability of different outcomes.

The MC method has become a powerful and commonly used technique for analyzing complex problems. It has been frequently applied to various fields from physical sciences, engineering to financial business. Among them, one powerful and popular application is in optimization. The problem is trying to find the extrema, which is either minima or maxima of functions. The early MC method implemented in computer was developed by Metropolis in the 50s.

In the classic Metropolis method, one configuration is created by the previous state with a transition probability which determined by the energy difference of two states. Different states follows a MC time step. The time dependent behavior according to a master equation in equilibrium condition is given by

$$P_n(t)W_{n\to m} = P_m(t)W_{m\to n}, \tag{2.9}$$

in which $P_n(t)$ is the probability of state n at time t, and $W_{n\to m}$ means the transition rate for $n \to m$. The probability of nth state can be expressed by

$$P_n(t) = e^{-E_n/k_BT}/Z \tag{2.10}$$

Because the partition function Z is typically not easily accessible, the probability can not be calculated directly. However, if the nth state come from m state , the

relative probability is proportional to the probability of the states, which means the denominator can be ignored, only considering the energy difference of two states

$$\Delta E = E_n - E_m. \tag{2.11}$$

The main implementation of Metropolis algorithm can be described by a simple recipe:

Metropolis Monte Carlo scheme

1. Choose a starting state A
2. Make a change from state A to state B, calculate the energy difference $\Delta E = E_B - E_A$ by doing this change
3. If $\Delta E < 0$, accept the change. Make B as a starting state and go to step (2)
4. Generate a random number p such that $0 < p < 1$
5. If $\Delta E > 0$ and $p < \exp(-\Delta E/k_B T)$, accept the change and do as (3), otherwise keep the state A and go to step (2)

- **Simulated Annealing (SA)**

  During the Metropolis optimization, it may take quite a long time to converge or be stuck in a local minimum state especially for those problems which have a great amount of degrees of freedom. The efficiency becomes a key issue in the optimization. We choose this simulated annealing method to improve the efficiency.

  Annealing is a technique commonly used in material science, that involves heating and cooling the material to change the structure so as to change

physical properties. This procedure is also a manner in which crystals often reach the state of minimum energy. If the temperature is high, molecules move freely. Reversely, at low temperature, molecules are tending to be stable as frozen. By slowly cooling the material, the system tend to move towards the minimum energy state.

Simulated Annealing has analogy concept with thermodynamics in material science. It is introduced by Kirkpatrick *et al.* [77] as a method used for the global optimization problem, especially for those where the global optimum is hidden among many local extrema. It is a modification based on the Metropolis algorithm.

The method starts at a relatively high temperature, which means the algorithm accepts worse solutions more often than the same it would at low temperature. As the temperature is decreased, more unsuitable solutions will be rejected so that when the temperature is low enough, only the good solutions will be accepted. This procedure lets the algorithm to have the ability to jump out of a minimum and then moving to another area which could be closer to optimum result. This cooling makes the optimization process to work more efficiently.

As the simulated annealing is implemented on the basis of the Metropolis algorithm, the biggest difference is that in the method with simulated annealing the temperature parameter is decreasing gradually and is not as constant as it is in the Metropolis algorithm. Considering the detailed modification according to the Methropolis algorithm procedure shown in the box above, we only need to make a small change like the following

Modification of SA on Metropolis Algorithm

. . . accept the change and **lower the temperature**,. . .

To make use of the simulated annealing MC method, the whole system have to satisfy the following four elements:

- **Initialization:** The starting model stays in a certain state and can be further improved for a good answer.
- **Modification:** A set of random changes are allowed over the initial configuration, potentially this change may generate all possible configurations while the temperature is decreasing.
- **Function:** A function gives out how good a certain configuration is after the change. Optimization of the output of the function is also the goal of the procedure.
- **Cooling:** It tells how the annealing procedure works. What should be the starting hot temperature, how much it should be lowered and when to stop has to be included here.

2.3 Refinement

All models of proteins contain errors. For experimentally determined models, these errors originate from errors in the data and its interpretation. For computationally predicted models, error arise from inaccurate energy function and insufficient sampling. In structural biology, no matter experimental or computational data, the models from those methods contain noise which generate structures with errors in different levels. To reduce these error, a general refinement procedure is needed. Nowadays, a number of methods exists for protein structure refinement. For structures produced by XRC or CryoEM, there are two different ways for the refinement:

- Force field driven refinement with experimental information;
- Rigid fitting refinement.

Force field refinement uses restraints for example for bonds and torsion angles restrains over the structures. The starting model for these methods do not have to be a high resolution structure, however, as the conformational change between starting model and the target is large, each small step in the molecular dynamic simulations will burden the computation.

The rigid body fitting is based on an existing structure. Splitted domains of this structure will be fitted independently. The rigid body fitting conserves the secondary structure information but of course does not allow the reasonable deformation changes within the domain.

For a better refinement, we can use an intermediate between both of these methods to refine the structure flexibly as introduced Section 1.5.2. However, since often the number of parameters (atomic coordinates) is much larger than the number of experimental observables, overfitting may happen during the fitting. To avoid overfitting, we choose DireX for model refinement, which can perform a good flexible fitting at the same time avoid the over-fitting problem by employing deformable elastic network (DEN) restraints.

2.3.1 Refinement with DireX

DireX is a real-space refinement program, which avoids over-fitting by refining only those degrees of freedom for which the density map actually provides information. To refine a starting model into the density map, DireX uses a very efficient geometry- based conformational sampling algorithm to yield ensembles of structures which are biased by different restraints on chemical bond, bond angles and planarity. Additionally, forces are that derived from the difference between target density and the calculated model density, which moves the atoms into the target density. Instead of controlling the degree of freedom of information from the density map and the initial structure manually, Direx adapts the restraints to the forces generated by the density map automatically.

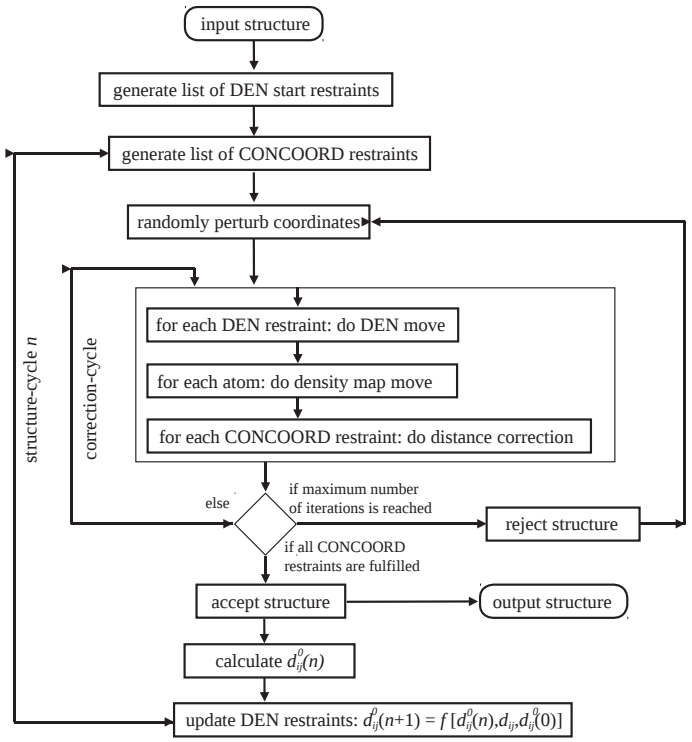

Figure 2.5: Workflow of the DireX. Diagram showing the refinement algorithm in DireX

Fig. 2.5 shows the general workflow. During the DireX refinement, three main forces are used that drive the conformational changes and eventually refine the starting structure into the density map:

- a conformational sampling algorithm which generates a random walk,
- a stochastic gradient of electron density map that moves the model into the density map,
- distance restraints within the secondary structure elements.

These three forces are described in the following three sections.

2.3.1.1 Conformational Sampling Algorithm

The CONCOORD algorithm [24] is used as the basis for the conformational sampling algorithm. From the starting structure, it generates a large number of distance restraints which are represented as allowed distance ranges and all the produced structures have to obey these distance ranges. There are two groups of distance restraints included in CONCOORD:

- bonded restraints: bond lengths, angles and planarity which keep correct stereo-chemistry of the structure
- nonbonded restraints: van-der-Waals restraints which avoid overlapping and also define an upper limit for the allowed conformational change.

Bonded restraints are generated from the starting structure and are kept constant afterwards. Nonbonded restraints are updated at every structure generation cycle. The number of CONCOORD restraints is usually about ten times larger than the number of atoms.

Afterwards, the coordinates of the structure are randomly perturbed by applying a Gaussian distribution function with a width of 0.5 Å. Then atom pairs which

violate the corresponding distance interval will be moved along their inter-atomic vector to a distance that is randomly picked from the allowed interval.

Distance corrections are repeated until all distances fulfill the allowed interval or the maximum number of cycles is reached (typically 500). Once the structure does not converge within 500 iterations, a new round of correction will restart with the same CONCOORD restraints but different random perturbations. As the perturbation in the beginning is relatively small, a new structure can be generated within 5 to 50 correction cycles and the new structure will be used as the starting point to calculate new CONCOORD restraints.

2.3.1.2 Forces from Density Maps

In addition to the conformational sampling, to refine a structure into the density map, forces from the maps are required. The forces are derived from the comparison between the target density map, $\rho_{exp}(\vec{x})$, and a density map, $\rho_{model}(\vec{x})$, computed from the current coordinates of the model.

In the beginning of each structure generation cycle, a density map from the current model is calculated by Fourier transform. Then $\rho_{exp}(\vec{x})$ and $\rho_{model}(\vec{x})$ are normalized to have a mean value of zero and a standard deviation of one, which are $\widetilde{\rho}_{exp}(\vec{x})$ and $\widetilde{\rho}_{model}(\vec{x})$. During the refinement, the model map $\widetilde{\rho}_{model}(\vec{x})$ is becoming as similar to the experimental target density map as possible.

In principle, computing the gradient of the density map overlap directly and move the atoms accordingly would work, however experimental density maps contain a large amount of noise, DireX use stochastic gradient algorithm which is more robust against noise effect than the standard gradient approach.

Firstly, the density difference is calculated by

$$\widetilde{\rho}_{diff}(\vec{x}) = \widetilde{\rho}_{exp}(\vec{x}) - \widetilde{\rho}_{model}(\vec{x}). \tag{2.12}$$

Atoms tend to move into regions with high density difference, where the model does not produce enough density and out of regions with low density difference, where the model has too much density. The value of $\widetilde{\rho}_{diff}(\vec{x})$ calculation is performed before the coordinate perturbation and kept fixed during the correction cycles. In each correction cycle, atoms are moved by adding a vector:

$$\vec{g}_i = v(s_c)\frac{1}{12}\sum_{j=1}^{12}\rho_{diff}(\vec{r}_j)\frac{\vec{r}_j - \vec{x}_i}{|\vec{r}_j - \vec{x}_i|} \tag{2.13}$$

$\vec{r}_j$ gives a random position chosen from an isotropic Gaussian function with a width of 1 Å around $\vec{x}_i$. $v(s_c)$ is a scaling factor which depends on the correction-cycle step s_c. It decrease linearly from 1 to 0 within the first 40 steps, which converges the structure generation.

2.3.1.3 Secondary Structure Distance Restraints

Secondary structure information showing the three dimensional form of the protein. In the secondary structure, the most popular two types are α-helices and β-sheets which are defined by the hydrogen bonds showing highly regular patterns in the protein. These structural patterns are usually more stable and more conserved in the protein.

In considering this, restraints are applied over the secondary structure elements (SSE) during the refinement, which improves the refinement result via producing more regular secondary structure segments. A factor w_{sse} (a value between 0 and 1) is multiplied the secondary structure restraints which controls the strength of the restraints. Additionally, the secondary structure restraints are deformable by tuning a factor to change its deformability. During the DireX refinement, the secondary structural information was read in from a independent file which contains all the distance restraints as shown in Appendix E.

To generate the secondary structure restraints file, a secondary structure prediction software is needed to collect the secondary structure compositions from the primary protein sequence information. There are many of programs like PSIpred [78], SAM [79, 80], PORTER [81], PROF [82] and SABLE [83]. Here we use PSIpred for secondary structure prediction. Then the distances restraints for those SSEs are generated by using the standard secondary structure distance information as shown in Appendix A.

2.4 Assessment of path

To assess the quality of the trace, we used and compared several parameters to measure the agreement between the tour we got and the target structure.

2.4.1 Density Map Correlation

The first one is the map correlation between the simulated model density map and the target density map. This correlation is a rough estimation of the correlation as for the bead model we only simulated the density over the pseudo Cα atoms ,however the correlation was calculated against the full atom target map. Despite of that, the map correlation value provides a good measure of how well the model fits to the density map. The map correlation between two maps are calculated by:

$$c.c. = \frac{\sum \rho_1 \rho_2}{\sqrt{\sum \rho_1^2 \sum \rho_2^2}} \tag{2.14}$$

ρ_1 and ρ_2 represents the data sets of two density maps.

2.4.2 Root Mean Square Deviation:

The most widely used measure to compare two structures is the root mean square deviation (RMSD). The RMSD is a measure that calculates the positional difference of corresponding atoms after a proper superposition procedure which is done by a least squares quadratic (LSQ) fit. It is calculated as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\parallel X_i - X_j \parallel)^2} \tag{2.15}$$

in which N is the number of equivalent atoms and the X_i and X_j are the corresponding pairs of atoms.

This value quantifies how similar the model is compared with the target structure. The lower the value is the more similar the compared two sets of atoms are. This value gives a good impression of the quality of a model. Although the RMSD gives a good approximation of the quality of model, minor local drift would result in a large variation in the score.

2.4.3 Topology Score

The topology score is a parameter to measure the topological similarity between model and target. This term is based on the directionality of the aligned sequences. It is a score from CLICK [84]. Its value ranges between 0 and 1, where 0 means two structures are have no similarity topologically and 1 means that the structures are topologically identical.

Within CLICK, the alignment was done independently from the topology between the model and the target structure. The important point for the topology score is that conformational variations would not have a big impact on the topology score.

CHAPTER 3

Results

3.1 Placing Pseudo-Cα Atoms

The goal of this work is to determine the backbone trace of a protein using a low-resolution density map and the known amino acid sequence. For this, we first place generic beads (point masses) into the density map, where the number of beads is chosen equal to the number of amino acids. Each bead is assigned the mass of an average amino acid.

At the same time, a density map that approximates the mass distribution of the protein was generated from the beads. This simulated density map was used to calculated the map correlation during the pre-bead refinement. Fig. 3.5 depicts the pseudoatom representation in the calmodulin protein.

Our simulated Cα atoms were located according to two criteria: density and distance.

3.1.1 Choice of the Density Threshold

We characterize the surface as a set of grid points with density value ρ_1, so that the volume corresponding to the density values that are larger than ρ_1 is close to

the volume of the protein. The calculation of the volume of the protein was based on an empirical value 1.21 $Å^3$/Da [85], which represents the average volume that a general protein occupies.

As we know the protein sequence, we can calculate the mass of the protein. Then the whole volume of the protein can be calculated by this equation:

$$\text{Vol}_{protein} = \sum_{i=0}^{n} M_{aa}^{i} \times 1.21. \tag{3.1}$$

The M_{aa}^{i} represent the mass of a certain amino acid in the sequence. The whole density map values were sorted from high to low value. From this sorted density map, we can use indices to find out the corresponding threshold. The index corresponding to the expected threshold was calculated from

$$\text{Index}_{threshold_1} = \text{Vol}_{protein}/\text{Vol}_{voxel}, \tag{3.2}$$

in which $\text{Vol}_{voxel} = \text{voxelsize}^3$. With $\text{Index}_{threshold_1}$, we can find out the corresponding density value $\rho_{threshold_1}$ from the sorted density histogram, as shown in schematic Fig. 3.1.

The isosurface ($\rho_{threshold_1}$) which correspond to the index ($Index_{threshold_1}$) identifies the volume of the full atomic structure and it is the lower boundary of the shape of the protein volume. However, regarding the structure of the backbone, there is no side chain information included as shown in Fig. 1.1. The backbone volume can be approximated by the mass of the backbone. After the amino acids forming the peptide chain, each amino acid can be represented by a simplified and repetitive molecular formula unit NC_2OH_2 which is shown in the blue box in Fig. 3.2.

The volume of the backbone can be calculated from

$$\text{Vol}_{backbone} = n * M_{NC_2OH_2} \times 1.21. \tag{3.3}$$

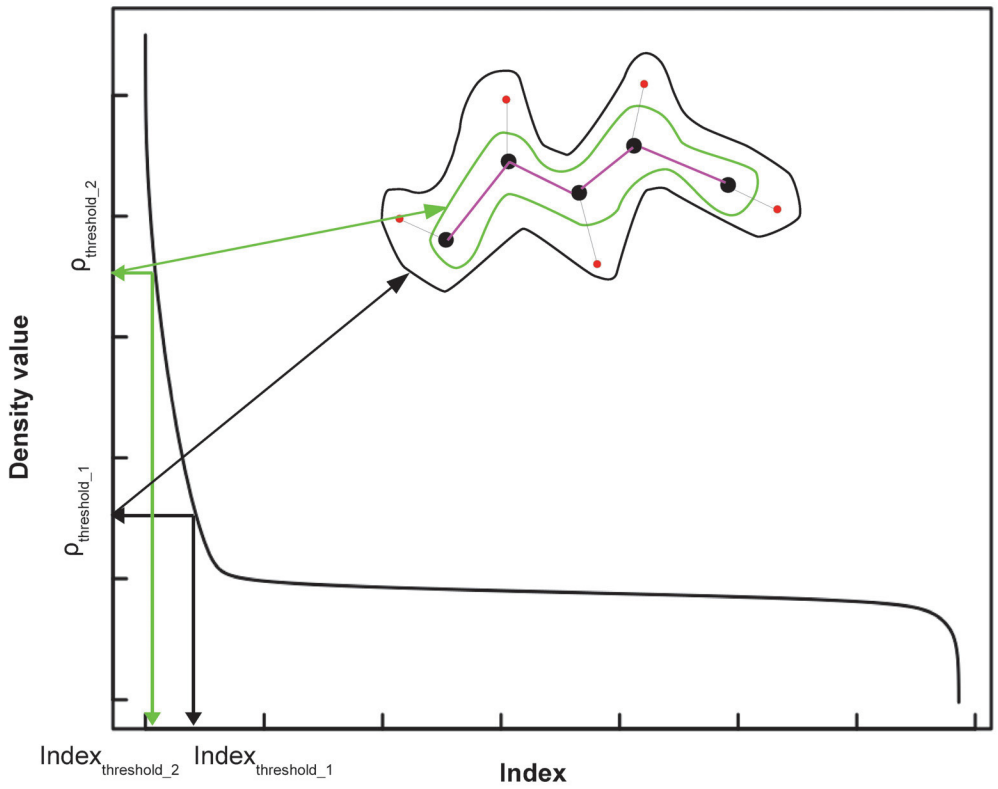

Figure 3.1: Density threshold and its index. The black curve is the schematic sorted density plot against indices. The green color highlights the backbone density, $Index_{threshold_2}$ and $\rho_{threshold_2}$. The black highlights the full atomic density, $Index_{threshold_1}$ and $\rho_{threshold_1}$

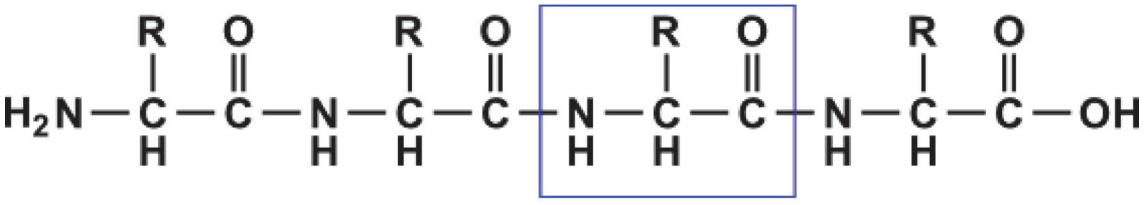

Figure 3.2: Protein Backbone. The blue box highlights the repeating unit of the peptide chain

in which n represents the number of amino acids in the sequence and the $M_{NC_2OH_2}$ gives the molecular weight of the repeating unit -NC2OH2- in the peptide chain.

The index that corresponds to the backbone threshold can be calculated with $\text{Index}_{threshold_2} = \text{Vol}_{backbone}/\text{Vol}_{voxel}$. From Fig. 3.1, the $\rho_{threshold_2}$ can be found with the $Index_{threshold_2}$ value. This $\rho_{threshold_2}$ yields an upper bound of the backbone volume. The value between $\rho_{threshold_1}$ and $\rho_{threshold_2}$ became the threshold interval. Theoretically, all the values inside this interval can be used as the threshold for locating the beads in the density. Although, the upper bound threshold encloses the whole backbone structure, the beads should not stay on the surface. To make the bead model more compact, the threshold was chosen as a value which has 10% more density volume than the upper bound $\rho_{threshold_2}$.

When locating the bead in the density, this calculated threshold is used as the outer border of the map, which means we do not choose any point that has a value smaller than this threshold as a potential bead position. So that we keep our beads in or at the threshold which is assumed to be the surface of the backbone density.

3.1.2 Distance

In addition to considering the threshold, we also use distance as another criterion to place the beads. After we determined the threshold, the density grid points with a density value larger than the threshold forms the protein space. This space defines the region that our beads will occupy. There is an empirical observation from hundreds of protein structures showing that the distances between Cα and Cα atoms are ranging from 3.5–4.2 Å. Based on these values, we randomly place beads into the density map while making sure that there is not any pair distance closer than 3.5 Å. And then a rough checking procedure was performed, which finds and deletes those beads that have no neighbours within a radius of 4.2 Å. Those beads will be complemented by new beads, for which except using the threshold and the distance criteria, an additional map correlation procedure (Section 3.1.3)

was also considered, which results in a bead model that has bond lengths of at least 3.5 Å, at the same time for each bead there will be at least another bead within a distance of 4.2 Å.

3.1.3 Map Correlation Refinement

While placing the beads into the density, we calculate simulated CryoEM maps by putting a three-dimensional Gaussian function centered at each bead position and integrating all these functions for each grid point at the same time to get the bead density function [35, 36]. This density function was calculated according to this equation:

$$\rho^{sim}(i,j,k) = \frac{1}{(\sqrt{2\pi/3}\sigma u)^3} \sum_{n=1}^{N} \int_{V_{ijk}} \exp\left\{ -\frac{3}{2\sigma^2}\Big[(x-x_n)^2 + (y-y_n)^2 + (z-z_n)^2\Big]\right\} dxdydz \tag{3.4}$$

where $\rho(i,j,k)$ is the simulated density of voxel (i,j,k), σ is one-half the map resolution [86], u is the grid spacing(the edge length of the cubic voxel), N is the total number of beads, and $V_{ijk} = u^3$ stands for the volume of a unit voxel,$(x_n, y_n, z_n) = (n,n,n) \times u$ are the Cartesian coordinates of atom n.

The resolution of this simulated density map does not exactly correspond to the resolution of the CryoEM experiments. A map filtering step as described in Section 2.1.2 was applied to this initially simulated density map, after which we get the expected resolution. Previous studies [35, 36] with coarse-grained models have shown that the details of how the maps were simulated do not affect the performance.

With the target density map $\rho^{exp}(i,j,k)$ and the filtered map $\rho^{sim}(i,j,k)$, the correlation coefficient between two maps is defined as

$$c.c. = \frac{\sum_{ijk} \rho^{exp}(i,j,k)\rho^{sim}(i,j,k)}{\sqrt{\sum_{ijk} \rho^{exp}(i,j,k)^2 \sum_{ijk} \rho^{sim}(i,j,k)^2}} \tag{3.5}$$

A pre-refinement procedure for the bead model was shown in Fig. 3.3. There were some beads removed from the original bead model because of violation of the distance criterion. While adding the new beads to complete the model, apart from using the threshold and distance rule, the correlation between the simulated map of the bead model and the target map is also considered. The new bead position has to satisfy the threshold and distance criteria, at the same time it also has to fulfill that the new model with the added beads has a better map correlation than the initial model otherwise the new beads would be relocated into the initial model. In the end, only the bead model that has better map correlation is accepted. This model will be used for the following steps.

3.1.4 Calmodulin Test Result

Threshold

The threshold here is used to define the border of the bead model and also the surface of the protein. As there are multiple ways to define these volumes which correspond to different constants [85, 87–91], here the one 1.21 $Å^3/Da$ [85] is one of the most classic and frequently used value. It was calculated based on the volume that residues occupy in folded proteins.

For calmodulin, we have the sequence information (Appendix C) and the corresponding weight can be calculated according to the element weight. In this 143 amino acids sequence, the overall weight is 16.15 kDa so that the volume of the atomic structure is 19541.6 $Å^3$. With this volume, from our simulated density map in which each cubic space has a volume of 1 $Å^3$, the corresponding $\text{Index}_{threshold_1} = 19542$ can be calculated. By using this index, the $\rho_{threshold_2}$ can be identified in the sorted density plot, that is 1.295. This value pair index and

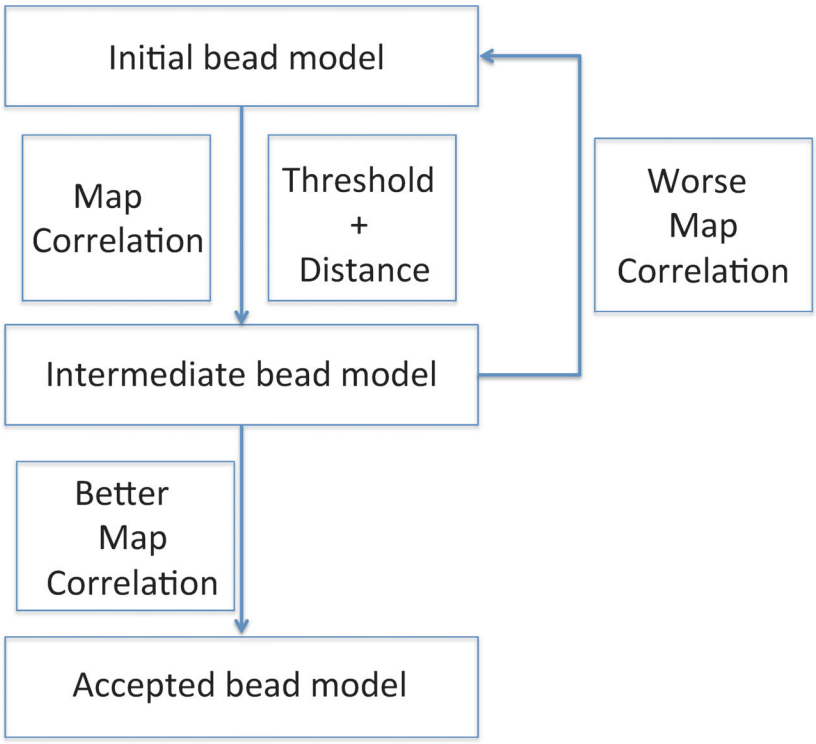

Figure 3.3: Pre-refinement of bead model. After random bead deleted, new bead will be added. If the new map correlation is better than the initial one, the model will be accepted, otherwise trying to start from the model after delete and find another new bead.

density was plotted as a green spot in Fig. 3.4. The density map and the full atomic model fit well.

$\rho_{threshold_2}$ is the lower boundary of the threshold interval. For the upper boundary, the backbone weight is used. Calmodulin has 143 amino acids so that the weight is calculated by $143 \times M_{C_2NOH_2}$ which equals to 8.028 kDa. Its volume is about 9714 $Å^3$ and the corresponding threshold is 3.3, this pair is shown as black dot in Fig. 3.4.

The threshold interval was identified by these two values, the final threshold that we use was 3.6 which is slightly (about 10%) higher than the upper boundary. The reason for doing this is that the upper boundary is close to the backbone trace and

to make the beads positions close to the backbone structure, a higher threshold than the upper boundary can be more efficient.

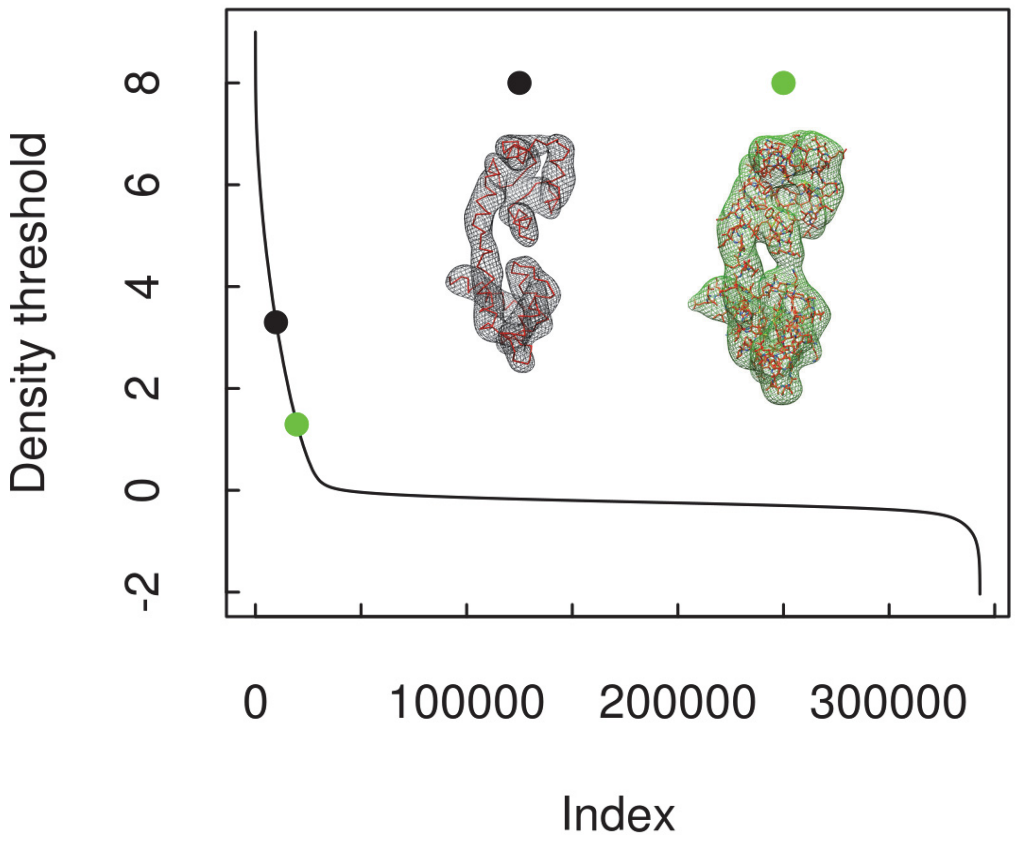

Figure 3.4: Calmodulin density threshold. The green dot represents the $\rho_{threshold_2}$ and $Index_{threshold_2}$ pair and the black dot is the $\rho_{threshold_1}$ and $Index_{threshold_1}$ pair. The corresponding density maps are depicted below the legend dot.

Distance

The empirical average distance between Cα atoms is around 3.78 Å which is a statistic result from several protein structures from the Protein Data Bank. However, if this value is strongly enforced while putting the beads, the model in the end will be too compact. To allow for more freedom, we used a distance interval between 3.5 to 4.2 Å, which make the bead have more spaces to fit into the map especially for low resolution data and it corresponding well to our energy function in the later steps.

The distance interval has a similar effect as choosing a slightly higher density threshold. There is a potential relation between the threshold and the average distance between beads. While choose larger threshold which means smaller density map scale, the smaller distance we have to use to pack the right number of beads inside the density and vice verse.

Map correlation

Initially, the density map simulated from the bead model was calculated by Eq. (3.4) at high resolution so that it can be further filtered to a expected lower resolution. Here we always generated the map with $\sigma = 0.5$, which means the simulated map has a resolution of 1.45 Å (according to the Rayleigh criterion). Then the map was filtered to 8 Å, the same as our target map.

After locating the beads with the threshold, the initial bead model has a map correlation of 0.9593, as considering the distance criterion, there were 41 beads deleted. While adding new beads, the final better bead model has a correlation of 0.978. The final bead model and the density map are shown in Fig. 3.5.

Although the bead model we get is not exactly the true backbone structure, the beads well represent the shape of the density map. Before tracing the backbone, we try to keep the topology of the beads as close as possible to the target map.The bead model will be further refined in the later steps.

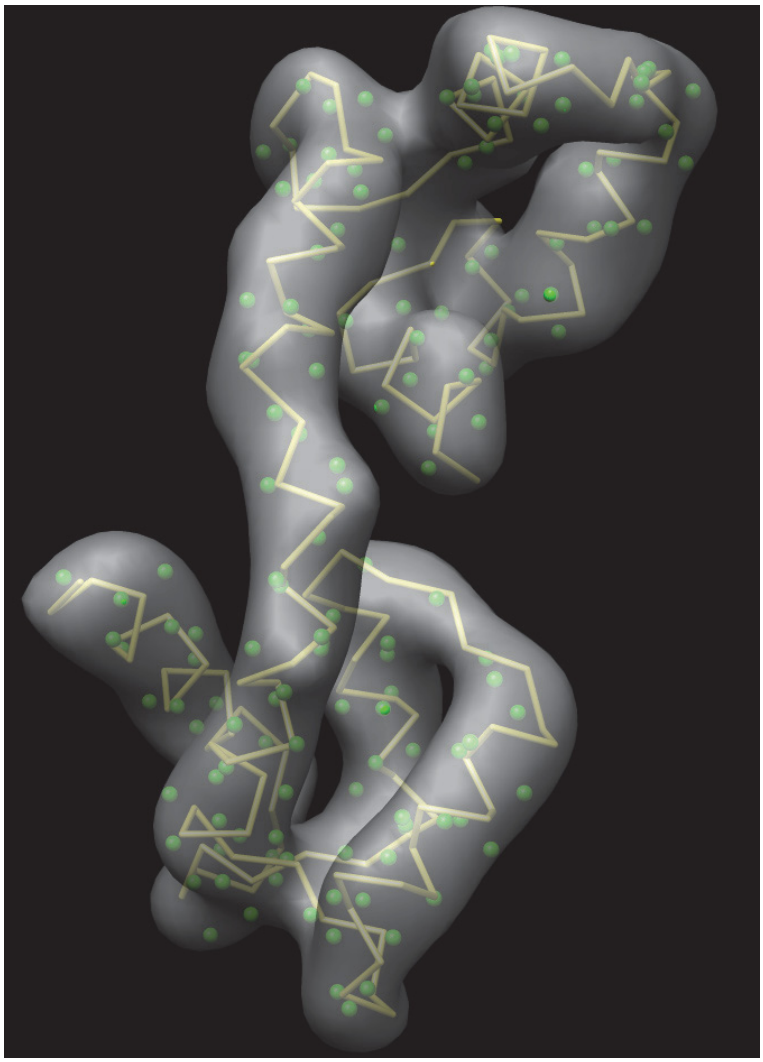

Figure 3.5: Bead model in density map. The bead model (green) is enclosed by the density map at 3.3 density threshold. The target backbone is shown as a yellow trace inside the map.

3.2 The Lin-Kernighan Heuristic using Pseudo-Cα Atoms

3.2.1 Monte-Carlo for Protein Backbone Tracing

In our case, to trace the backbone, for the beads we have already located, there are a great amount of configurations we can search in the space. Among all these different configurations, a lot of local minima exist. To identify the optimum one

is getting more difficult especially as the number of beads is increasing. To help the optimization procedure, we decided to use the SA method described above as the main optimization frame. At the same time, it combines with LK algorithm making two heuristic together, resulting in a more flexible and powerful method.

As we use the SA method for our backbone tracing, our bead model system has to fit into the SA four necessary requirements shown previously. Our tracing problem fit as the following way:

- **Initialization:** All the beads are randomly ordered and connected. Each bead is connected with other two beads that are arranged one as the previous and the other as the next. The starting tour becomes a one direction oriented tour.
- **Modification:** Here we do the changes by $2-opt$, $3-opt$ and the $k-opt$ moves which have been introduced in Section 2.2.2.
- **Function:** The function here which we aim to minimize is a hybrid energy that combines with several different energy terms. It can be written as:

$$E_{total} = \alpha E_{lk} + \beta E_{sse} + \gamma E_{mj} + \delta E_{dens}, \tag{3.6}$$

  in which the E_{lk} is a term used to describe the energy change of the LK procedure, E_{sse} is a term for the secondary structure energy change, E_{mj} is a term representing the residue-residue interaction energy called Miyazawa-Jernigan energy as described below and the E_{dens} is a term describing the electron density energy change during the optimization. The parameters combined with these energy are the weighting factors which balance the information between each other. All these energy terms are further introduced in the next section.
- **Cooling:** Our annealing strategy needs some tests before execution. We randomly generate some tours which give an estimate of the range of values ΔE. From there, we choose a T value significantly larger than the largest

ΔE. The temperature will be decreased by 1 or 10 percent from the previous step depending on different situations. If the energy landscape is relatively smooth, then we can drop the temperature quickly which also makes the structure converge fast. Otherwise, we have to slowly lower the temperature to let them have time and higher chance to jump out of the local minima.

3.2.2 Energy Function

As we have described in Section 2.2, the original LK algorithm is trying to search the shortest path among all the cities (Cα-positions), then we can easily figure out that the function need to be minimized is just the total length of a journey. However, coming back to the backbone tracing problem, we not only need to search through a large amount of tours to get the correct one which might not be the shortest of all, but also have to take the protein structure information into account so that we finally find out the optimum result of backbone but not only the shortest path. We use those protein structure information to balance the searching procedure. The purpose of designing this energy function is to quantitatively show the optimization result. At the same time, each step of the modification will be directly illustrated by the energy in that state.

The entire energy function can be split into three parts:

$$E_{total} = \underbrace{\alpha E_{lk} + \beta E_{sse}}_{\text{structure based}} + \overbrace{\gamma E_{mj}}^{\text{sequence based}} + \underbrace{\delta E_{dens}}_{\text{density based}} .$$

The corresponding functionality is shown with the formula.

- **Structure Based:**

  That means all terms in this group are structure related. E_{lk} focuses on the local structural information and E_{sse} emphasize more on the protein secondary structural information.

***E_{lk}*:** This is a term representing the energy of the LK procedure and it is written as:

$$E_{lk} = \sum_{i=1}^{n-1} (d_{i,i+1} - d_{average})^2 \tag{3.7}$$

in which i is a index for the beads that currently taken into account, $d_{i,i+1}$ is the Euclidean distance between two consecutive beads and the $d_{average}$ is the average distance between the beads in the current path. It can be calculated by

$$d_{average} = \frac{\sum_{i=1}^{n-1} d_{i,i+1}}{n-1}. \tag{3.8}$$

The LK energy is a harmonic potential which can be viewed as a distance restraint between two connected beads. It is the most important energy term among all the four and the amount of information E_{lk} contributes to E_{total} is weighted by the parameter α as shown in Eq. (3.6). When suppressing all the other energy, the meaning of minimizing the total energy is "shorten" this harmonic potential for the tour and aiming to identify the best structure by finding out the tour with the lowest potential energy. As the optimization is running, the $d_{average}$ is decreasing and the E_{lk} is also converging. This energy change is attached with the LK procedure that is either $2-opt$ or $3-opt$ moves as we described before.

***E_{sse}*:** This group describes the secondary structure energy of the tour. The energy is calculated as

$$E_{sse} = \sum_{k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} (d_{i,j} - d_{ij}^{seq})^2 \tag{3.9}$$

in which $d_{i,j}$ is the distance between two beads within the same secondary structure element and d_{ij}^{seq} is the reference distance of the corresponding Cα atoms in the secondary structure. For each secondary structure in the sequence (sum over k), the distance between each pair of atoms ($d_{i,j}$) is compared to its reference distance d_{ij}^{seq}. The whole energy term is a summation of two parts one is the helix energy E_{helix} and the other is the β-sheet energy E_{beta}. The amount that E_{sse} contributes to the E_{total} is controlled by the

weighting factor β as shown in Eq. (3.6). Those two parts in E_{sse} are calculated in the harmonic potential way by using the reference distance. This reference distance was introduced by Frederic M.Richards [92]. It was used as a template to compare with a particular secondary structure type of a model structure to assign the right secondary structure in the model. The concrete distances that were used in this work are listed in Appendix A. The table contains standard Cα-Cα distances information of most secondary structure types like α-helix, β-sheet and it also has Cα-Cα distances information of neighbouring β-sheet.

Similar to the LK energy, E_{sse} is also a harmonic potential, which restraint the secondary structure elements but not the neighbouring two beads. The optimization procedure will try to converge all the connections within a local region which is supposed to be the corresponding secondary structure region.

- **Sequence Based:**

  There is only one term in this group called E_{mj}. The sequence based energy is meant to include sequence information. The primary sequence information is usually known when studying a particular protein. At the same time, the sequence information can be combined with the structure of the bead model to understand the contact potentials.

  ***E_{mj}*:** This energy is a term describing the information about the preference of residue pairs to be within a certain range. The letter "mj" represents the inter-residue contact potential calculation method which was introduced by Miyazawa and Jernigan [93, 94]. They developed a method called quasi-chemical approximation to estimate the inter-residue contact energies from a large amount of observed residue-residue contacts (within 6.5 Å) that exist in crystal structures. This empirical energy function also takes the solvent effects implicitly into account. Our sequence based energy can make use of this empirical energy and is calculated by:

$$E_{mj} = \sum_{i<j} M(a_i, a_j) D_{ij} \tag{3.10}$$

in which $M(a_i, a_j)$ is the Miyazawa-Jernigan (MJ) contact energy [95] between a_i and a_j and the energy information is shown in the Table 3.1. This statistical potential tells how much one amino acid favors the contact with the other one. D_{ij} is the contact matrix element. It is often used when describing proteins at a coarse-grained residue level to evaluate the total conformational energy. The contact was defined according to the Miyazawa-Jernigan potential: if two amino acids are within 6.5 Å, D_{ij} is 1, otherwise D_{ij} is 0. Here in our model, the pseudo Cα-Cα distance is considered as the corresponding amino acid distance. The contribution of the MJ contact energy to the total energy is balanced by the parameter γ as shown in Eq. (3.6).

Although the E_{mj} was calculated on the base of the sequence information, it also has to use the structural frame of the beads. The energy term E_{mj} restraints the connections between beads within a certain range by using the MJ contact energy. Each time we optimize the connection, we need to reassign the sequence twice, once from the starting bead to the end and the other one assign in the opposite direction. The one that has lower energy potential will be considered for further optimization during the process.

- **Density Based:**

  We have included the structure-based and sequence-based terms for the whole energy calculation. Those are using information either from the protein sequence or from the bead structure. None of them is using the density map information, which provides a lot of useful structural information about the protein. We therefore devised the last energy term

  $\boldsymbol{E_{dens}}$**:** The E_{dens} describes the interaction between any pairs of connected beads, which is defined by the density map information. It is given by the following formula:

	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO
PRO	-0.18	-0.13	-0.19	-0.05	-0.12	-0.05	-0.37	-0.25	0.15	0.02	0.13	0.20	0.17	0.18	0.37	0.33	0.01	0.17	0.47	0.11
LYS	0.33	0.29	0.19	0.24	0.22	0.29	0.09	-0.05	0.41	0.29	0.33	0.36	0.28	0.22	-0.06	-0.01	0.38	0.66	0.76	
ARG	0.08	0.03	-0.05	0.00	-0.04	0.08	-0.21	-0.25	0.24	0.09	0.11	0.16	0.09	0.10	-0.22	-0.24	0.05	0.19		
HIS	-0.36	-0.29	-0.34	-0.13	-0.18	-0.06	-0.37	-0.30	0.07	0.00	-0.03	0.04	0.15	0.00	0.00	-0.10	-0.40			
ASP	0.12	0.30	0.18	0.22	0.27	0.36	0.07	-0.07	0.27	0.11	0.11	0.10	0.24	0.02	0.44	0.29				
GLU	0.20	0.12	0.14	0.17	0.17	0.26	-0.00	-0.08	0.38	0.32	0.16	0.18	0.27	0.12	0.46					
ASN	-0.01	0.04	-0.01	0.14	0.04	0.12	-0.10	-0.11	0.15	-0.01	0.04	0.09	0.06	-0.06						
GLN	-0.07	-0.06	-0.11	-0.01	-0.04	0.08	-0.02	-0.14	0.22	0.13	0.12	0.22	0.20							
SER	-0.13	0.05	-0.12	0.03	-0.02	0.04	-0.01	-0.08	0.10	-0.01	0.04	0.05								
THR	-0.15	-0.11	-0.15	-0.15	-0.15	-0.07	-0.02	-0.09	0.04	-0.04	0.03									
GLY	-0.31	-0.17	-0.19	-0.13	-0.16	-0.15	-0.25	-0.22	-0.08	-0.29										
ALA	-0.33	-0.27	-0.36	-0.37	-0.38	-0.32	-0.27	-0.20	-0.12											
TYR	-0.39	-0.56	-0.58	-0.49	-0.55	-0.38	-0.49	-0.45												
TRP	-0.66	-0.73	-0.68	-0.60	-0.62	-0.51	-0.64													
VAL	-0.59	-0.51	-0.67	-0.67	-0.74	-0.65														
LEU	-0.65	-0.70	-0.80	-0.81	-0.84															
ILE	-0.64	-0.66	-0.73	-0.74																
PHE	-0.67	-0.83	-0.88																	
MET	-0.61	-0.70																		
CYS	-1.19																			

Table 3.1: Contact energies derived from protein crystal structures. The smaller the score, the more frequent is the contact observed in the Protein Data Bank.

$$E_{inter} = \begin{cases} 1 & \rho_{inter} > \rho_{cutoff} \\ 0 & \rho_{inter} <= \rho_{cutoff} \end{cases} \tag{3.11}$$

$$E_{dens} = 1 - \frac{\sum_{i=1}^{n-1} E_{inter}(i)}{n-1} \tag{3.12}$$

in Eq. (3.13) the ρ_{inter} is the density interpolated along the connection between two beads. The value was interpolated by the method of trilinear interpolation which is shown in the Appendix. B. E_{inter} is the density energy term assigned to that corresponding pair of beads. If the interpolated density value is greater than the ρ_{cutoff}, then the density energy is set to 1, otherwise it is equal to 0. The ρ_{cutoff} is set as the same value as the threshold when locating the beads. The energy here only represents the binary energy of a certain point which either informative or none. The virtual energy E_{dens} that used to describing the edge is calculated by Eq. (3.14) which gives the average value of the interpolated points subtracted by 1. The amount of information it provides to the E_{total} is weighted by the factor δ as shown in Eq. (3.6).

The way to set up the density energy is shown in Fig. 3.6. In both of A and B, the red circles represent the beads that we placed into the density (depicted by the curved blue lines), so when the density values are always above the cutoff, the E_{inter} will be 1. The blue and green circles which evenly distant (1Å) and distributed between them are the interpolated points along the vector of these two beads. In panel A, it is shown that when the interpolated points are staying outside the density, the $E_{dens} = \frac{5}{7}$ and panel B shows when all beads stay in the density map then $E_{dens} = 0$. The more interpolated points are outside the density map, the higher the density energy will be. This energy value is always in the interval $[0, 1)$.

The density energy restraint are used to restraint the connections with higher

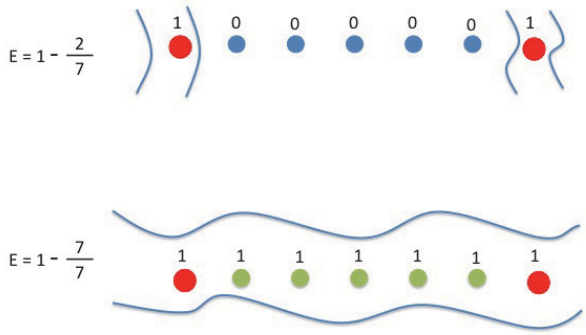

Figure 3.6: Density restraints. The red circles represents the beads located in the density (blue curves) and the blue and green beads represents the interpolated points along the corresponding vector between the red circles. The number on top of the beads are the E_{inter} values.

density energy. Fig. 3.7 exemplifies the effect of applying the density restraints. State A has a higher density potential, with applying the density restraints during the optimization, it tends to choose those connections with lower density potential which fits better into the density map.

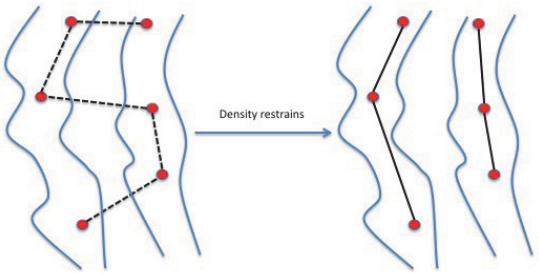

Figure 3.7: Effect of Density restraints. The red circles represent the beads. Dotted lines are the connections that cross the density map and the solid lines are the connections that stay inside the density map.

3.2.3 Calmodulin Test Result

- **LKH Test with Correct Cα-Positions**

  2-opt Move Only

  To test the ability of the LK algorithm to generate the correct path with a collection of pseudoatoms, we take the Cα atoms from the calmodulin crystal structure and use this correct Cα atoms as a first test model. As all the Cα-atoms stay in the right positions, in the energy function Eq. (3.6), we only use the LK term and set all the other terms to zero. That means only the LK algorithm has an effect during the optimization procedure. To test the LK algorithm, $2-opt$ move was used as the fundamental step, we try to use $2-opt$ move alone to solve the trace.

  With the derived perfect Cα positions, we make 200 independent optimization runs. For each single test, 5 million iterations of optimization was implemented and all starting structures had the same initial connections. From all the 200 tests, we chose the structure with the lowest energy. The energy of these assemble of structures range from 0.000414 to 1.947 with a mean value of 0.897 and variance at 0.165 and the RMSD value between 0 and 16.846 with mean value at 11.213 and variance at 10.29.

  According to the 200 tests, all the tests can find out the correct trace, however in each test the hit rate is about only 1.5% that means about 3 of 200 can reach the target energy which calculated from the perfect backbone trace. Comparing with the target structure, the topology score is 1 and RMSD value is 0 that means these traces are identical to the target. On average, it took 2.097s for one test run and the typical energy and RMSD change was shown in Fig. 3.8. The plot shows the optimization processes in which correct trace was figured out. The energy decreases gradually, however, the RMSD value has larger fluctuations comparing with the energy change. 2-opt move can optimize fast and the energy converges also quickly.

Although the 2-opt alone can find the optimum structure in the end, the efficiency is still low even for the perfect Cα positions. As the starting models have the same connections for all the 40000 runs, there is not adequate randomness and the sampling result is not sufficiently general.

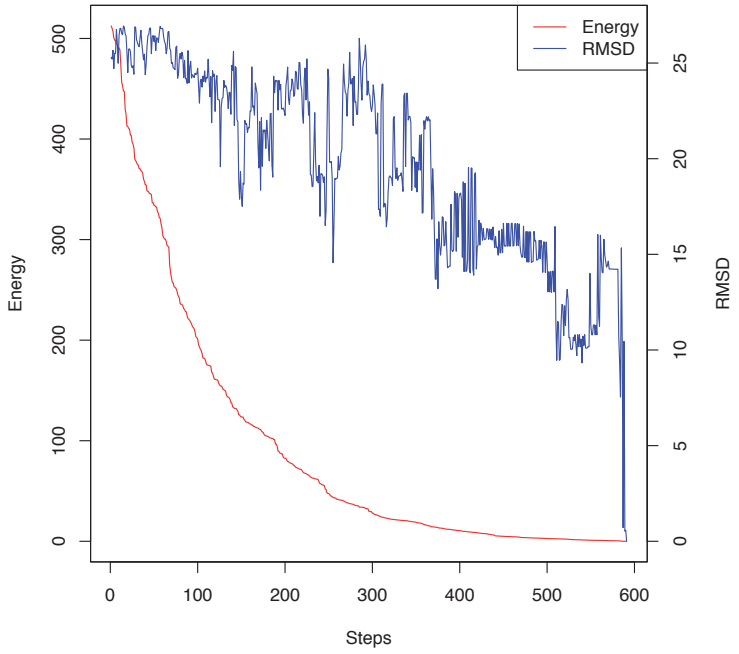

Figure 3.8: Energy and RMSD change during the path optimization. The left Y axis shows the energy value (red) and the right Y axis gives the RMSD value.

3.2.3.1 Further Optimization

The 2-opt move itself can find the correct backbone trace in a short time. However limitations still exist. In the 2-opt move, even though everything was completed based on the perfect Cα positions, the successful rate to get the right trace is still low and the majority of solutions are distant from the correct structure. Additionally, each single test run starts with the same connection which also limited the sampling space.

All these limitations because the structures got trapped in local minima which are difficult to jump out of to reach the global minimum. It is schematically shown in Fig. 3.9, the configuration A was stuck in the local minima state which is still far away from the global optimum configuration B. To make the entire optimization processes work more efficient, we add a few more points from which we could get some improvements to our algorithm.

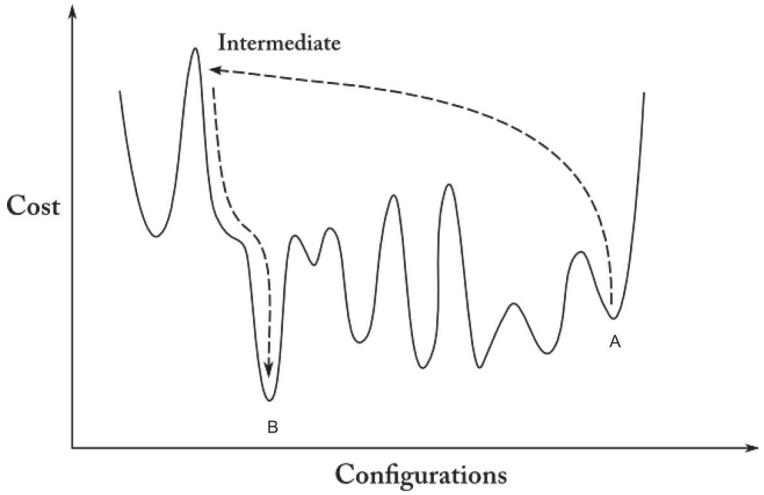

Figure 3.9: Schematic energy landscape. A is the staring energy state which jumping to the intermediate state and then converge to the global minimum state B.

Random restart

2-opt move alone can converge fast, however in the 200 structures, most of them (98%) drop directly to a local minimum and get stuck there. To improve the probability of reaching the global minimum, we add some randomness to the starting model.

The most direct way to improve the optimization procedure is to repeat the processes with a new random start. In this case, all the local minima results are independent and this random start makes it possible to reach a lower energy state of all the local minima distributions. The random restart methods works well for searching the optimum result, however as the size of the system is increasing there is also a large number of local minima, it

becomes more difficult to find the optimum as the probability of finding it by this random sampling is getting lower and lower [96]. Regarding the size of proteins, most of the protein sequences has a length less than 300 amino acids, so random restart is worthful to be tried out.

For each single bead, each test run was started with random connections. With the random restart method, each search starts from a different energy state so that test runs are independent from each other and multiple trials would increase the chance to reach the lower energy state.

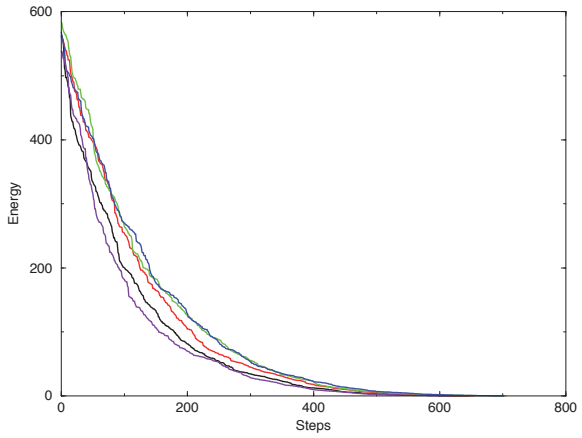

Figure 3.10: Energy convergence by random restart. Energy convergence of five typical successful runs with random restart. Each different color of the curve correspond to different starting energy state.

In the next step, the 2-opt move was combined with the random restart The same tests as for 2-opt alone were performed. With random restart, each single test took 2.456s which is slightly longer than 2-opt method. However, among all the 200 runs result, the successful hit rate increased to 4.5% which is about 9 correct structures. The energy distribution of these two tests are shown in Fig. 3.11. The energy distribution from the method with random restart is slightly shifted towards lower values. It shows that more structures converged to a lower energy states than the structures from 2-opt alone. The

mean value shifted from 0.897 to 0.792 and the variance increased from 0.165 to 0.176 which shows the randomness.

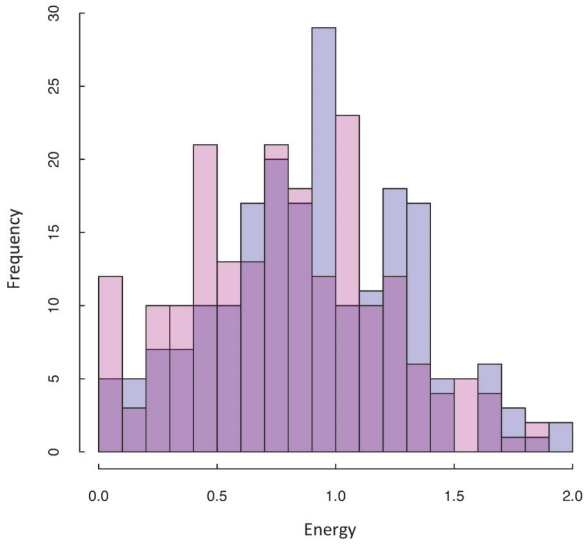

Figure 3.11: Histogram of energy distribution. Blue bars shows the energy distribution of all models without random restart; pink colored the same after adding random restart feature and the overlapped region shown as purple.

Perturbation of path

After combining with random restart, the algorithm works slightly better than before, however the low efficiency problem still exists because of the local minima trap.

Except doing a random restart, we can also try to make big jumps which make it out of the local minima like is shown in Fig. 3.9. The local minima was taken out from the local minima to a higher energy state C and from there the optimization procedure can proceed to reach the global minimum.

The method here we use called perturbation which means a modification was applied on the current local minima configuration so that it can jump out of that region and move to an active state for further optimization. The strength of the perturbation can be defined as the significance level of modification. In our tracing algorithm, that can be the number of exchanges.

To give a proper perturbation for our tracing algorithm, we have to satisfy one basic condition that the LK optimization procedure should not be able

to undo the perturbation. If it does, then the procedure will return back to the local minimum state from where it just came. The easiest way would be making some exchanges to perturb. However, if the perturbation is too strong which means a large number of exchanges, it will behave similar as the random restart in which better solution may be found with low probability. If the perturbation strength is too small like the basic $2-opt$ move, then it will easily move back to the local minima which will limit the search efficiency for the global minima structure.

As the optimization procedure always consists of structured sequential changes, the perturbation should better not be applied in a sequential way, because the aim is to transform a local optimum into a good start for further optimization. The original LK paper [65] introduced the first non-sequential exchange which is a $4-opt$ move as shown in Fig. 3.12, the left figure depicts the $4-opt$ move in the real way and the right one shows it in the circle way. This perturbation procedure can be seen as two $2-opt$ moves, however different from the $2-opt$ move, the first step breaks the connection and the second step reconnects from other location to make a full path. The perturbation is generated randomly.

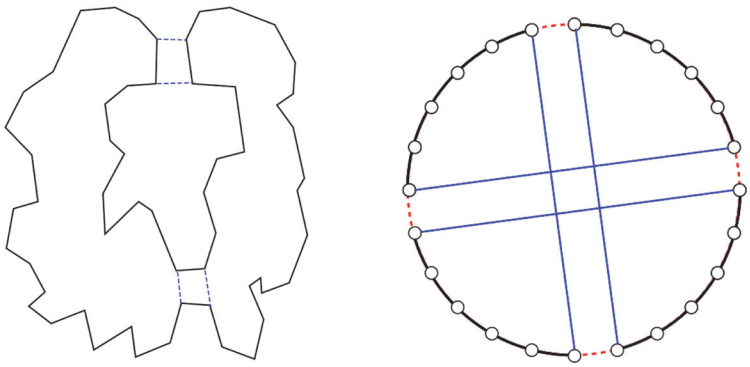

Figure 3.12: $4-opt$ **perturbation.** The dotted lines were deleted and replaced by the solid blue lines. The right panel shows the circle style of the left path and dotted red line are deleted and replaced by the solid blue lines.

This perturbation can not be undone by the $2-opt$, $3-opt$ or LK and it can change the topology dramatically by the two set distant exchanges. Additionally, this change does not increase the length of the tour that much so it still in a good state.

To test the method combined with perturbation, we also made the same runs as before. Differently from before, as the structure reaches a minimum after a given number of iterations, non-global minima structures will be perturbed and a further optimization is done without any random restart. If random restart would be reused here, then the perturbation effect would be killed.

After adding the perturbation method, each single run took about 3.135s which is a bit longer than before. However the success rate increased dramatically from 4.5% to 24.5%, the mean value of the energy distribution was 0.359 with a variance of 0.082. In Fig. 3.13, the plot shows perturbation effect on the perfect Cα positions. Most of the energy values shifted to the left which means a number of higher energy models were optimized to a lower energy states. With the perturbation, the energy would jump up to a higher energy level and was further optimized from there. Panel B exemplifies the energy change of one successful run with perturbation. After perturbation, the energy quickly dropped and converged to the global minimum state.

MC-SA

Perturbation plays a positive role during the optimization process, however, a success rate of 24.5% is low still. The reason for this is that the configuration space of the tours is large and an exhaustive search would result in a low probability to hit the correct one.

The aforementioned MC-SA (Section 2.2.3) is a method used for the global optimization problem. It is trying to make a good approximation to the global optimum while searching in a large space. It can find an optimum or near optimum solution within a reasonable time range.

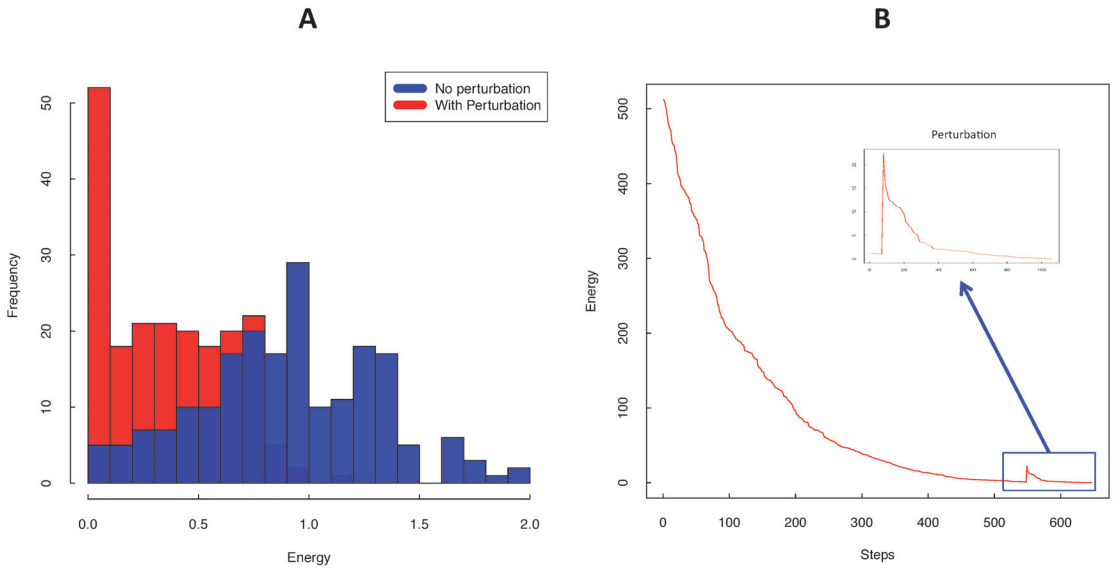

Figure 3.13: Perturbation effect on perfect Cα positions. Panel A shows the energy distribution of optimized models from the method with and without perturbation. Panel B exemplifies the energy converging process with the inset plot showing the energy change after perturbation.

With the same test as before, among the 200 output paths, the starting model of each optimization was randomly connected as in random restart. Fig. 3.14 A shows one of the typical random restart models. In the end, most of the structures (89.5%) converged to the global minimum structure which is also the correct Cα trace with energy 0.000414 and its corresponding trace is also shown in B of Fig. 3.14 which exactly matched the correct Cα trace. The energy change during the optimization processes is similar as the previous one, is shown in Fig. 3.11. The RMSD value between the final model and our target is 0 and the topology score from CLICK is 1, which means they are identical. The time it costs for each single run is approximately 2.771s nearly the same time as before but the efficiency is greatly improved.

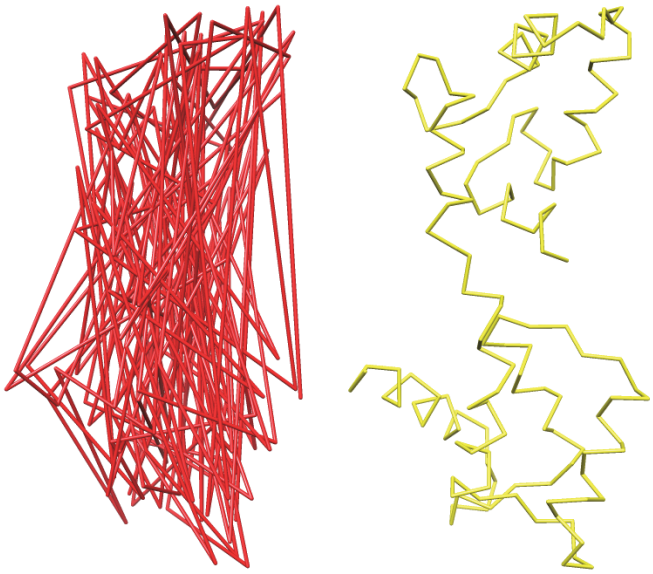

Figure 3.14: Tracing with perfect Cα positions. The left panel shows the starting model with random connections (in red). The optimized model (yellow) is shown on the right panel.

- **LKH Test with Bead Model**

  The LKH algorithm works well with the perfect Cα positions. However, in the real case, the beads can not be placed exactly at the right Cα positions especially for the low resolution data, so we have to test the algorithm with our unprecise bead model. With the aforementioned bead placing method, we also generate 200 bead models and for each of them the same LKH test as for perfect Cα model was performed. For each model we make 200 runs, among all those structures, we pick out the one with the lowest energy. The value of E_{total} was ranging from 0.321 to 0.666, with a mean value of 0.398 and variance of 0.004. The corresponding RMSD values calculated between the lowest energy model and our target model was ranging from 5.6 to 16.47 Å, with a mean value at 9.26 Å and standard deviation at 6.58 Å^2. In Fig. 3.15, the pair value of RMSD and total energy of one protein model was plotted. There is a correlation between the RMSD value and the total energy in which here is only the LK energy and its value is 0.428. The model which has the lowest RMSD value 5.6 corresponding to the total energy 0.399 which is close to the lowest energy.

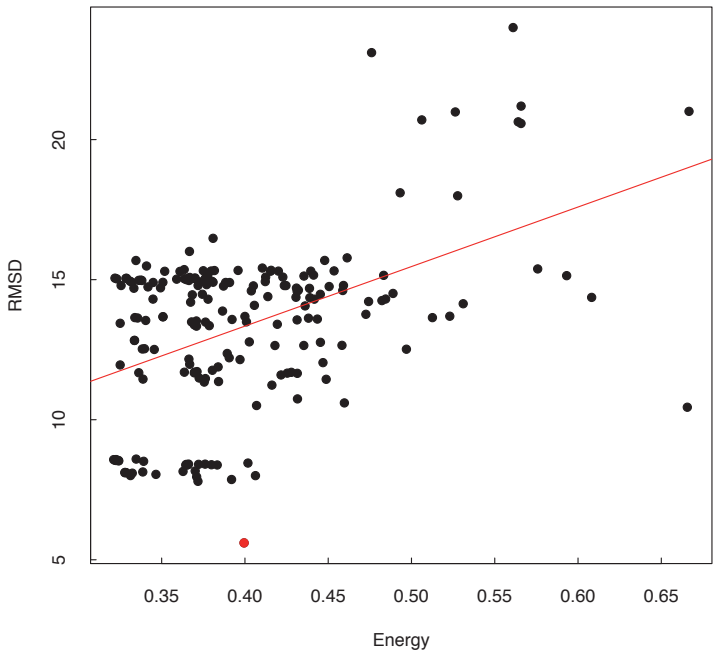

Figure 3.15: RMSD and total energy correlation. The black dots show the value pair of RMSD and total energy. The red line gives the correlation between them. The model with the best RMSD value was highlighted as red dot.

From all the optimized trace models, we choose the one with the lowest RMSD value to check the detailed information of the structure. The topology score is 1 which means topologically the "best" model is correct as our target. However, if we look into the detail of the structure as shown in Fig. 3.16, irregular connections still exist. In Fig. 3.16, the Panel A shows the best model and the target structure together with coloring method corresponding to the order of the residue. The topology match can be checked from the color coding. Two major wrong connections from two different domains are listed in panel B and C. The green ribbon showing the correct trace and the black one represents our resulting model. Those are the main affections of the quality of the final result.

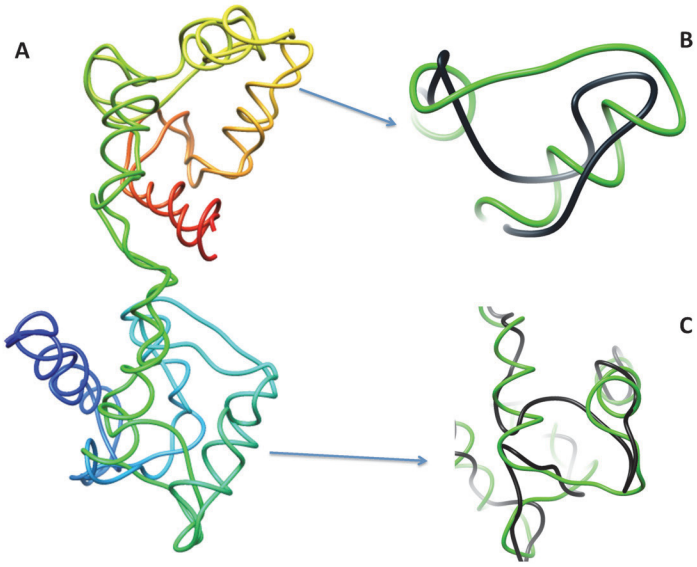

Figure 3.16: Comparison between optimal model and the target. Panel A shows the two structures with color scheme showing the order of the amino acids. B and C pointed out the wrong connection areas.

3.2.3.2 Energy Function Optimization

We have tested the LKH algorithm which works well for tracing the backbone over the perfect Cα positions , however when we try to identify the trace for unprecise position—bead model, the optimized model in the end still have wrong connections. The main reason for that is because of the unprecise bead positions make the whole energy landscape more rugged so that during the optimization the structure gets even more chances to be trapped in the local energy minima. In case of this, some biased sampling methods are needed to improve the efficiency of optimization. The biased sampling method would increase the probability to generate the model it preferred and guides the optimization to the direction of the global minimum.

For our backbone tracing problem, to bias our sampling procedure we designed a new energy function as described in Eq. (3.2.2). Except the energy term E_{lk} we used previously, three additional energy terms were added and they are E_{sse}, E_{mj} and E_{dens}, which play different roles in the entire energy function.

- **Structure based—Secondary structure energy**

  Secondary structure introduced in Section 1.1.1, as the very first level of protein structure which offers the three dimensional information, it can be used to provide useful geometry information.

  We designed a structural based secondary structure energy term:

$$E_{sse} = \sum_{k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} (d_{i,j} - d_{ij}^{seq})^2$$

  The energy calculation is based on the fluctuations between the pairwise atom distances belonging to the same secondary structure element (SSE), with respect to the reference values. Then this energy term is normalized by the total number of atom pairs appearing in the energy.

  To get the secondary structure information, the result can be directly calculated with DSSP [97] if crystal structure is available. Otherwise, there are a several tools such as PSI-pred [78], JPRED [98], PREDATOR [99] and YASSPP [100] etc. The best modern SSE prediction method is reported to reach about 80% accuracy. With such a high accuracy, model building with the prediction information would be very helpful.

  For our calmodulin test, we use PSI-pred to predict the SSE information from the sequence. The predicted result can be see from Appendix D and it is almost the same as the DSSP result if we only consider the helix and β-sheet information.

  SSE energy tests first was implemented with the bead model alone, then we take out the one with the lowest energy which only contains the E_{sse} that means the configuration contains the secondary structure that is closest to the target structure. We made a topology comparison between the lowest energy model and the target structure as shown in Fig. 3.17. This model has a RMSD value of 24.4 Å as most of connections were broken and reconnected with distant bead. Although the structure seems messed up entirely as shown

in Fig. 3.17 A, if we check the detail information especially the secondary structure, several local sections of the model are optimized well and match to the corresponding target secondary structure. From the target sequence (Appendix D), eight segments of secondary structure were identified by the prediction and 5 fractions of them were figured out after optimization. The distortions of the SSE in the optimized model came from the unprecise bead positions.

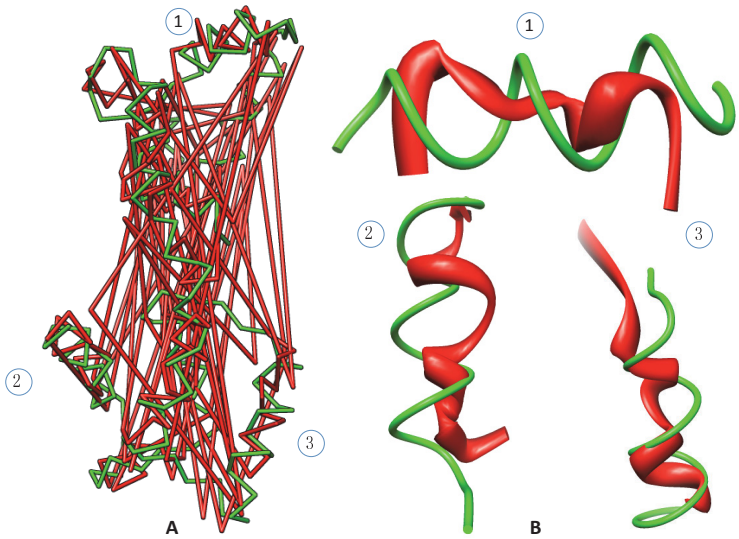

Figure 3.17: Effect of SSE energy. The A panel shows the optimized model (red) and the target (green). In B, three helices with the number corresponding to the number in A.

As the SSE energy alone has a beneficial effect in sampling the structures which have better SSE, so we combine it with the basic LK energy and investigate how they work together. To check that, for the two weighting factors α and β in Eq. (3.6), we chose two sets of different values and test with the bead model. As shown in Fig. 3.18, the best RMSD value approached when using larger α and smaller β. The LK energy dominates the optimization procedure and SSE energy would have an effect for localizing secondary structure regions. High weight for the SSE alone would not converge to a good structure. The best result was obtained with $\alpha = 1.0$ and $\beta = 0.1$.

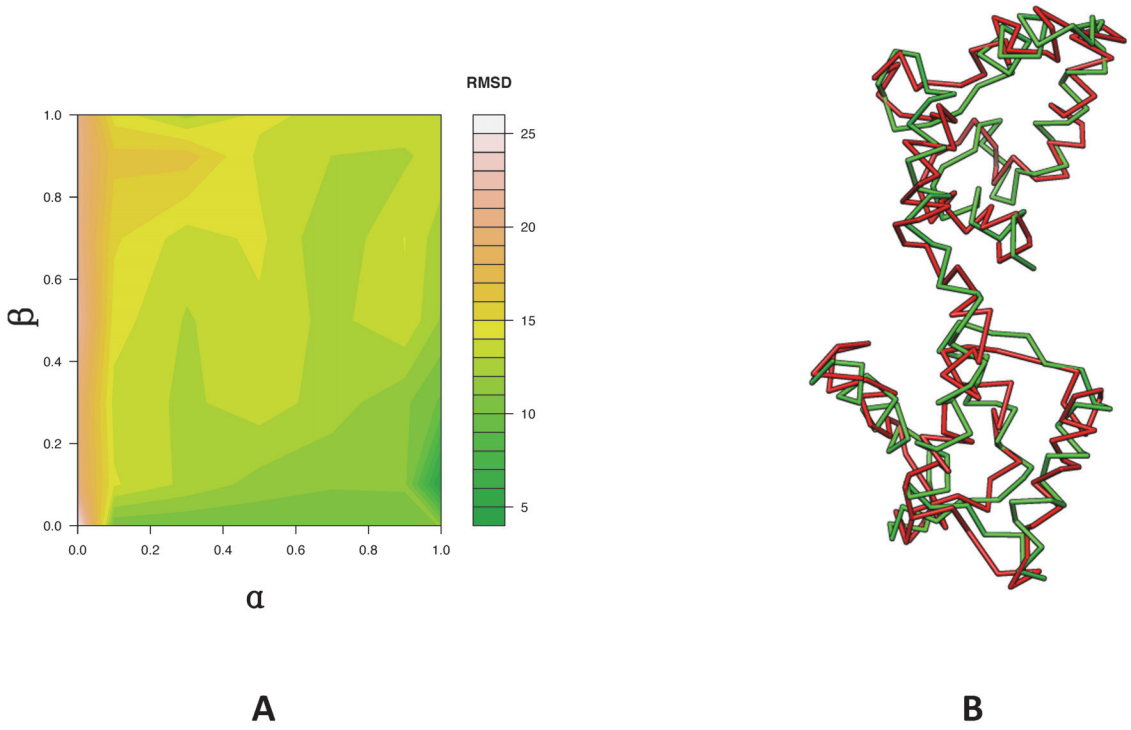

Figure 3.18: SSE result and $\alpha-\beta$ relation. Panel A shows the relationship of α and β. Panel B shows the best model by using SSE energy and aligned with our target structure.

By combining the LK energy and the SSE energy, the optimum result we obtain has the lowest energy of 1.331 and a RMSD of 4.7. The comparison between the model and the target is shown in Fig. 3.18 panel B. In the optimum result, the old irregular connections shown in Fig. 3.16 were repaired.

However, the running time by combining with SSE energy was not as good as before. For each single run, the time spent is about 94s which is almost 45 times longer than using LK energy alone. There are two reasons for this: one is for the SSE energy calculation. The SSE energy have to include all pairs of atoms in each secondary structure segments and for each optimized step, this energy has to be calculated over the entire model. The other reason is for the amino acid reassignment. After each optimization step, the sequence has to be reassigned to the beads for both directions of the path (corresponding to the sequence C$\rightarrow$ N and N $\rightarrow$ C). This point doubles the

calculation time of the first point so that the efficiency of this procedure is lower.

- **Sequence Based—Miyazawa Jernigan Energy**

  As the SSE energy term depends strongly on the secondary structure prediction, in cases when those predictions of SSE is not accurate, the optimization process could be easily guided to the wrong direction. In that case, using the SSE energy term is not recommended.

  For this reason, we designed the SSE free energy term which is a sequence based energy. Sequence based energy, just as the name implies, is an energy term based on the sequence content which is the primary information of the protein. Although it is also based on a structure frame, it is independent of any regular structure patterns. Our energy function was designed as:

$$E_{mj} = \sum_{i<j} M(a_i, a_j) D_{ij}$$

  The energy calculation is based on the MJ contact energy Eq. (3.2.2). This energy shows how well certain amino acid pairs favor the interaction within a certain distance.

  An initial test was performed only using the MJ energy term. The model with lowest MJ energy -46.07 was picked and is shown in Fig. 3.19, most of the connections seem rather random. It looks a bit similar to the SSE energy result, however, the SSE energy effects are gone and substituted with local connections mostly. As shown in Fig. 3.19, most of connections are converged to a local area and the those long connections which span over two domains as in the starting model in Fig. 3.14 do not exist anymore. However, the right topology of a protein structure is still away from that. If we put the density map together with the models, clearly there are some connections cross through the density and leave most of the bonds outside the map.

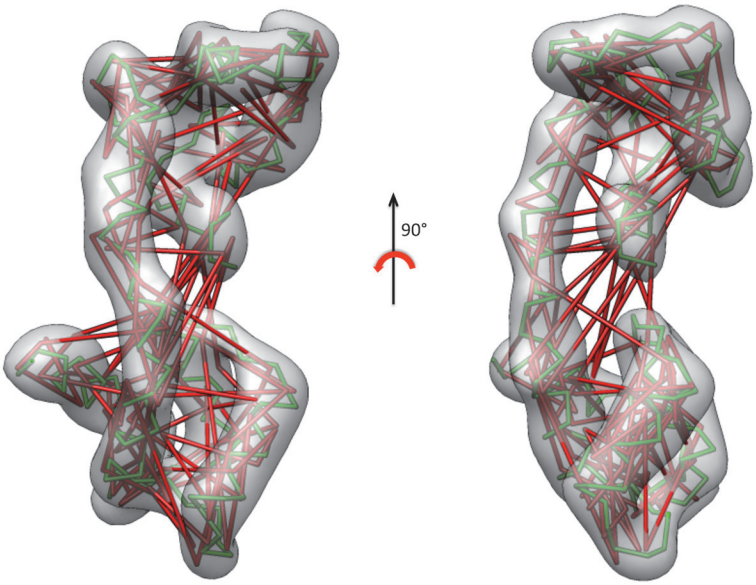

Figure 3.19: Effect of MJ energy. The left panel shows the model with the lowest MJ energy (red) and the target (green). Right panel shows the view of same combination by a 90 degree rotation.

As the LK energy plays a dominant role while combining with SSE energy, we also did tests for LK and MJ energy together. To check that, for the two weighting factors α and γ in Eq. (3.6), we chose the same value sets as for SSE and test it with the bead model. As shown in Fig. 3.20, the best RMSD value was obtained when using larger α and smaller γ. The same trend as the SSE result, LK energy dominate the optimization procedure and MJ energy has some effect on the local regions. High weights for the MJ alone would not converge to a good structure. The best result with RMSD at 8.1 Å and an energy of 60.29 was generated with $\alpha = 1.0$ and $\gamma = 0.1$. That shows that still LK is most important even when combined with the MJ energy. However, by reducing the information amount from LK and moderately increasing the weight of MJ energy, good results can be also achieved. High value for the weight of MJ and low for LK would not converge to a good model.

Regarding the running the speed of LK and MJ together, it was 87s which is close to the SSE result because the calculation also has to consider the double directions of the sequence and each time the neighbouring beads energy have

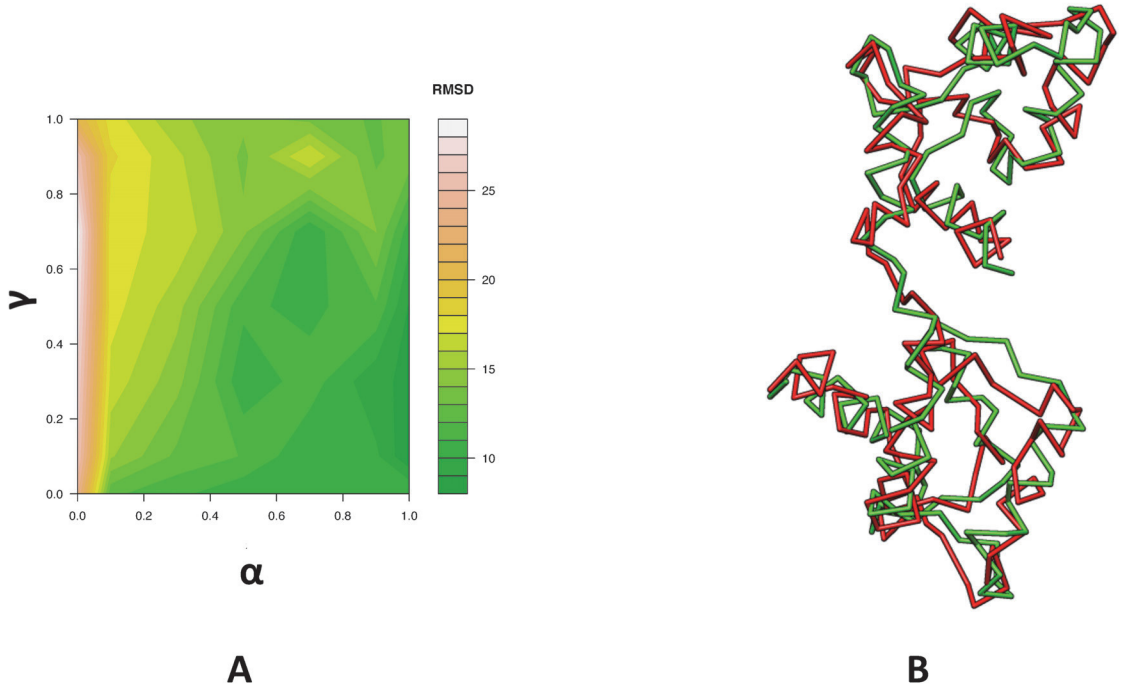

Figure 3.20: MJ result and $\alpha - \gamma$ relation. Panel A shows the relationship of α and γ. Panel B shows the best model by using MJ energy and aligned with our target structure.

to be recalculated.

- **Density Based—Density Map Energy**

  As we have seen from the MJ energy optimization, there are some cross connections inside the molecule and those connections destroy the overall topology. These cross connections not only happen when using MJ energy but also exist in other optimization procedure even when using LKH algorithm alone. If we look at those cross connections without any target, it might be difficult to decide whether they are wrong or not. When we put the structure into the density map, those irregular connections clearly stand outside the density map. At the same time, the energy terms we used so far (either sequence or structure related) have nothing to do with the density map.

  Owing to these reasons, we devised the third energy term with the aid of density map and categorized as the density based group. This energy term is

independent of any sequence or structure information. It can be calculated as:

$$E_{inter} = \begin{cases} 1 & \rho_{inter} > \rho_{cutoff} \\ 0 & \rho_{inter} <= \rho_{cutoff} \end{cases} \tag{3.13}$$

$$E_{dens} = 1 - \frac{\sum_{i=1}^{n-1} E_{inter}(i)}{n-1} \tag{3.14}$$

It is weighted by a factor δ as shown in Eq. (3.6). This energy term provides a pseudo—energy between pairs of beads. For this calculation, several pseudo—grid points are interpolated by the trilinear interpolation method. Afterwards, the pseudo—energy would be accumulated to represent the energy for the connection between two beads.

By importing the density map, we aim at removing the wrong connections that should not show up where there is no density between the beads. To set this up, we have to define a certain region so that the beads can recognize where the density information needs to be considered. For that, we specify a density threshold and the way to calculate this threshold is introduced in Section 3.1.1. As in the other tests, the density energy term was first implemented alone to analyze its effect and the density threshold for the 8 Å calmodulin map is Section 3.1. The optimized bead model and the target were shown in Fig. 3.21.

As shown in Fig. 3.21, similar as the MJ energy result, all the connections are fined into the local regions and differently, there is no connection which span over the undefined density area. The final model contains only the connections that are inside the density.

To check how the DENS energy E_{dens} works in combination with the main LK energy, we made the tests with the two terms together. The factors α and δ, were set to the same value as for the SSE and MJ energy terms , as described above. The best RMSD value was 6.8 Å its final energy was 1.165

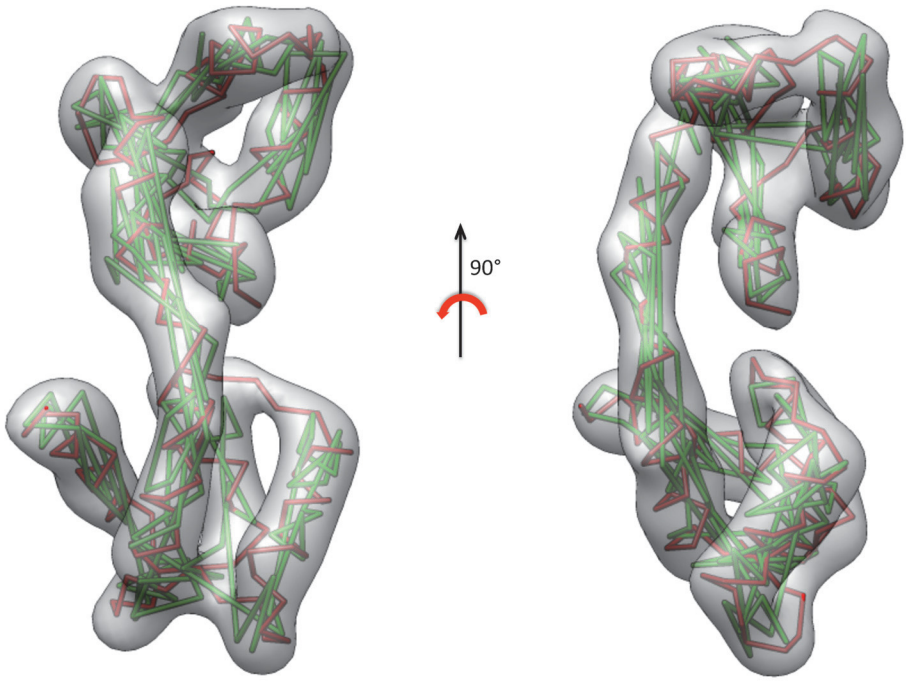

Figure 3.21: Effect of density energy. The left panel shows the optimized model (red) and the target (green). On the right, same content was shown by 90 degree rotation.

and the best model was obtained with $\alpha = 0.9$ and $\delta = 1.0$. In general, as shown in Fig. 3.22, the optimization tends to yield better results when α and δ are both large, and vice versa.

Except for the different factor setting, the energy term E_{dens} works more efficient than the other two. For a single test, it takes 35.59s. The calculation for this is unrelated to the sequence change right now and only corresponds to the two paired beads.

- **Combination of Multiple Energy Terms**

  As previously tested, all the different energy terms have different effects during the structure optimization. To check the integrative effect of all the restraints, we did a test with the combination of all three energy terms. The function was written as:

$$E_{total} = \alpha E_{lk} + \beta E_{sse} + \gamma E_{mj} + \delta E_{dens},$$

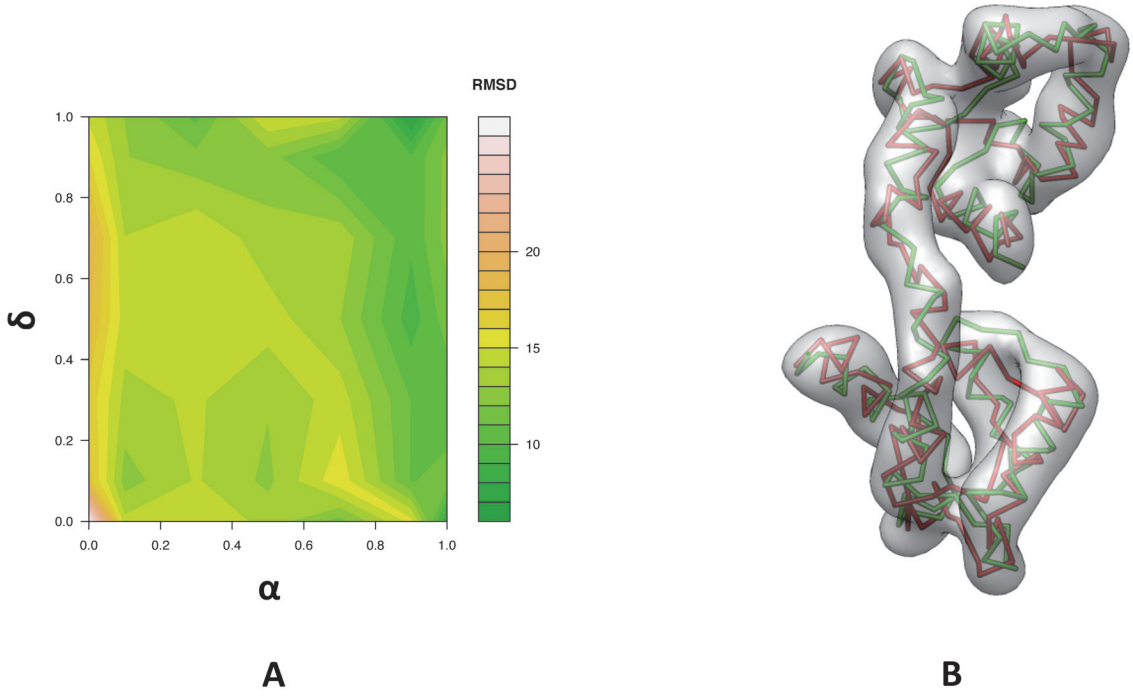

Figure 3.22: DENS result and $\alpha - \delta$ relation. Panel A shows the relationship of α and δ. Panel B shows the best model by using DENS energy and aligned with our target structure.

In Section 3.2.3.1, without different restraints, we had found the correct topology from the bead model. To check the effects of the entire energy term, we chose a bead model which was shown before to result in the correct topology in our test case so that the effect of the restraints can be analyzed by the improvement over the model from the optimization procedure without restraints.

In Eq. (3.6), different restraints are weighted by different factors which control the amount of useful information each parameter contributes. The result from the previous sections shows the pairwised relationship between different restraints factors and the main LK energy factor. The LK energy is the main driving force and its weight should be large to obtain better results as shown in the relationship result in Section 3.2.3.2. The best result came with the factors assigned by the following values $\alpha = 0.9$, $\beta = 0.1$, $\gamma = 0.1$ and $\delta = 0.3$. The optimal optimized model is shown in Fig. 3.23 and the RMSD value between the model and target is 4.6 Å. The topology score

between these two structures was 1 and the identical topology was shown as the same color distribution by color coded amino acid from dark blue which represented first amino acid to red which means the last amino acid.

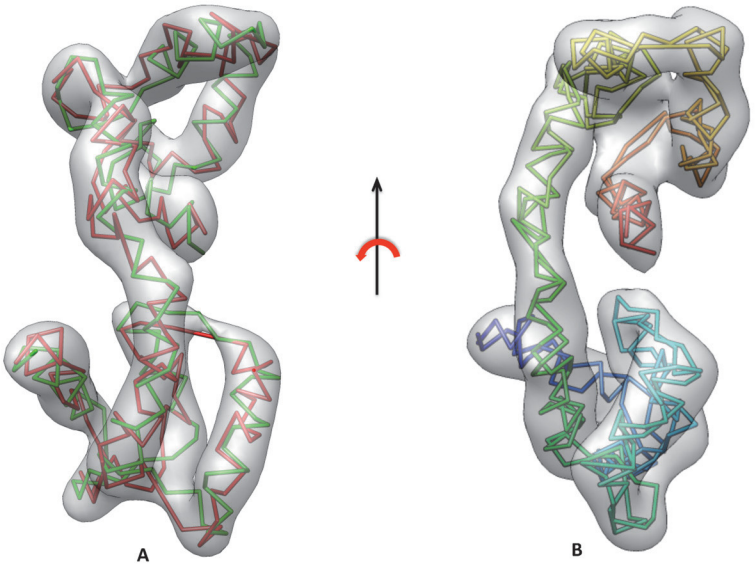

Figure 3.23: Optimal model with all restraints. Panel A shows the optimal bead model (red) superimposed with the target structure (green). Panel B shows the same content with a 90 degree of rotation but colored according to the order of amino acids.

The best result without the restraints were shown in Fig. 3.16 and the RMSD value was 5.6 Å. By applying the restraints, the resulting model has a RMSD value of 4.6 Å. That means the restraints could guide the optimization towards traces which are restrained by different potential energies. However, as three energy calculations were added, the running time of one single test run was about 120 seconds.

3.3 Refinement

All the previous sections described the methods we used for the tracing algorithm. To further improve the resulting model, we need to do a refinement of the connected beads. There are mainly two parts where we can consider to do refinement: One is the bead positions refinement and the other is the refinement after the optimization which is the final refinement.

3.3.1 Bead Refinement

The bead refinement here is different from that in Section 3.1.3 which focuses on refining the beads to better positions. The bead refinement here emphasizes more on refining the beads on the whole energy landscape and making the entire arrangement of the beads easier for the tracing algorithm.

Here we move the beads into the spatial central axis of the corresponding density segment by using Direx which was introduced in Section 2.3.1. Here the map values were scaled by

$$\rho_{dens} = \rho_{dens} - \alpha\rho_{currentmap}. \tag{3.15}$$

The bead model was refined to the scaled map ρ_{dens}. By this, the beads tend to move to the center of the density map and it ended with the beads lined up in the density. However, this refinement step would not consider the distance restraint we added in Section 3.1, that means even some beads may even clash together . We use the real absolute distance difference instead of harmonic potential for the LK energy term. This makes the LK optimization the same as the TSP.

The rational idea behind this method is that by implementing this refinement the lined beads would have a smoother energy landscape than it has before. This makes it much easier for the optimization to find the optimum result.

3.3.2 Final Refinement

The previous sections introduced all the optimization procedures, after that we pick out the best structure from the ensemble of tours. However the best tour may not have the best Cα backbone, because the bead positions are not perfect and some local areas are not in the correct topology state either. To improve the geometry of the tour, we need to perform one more refinement step.

We refined the optimized tour with Direx which is combined with the secondary structure restraints. The distance information for the secondary structure restraints is the same as the distances used for the energy term E_{sse}, as is shown in Appendix A. This refinement is based on the secondary structure assignment for which we use the program DSSP [97]. The assigned secondary structure information is combined with the distance information and used in Direx refinement. All the distances within the helices and β-sheets are restrained by the standard distances.

This final refinement focuses on the topology refinement locally by applying the secondary structure information. The structure after this final refinement has a correct topology which fits better to the density map than it does before.

3.3.3 Calmodulin Test Result

3.3.3.1 Bead Refinement

In our calmodulin test, we use the beads generation method described in Section 3.1. Afterwards, the beads were refined by DireX according to the method introduced in Section 2.3. The Direx refinement here was aiming to move the beads to a relative higher density region. By setting the perturbation factor to a large value of 0.3, DireX would allow the bead to make a large movements which are then corrected by the CONCOORD restraints . At the same time, the density difference factor α in Eq. (3.15) was set to 0.3. After 10 of steps DireX refinement,

the beads were moved into higher density region towards the central axis of the density as is shown in Fig. 3.24 B.

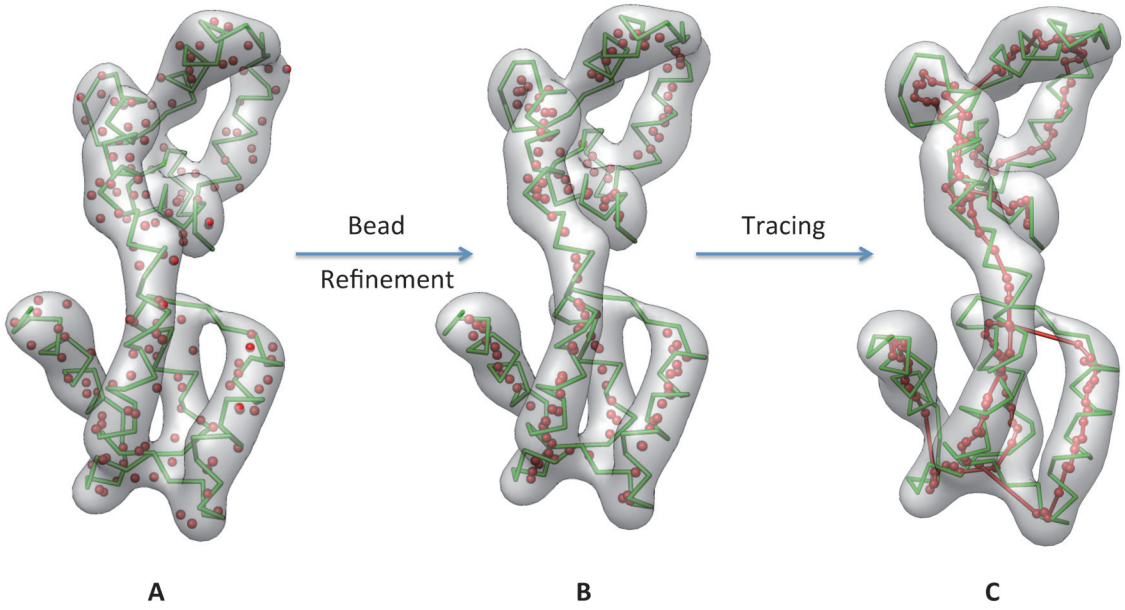

Figure 3.24: Bead refinement. Panel A shows the bead model (red) and the target backbone (green); B shows the refined bead model (red) which is centralized in the density and the target (green); C shows the traced backbone (red) and the target backbone.

The initial bead model and the refined bead model shown in A and B in Fig. 3.24 were both randomly connected. The DireX refinement procedure disarranged the initial distance relations so that the original LK harmonic potential (Eq. (3.7)) was not suitable to be used as the optimization function. Instead of using the harmonic potential, we simplified Eq. (3.7) to:

$$E_{lk} = \sum_{i=1}^{n-1} d_{i,i+1} \tag{3.16}$$

in which means E_{lk} is just the length of the trace. As the distances between beads were destroyed, the distances between the amino acids and within the SSEs do not

fit with the MJ energy and the secondary structure energy anymore. Therefore during the optimization procedure, only the LK energy term was used. After tracing with our algorithm, the random beads were threaded with a trace which has the shortest connections as shown in Fig. 3.24 C. The RMSD value dropped from 24.4 Å to 4.7 Å comparing with the target backbone structure. The topology score between the best trace and the target protein backbone is 1, which means they were topologically identical.

3.3.3.2 Final Refinement

From the aforementioned results, the beads after being processed by LK optimization alone and processed by DireX refinement first and then LK optimization are both topologically correct. However neither of them had the correct backbone structure, especially for the model after the bead refinement in which the tracing result was just a single string like structure. None of them showing a proper protein structure as shown in Fig. 3.25 B and C and model B and C had a RMSD value of 4.6 Å and 4.7 Å respectively.

The aim of the final refinement is trying to refine the bead model as close as possible to the target structure based on the correct topology. The inaccurate positions of the beads is responsible for the model distance from the target structure, which does not show a proper protein structure. The final refinement accomplishes this by refining the bead positions with DireX.

To refine a model closer to the target, the secondary structure information is very useful for DireX. The secondary structure was predicted by PSIpred using the primary protein sequence (Appendix C) as the only input. The predicted result shows which segment in the protein belonged to which secondary structure element as shown in Appendix D. Based on the predicted secondary structure information and the standard Cα-Cα distances shown in Appendix A, a file containing the distance restraints for all the secondary structure elements were produced and formatted as shown in Appendix E. It was used as the distance restraints input file for the DireX refinement.

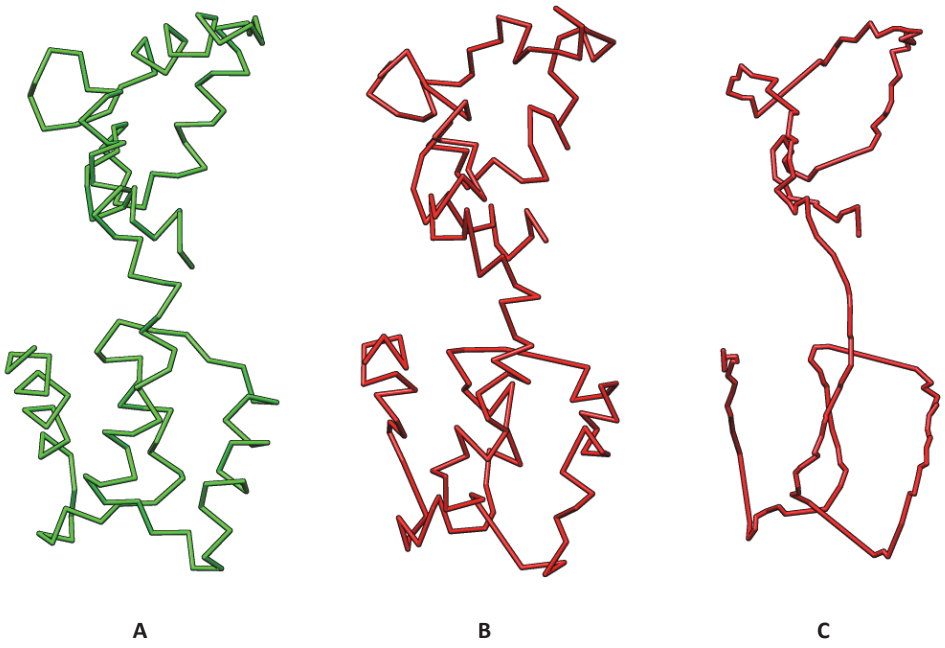

Figure 3.25: Comparison of optimized model and target. Panel A shows the target backbone structure (green); Panel B is the model after optimization with entire energy function (red); Panel C shows the model from DireX refined first and then optimized result (red).

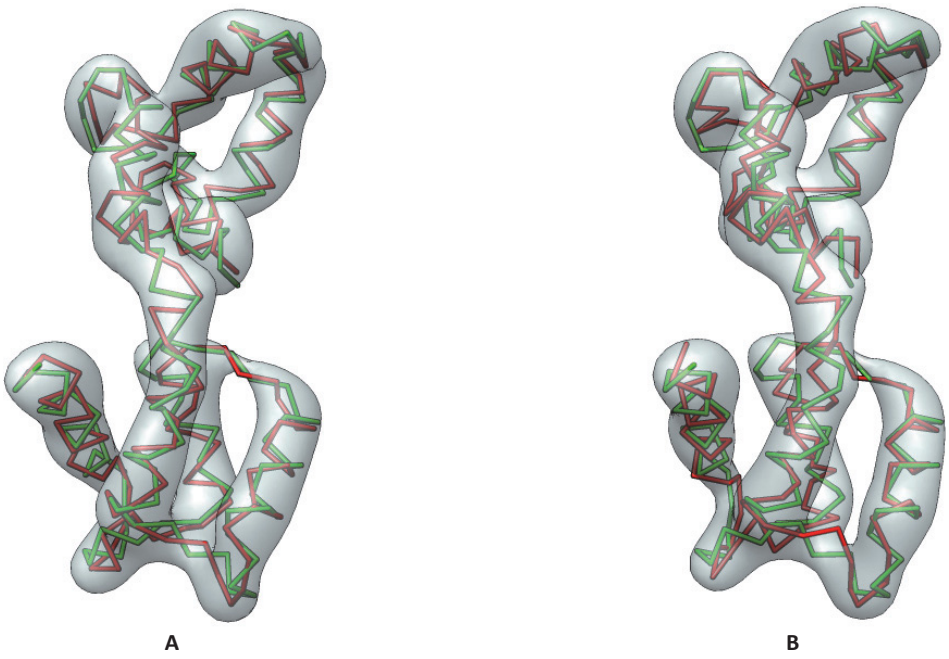

Figure 3.26: Comparison of refined models with target. Panel A shows the target backbone structure (green) and the refined model (red) from model B in Fig. 3.25; Panel B shows the target (green) and the refined model (red) from model C in Fig. 3.25.

Our DireX refinement was implemented on both model B and C as shown in Fig. 3.25 with the same set of secondary structure restraints. The refined result was shown in Fig. 3.26. The RMSD value between the two structures in A of Fig. 3.26 is 2.97 Å and in B of Fig. 3.26 was 3.29 Å. The RMSD values were improved by 1.612 and 1.416 Å respectively. The map correlation is 0.93 and 0.937 for the two models. Structurally, both of them became better and the secondary structure elements showing the right pattern and stay closer to the target.

3.4 Optimization Protocol

In the previous chapter, we introduced all the methods that have been used in our tracing algorithm. To test our algorithm with different systems, we generalized the protocol. It includes three parts: initialization, optimization and refinement. The whole process is depicted in Fig. 3.27.

In general, the initialization step contains the procedures to process the map and generate the beads. The refinement step includes the refinement of the beads which is implemented before the optimization and the final refinement of the topology happens after the optimization. The optimization is the core algorithm which is the combination of LK and MC-SA as we described in the method part. In Fig. 3.27, these three different sections are highlighted in different colors.

Initially, the input density map is processed and then each bead which represent one amino acid is placed into the processed density map. Afterwards, the bead locations are refined either slightly or more aggressively to the central axis of the density segments. Refined beads are directed as input for the LK-MC-SA optimization. After a certain number of optimization steps, one optimum result is picked out of the ensemble of structures. This optimum structure is further refined using the predicted secondary structure information with Direx.

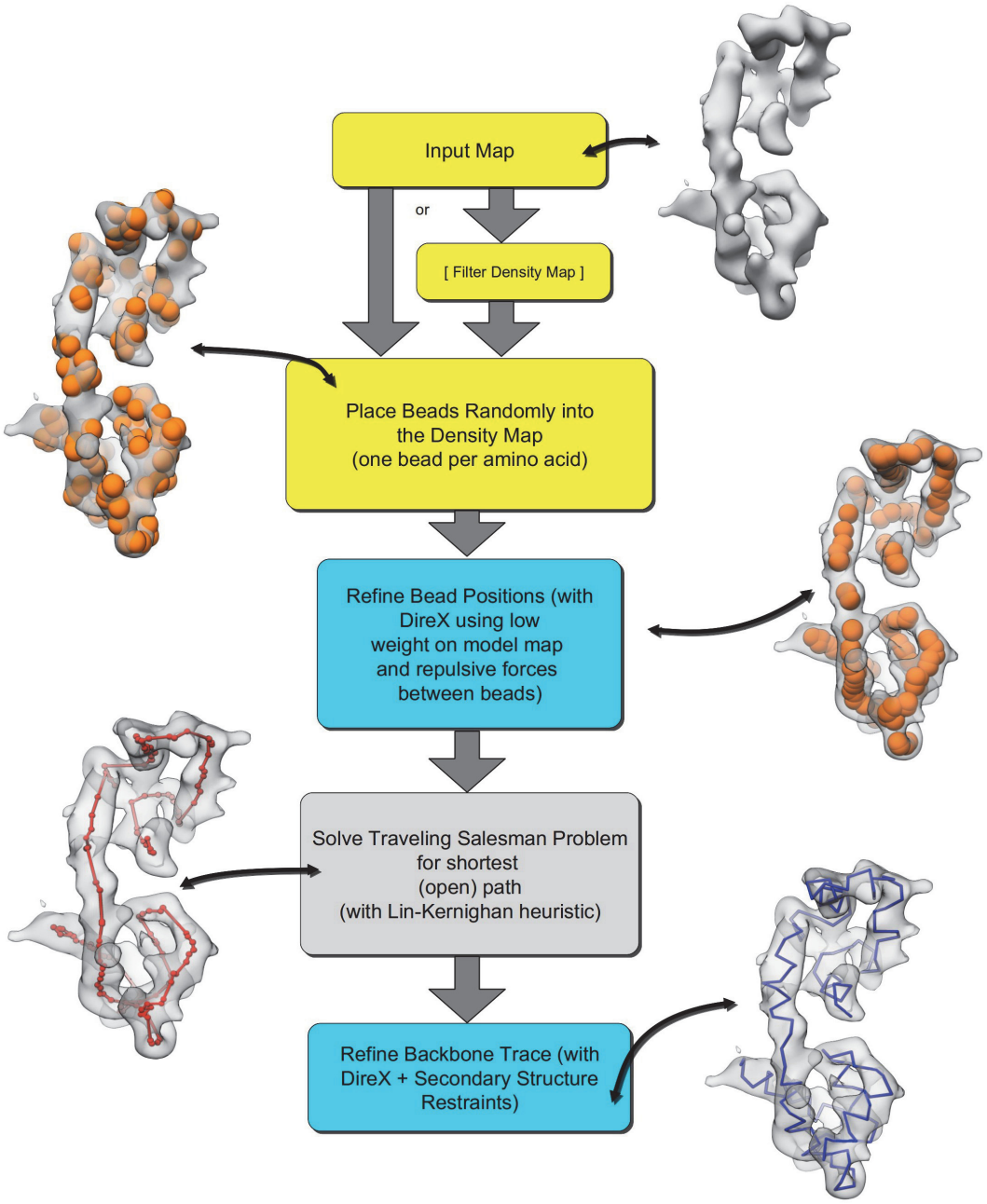

Figure 3.27: Tracing Protocol. The yellow background stands for the initialization,blue color highlight the refinement and the gray means the optimization procedure. The density map is shown in gray and stylized as transparent and balls inside represent the beads.

3.5 More Test Cases

As a test model, calmodulin has been used for tests in different situations. However, to demonstrate that our protocol can be widely applied, we did several additional tests with other proteins. For these tests, we chose our target structures according to the protein classification defined by CATH [101]. In CATH, protein structures are classified by their secondary structure composition. The three main groups are structures that mainly contain α-helix, β-sheet and a mixture of α-helix and β-sheet mixed.

According to the CATH classification, we selected 6 structures in total, 2 for each class. The results of each part are shown in the following sections.

3.5.1 α-Helical Structures

In this category, the two structures we selected from the PDB were 1AEP (PDBID) and 4GOW (PDBID). 1AEP is an apolipoprotein which isolated from the African migratory locust Locusta migratoria [102]. It consists of five long α-helices connected by short loops. The target structure was determined by XRC to a resolution of 2.5 Å. 4GOW is also a calmodulin which is in charge of regulating voltage-gated potassium channels. It was determined by X-ray crystallography at the resolution of 2.6 Å [103]. Both 1AEP and 4GOW are composed of α-helical elements. In contrast to 1AEP, 4GOW has a structure with short helices which are closely packed to each other. Comparing with our calmodulin test model, 4GOW is a binding protein which does not contain the short β-sheet segment as in the test structure. Additionally, in the structure of 4GOW, there is a small gap between residue 78 to 81 while are missing, which is also visible in the original density map. Both target structures are shown in Fig. 3.28.

For both of these two helical structures, we simulated their density map at a resolution of 8 Å as shown in Fig. 3.28. With the density map, the corresponding backbone density threshold could be calculated as described in Section 3.1.1. The

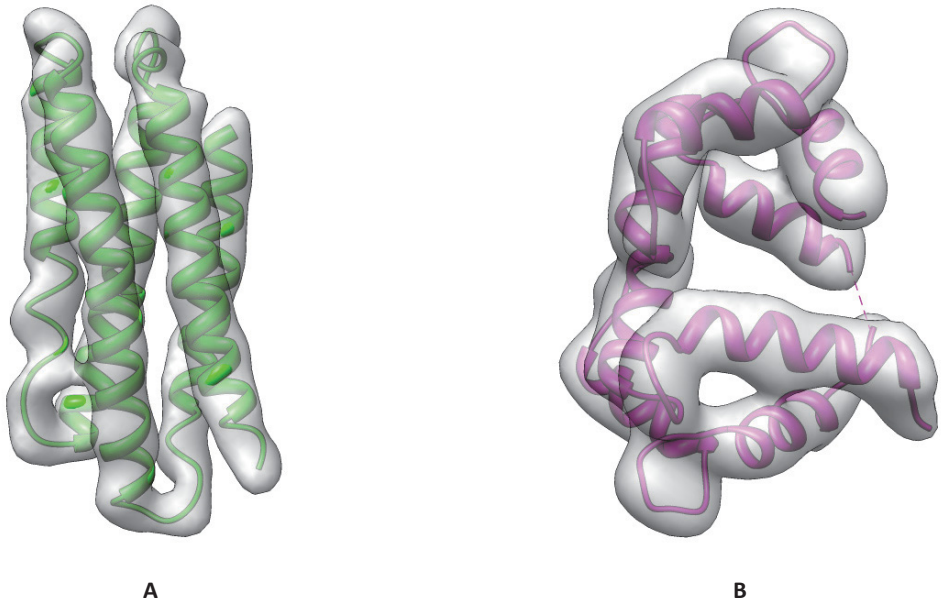

Figure 3.28: α-helical targets. Panel A shows the 1AEP target structure (green); Panel B shows the 4GOW target structure (purple). The red dashed line in B marks the gap between residue 78 and 81.

density threshold for 1AEP and 4GOW were 2.41 and 2.53 respectively. Corresponding bead models were generated using these values and the bead models are shown in Fig. 3.29.

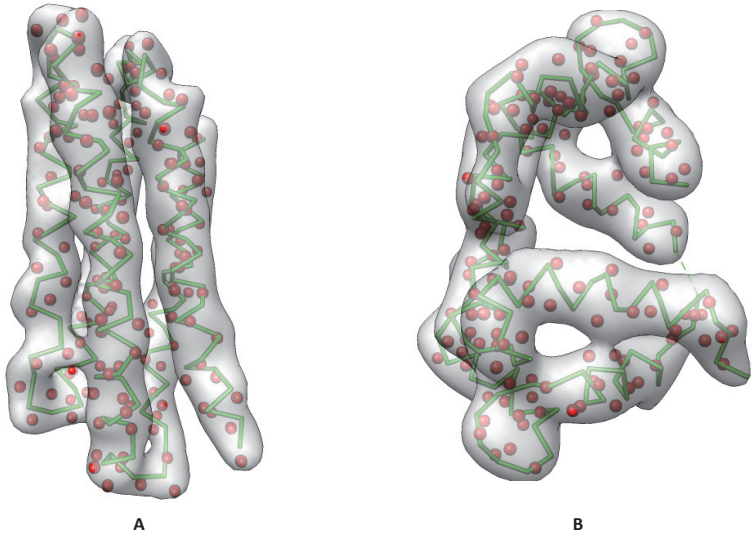

Figure 3.29: Bead models for α-helical structures. Panel A shows the bead model (red balls) for 1AEP and the target backbone trace (green); Panel B shows the bead model (red balls) for 4GOW and the target backbone trace (green).

Before tracing the backbone with our bead models, we made trial runs with the perfect Cα models. With the LK energy alone, we could find the perfect trace for both 1AEP and 4GOW. However, when using the bead models instead, the LK energy alone was not sufficient to find the right trace. Even though the starting and ending beads could be determined easily, the connections inside were not always correct. Most typical cases of wrong connections are shown in Fig. 3.30.

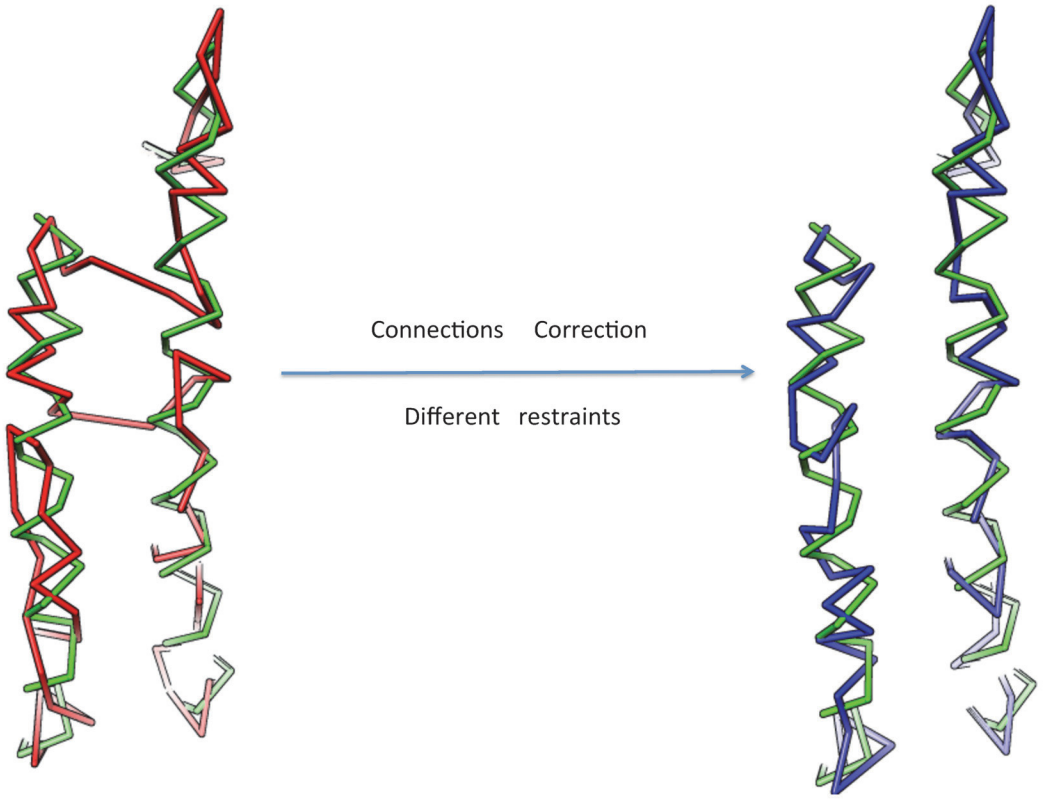

Figure 3.30: Typical wrong connections in bead model. Panel A shows the bead model with wrong connections aligned with the target structure; Panel B shows the bead model with correct connections aligned with the target structure.

This type of wrong connections exist when the distances between the beads are small so they tend to be connected by the algorithm. Wrong connections would also make the whole topology incorrect as shown in A of Fig. 3.30. After adding the restraint energy terms, those wrong connections were corrected because they would not be in agreement with either the SSE distance restraints or the density restraints.

After running with our optimization protocol, both bead models yield their correct topology as shown in Fig. 3.31. The panel A shows that both bead models had a correct topology score of 1 and are colored as the sequence order. The bead models with random initial connections had RMSD values of 23.1 Å and 22.4 Å for 1AEP and 4GOW respectively. After optimization, the corresponding RMSD dropped to 5.3 Å and 4.8 Å respectively. Both sequences information (Appendix F) were used for secondary structure prediction by PSIpred. The secondary structure restraints were generated from the prediction (Appendix G). By using the secondary structure restraints, the final RMSD values reached 4.2 Å and 3.8 Å respectively.

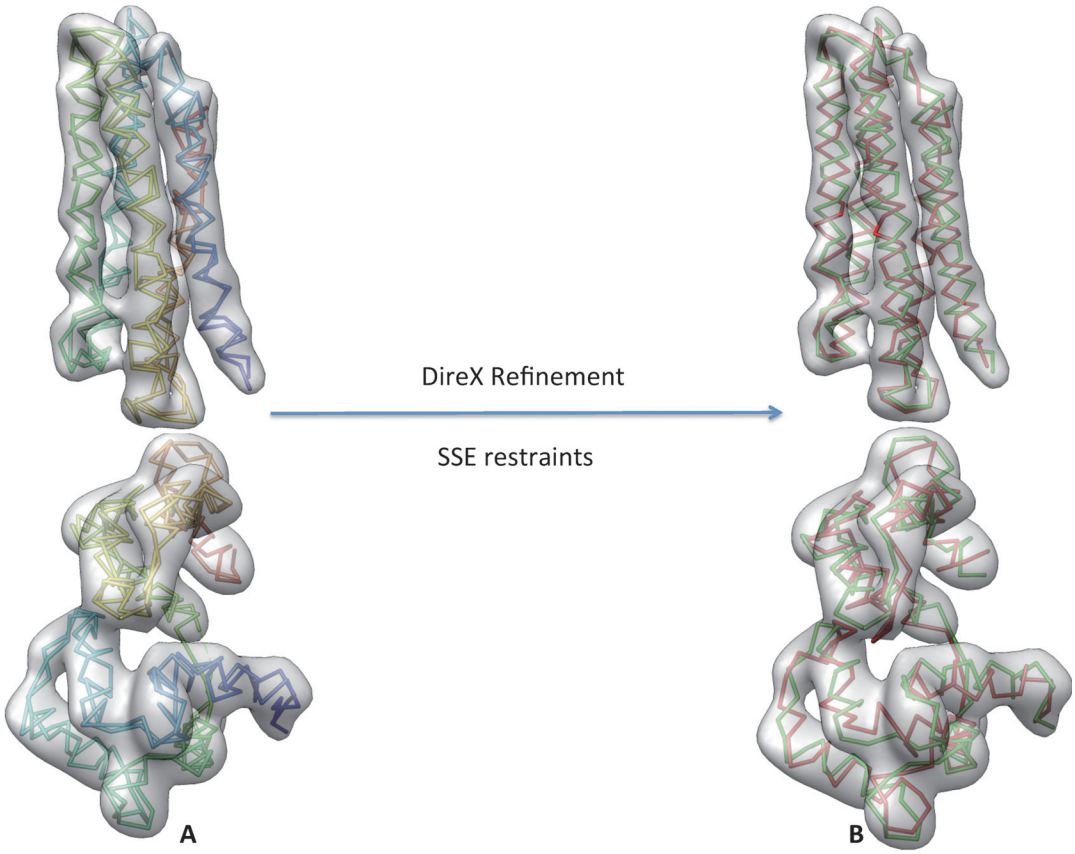

Figure 3.31: 1AEP and 4GOW tracing and refinement result. Panel A shows the bead model with correct trace and together with the target and both are colored according to their amino acid order; Panel B shows the refined bead model (red) with the target structure (green).

Even though the structure of 4GOW has a small gap between residue 78 and 81 and the corresponding region had no density, our tracing algorithm could determine the correct connection which fills the gap. It can be seen from the lower model

of Fig. 3.31. After refinement, both of the models had distinguishable α-helices showing up when imposing the α-helical distance restraints. Although both models improved after DireX refinement, there was still a considerable shift between the trace and the correct backbone structure, which means the bead positions were not correctly assigned inside the density map.

3.5.2 β-Sheet Structures

For the β-sheet tests, the two structures we selected from the PDB were 1DC9 (PDBID) and 3EMM (PDBID). 1DC9 is a rat intestinal fatty acid binding protein which contains a β-barrel fold. The original protein has 131 amino acids and to make it fit into the β-sheet group we shorten the sequence to 98 amino acid. The structure was determined by XRC at the resolution of 2.1 Å [104]. 3EMM is a protein from a Arabidopsis thaliana gene, it is a heme binding protein. The structure was also determined by XRC at the resolution of 1.36 Å [105]. To fit the structure into the β-sheet group, the 160 amino acid original sequence was shorten to 143. Both target structures contains mostly β-sheet except for a few turn regions which connect the β- sheets. The structures are shown in Fig. 3.32.

As introduced in Table 1.1, β-sheets can be built reliably at a resolution higher than 4Å. For the 5 Å resolution maps, the density for the parallel β-sheets may have severe overlap, which may lead to artifacts while placing the beads. To make the β-sheet density relatively clear and distinguishable between parallel β-sheets, we simulated the density map for both structures at the resolution of 4.5 Å, which is not a good resolution for β-sheets structure but it is still possible to trace the backbone. Both maps were shown together with their target structures in Fig. 3.32.

The beads were supposed to be put over the backbone density. However, the density threshold we chose would also including some side chain density so that beads might be located in some big side chain density blobs. For those beads, we

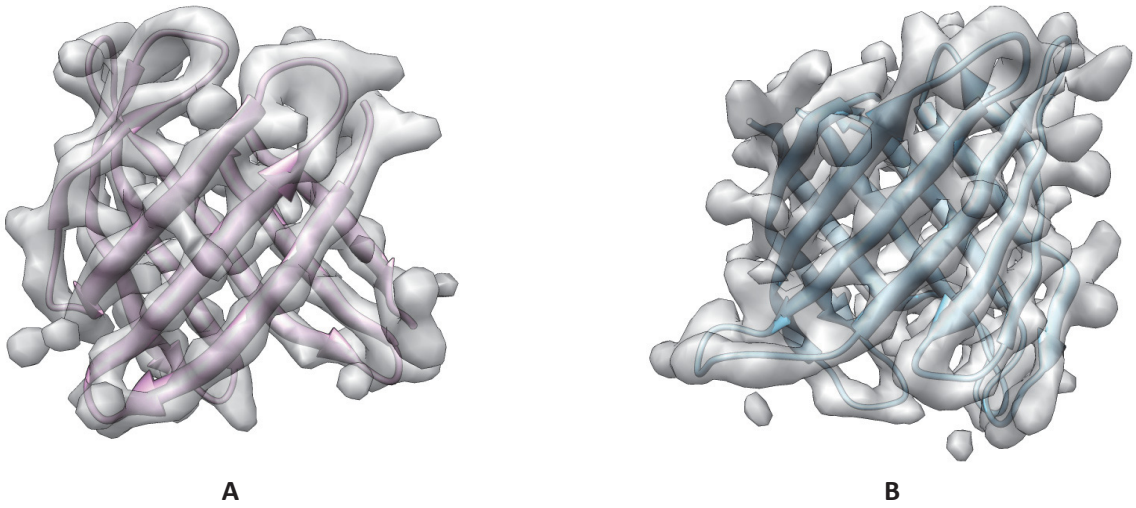

Figure 3.32: β-sheet targets. Panel A shows the target structure of 1DC9 (pink); Panel B shows the 3EMM target structure (cyan). Both targets were stay inside their corresponding 4.5 Å simulated density maps.

checked with Chimera [106] and manually moved them from the side chain regions to the main chain area. Both bead models are shown in Fig. 3.33.

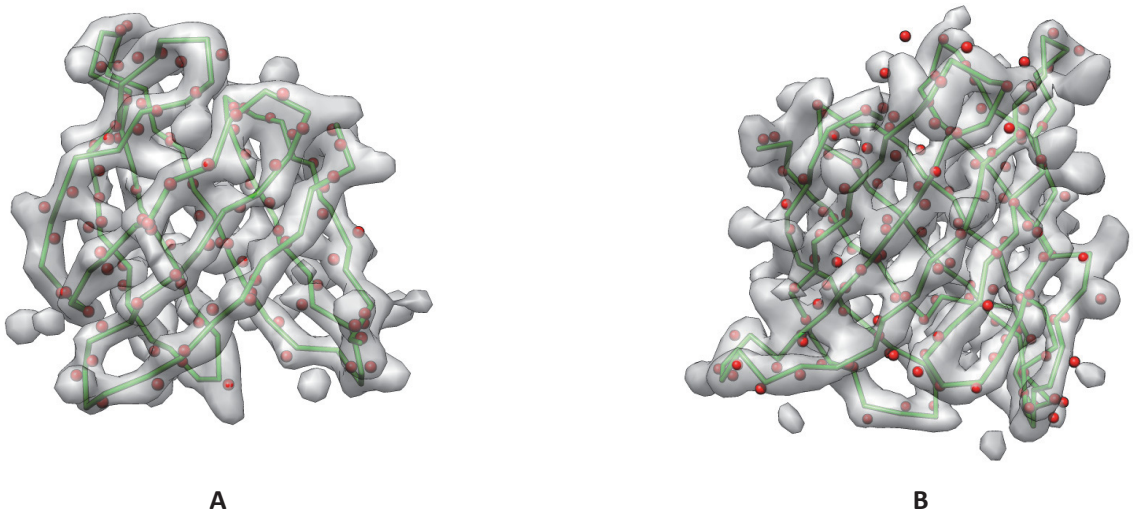

Figure 3.33: β-sheet bead models. Panel A shows the bead structure of 1DC9 (red) and the target backbone (green); Panel B shows the 3EMM bead model (red) and its target backbone (green).

During the optimization with the LK energy alone, the typical wrong connection as seen in the α-helical tests happen more frequently in the β-sheet tests. There

are two reasons for this: First, the parallel β-sheet is close to each other so the beads inside one sheet is close to the beads in the neighbouring one. The other reason is the beads are not located at precise positions. This error may lead to two beads from different sheets being placed really close to each other. However, with our density restraints, this effect can be well re-solved. The optimization result can be seen from A in Fig. 3.34.

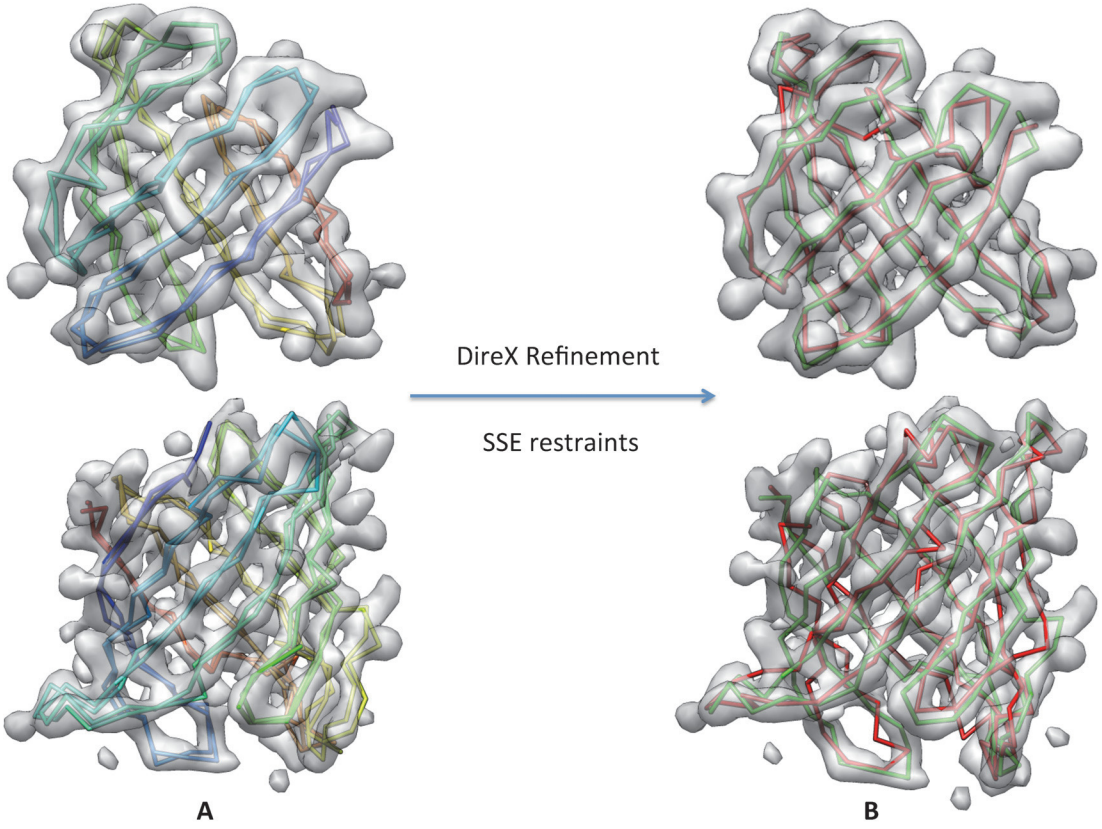

Figure 3.34: 1DC9 and 3EMM tracing and refinement results. Panel A shows the bead tracing results of 1DC9 and 3EMM and models were colored by the order of amino acids; Panel B shows refinement results of 1DC9 and 3EMM bead models. Refined models are colored in red and the target backbone is shown in green.

For the optimization of the 3EMM bead model, because the starting bead was close to the other beads and the density for the starting position was also close to nearby density, it was difficult to identify the starting and ending positions. We added three more beads to the model, which extended the starting positions and moved it far away from the beads around. After that, with our identification of the longest edge method, the starting position and the ending position could

be determined easily. The backbone trace was shown after deleting the three additional beads.

After adding the restraints, the beads could be connected straightly following the β-sheet density as the connections between neighbouring sheets have lower chance to be chosen comparing with the connections lying inside the density. The RMSD value for 1DC9 and 3EMM decreased from 17.1 Å and 20.4 Å to 3.7 Å and 4.3 Å respectively. At the same time both backbone models have the topology score as 1. After DireX refinement with the SSE restraints (Appendix G), the RMSD values dropped to 3.1 Å and 4.2 Å.

Comparing with the α-helical tests, β-sheets group did not have a better improvement according to the RMSD change. The main reason is the side chain density would have a side effect on moving the bead during the refinement especially for the map at resolution of 4.5 Å. At the same time, as the distance restraints i are rigid, some beads came out of the density as there was not enough density to fit all beads.

3.5.3 α and β Structures

To further demonstrate the performance of our tracing algorithm, in addition to α-helical and β-sheet tests, we also examined our methods with the α-helix and β-sheet mixed structures. Two structures 201L (PDBID) and 3QDD (PDBID) were selected from the PDB. 201L is a T4 lysozyme protein in charge of breaking bacterial cell walls. The 2.0 Å structure was determined by XRC method [107]. It contains 155 amino acids with 62% helical elements and 9% β-sheet. 3QDD is a protein functioning as an inhibitor of heat shock protein 90. The structure which has 203 amino acids was determined by XRC at the resolution of 1.79 Å [108]. 39% of the structure are composed by helical structure and 20% are β-sheets. Both structures are shown in Fig. 3.35.

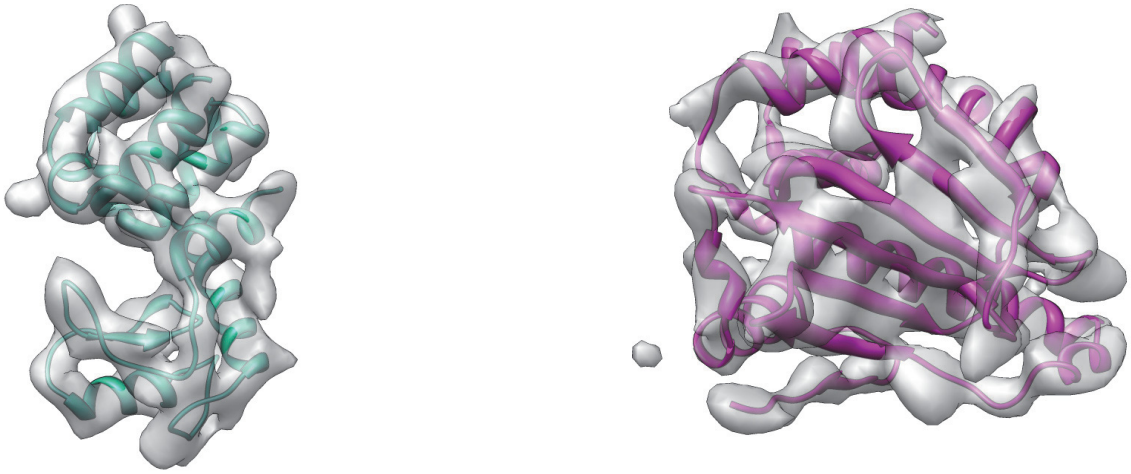

Figure 3.35: α and β mixed targets. Panel A shows the target structure of 201L (blue); Panel B shows the target structure of 3QDD (purple).

As these structures are α-helix and β-sheet mixed, the resolution of the simulated density maps should be a compromise between only α-helical structure and only β-sheet. Otherwise, the β-sheet density would be tangle together. For this reason, we simulated the density map at 6 Å as shown in Fig. 3.35.

With the sequence information (Appendix F), the threshold for 201L and 3QDD were determined at 2.9 and 2.41 respectively. In the same way as we did for the other tests, beads were placed according to their corresponding threshold and the number of beads was set to the number of amino acids in the sequences. Bead models are shown in Fig. 3.36.

With the random connections, the starting bead models for the two structures had RMSD values as 22.1 Å and 21.0 Å respectively. The correct topology was determined by running through the tracing algorithm as shown in A of Fig. 3.37. After optimization, the RMSD values dropped to 4.6 Å and 8.9 Å for 201L and 3QDD separately. DireX refinement further decreased the corresponding RMSD values to 3.2 Å and 3.8 Å.

Although the shift problem as in the other tests also happened here, as each segment of the secondary structure was not as long as the other tests, the shift effect

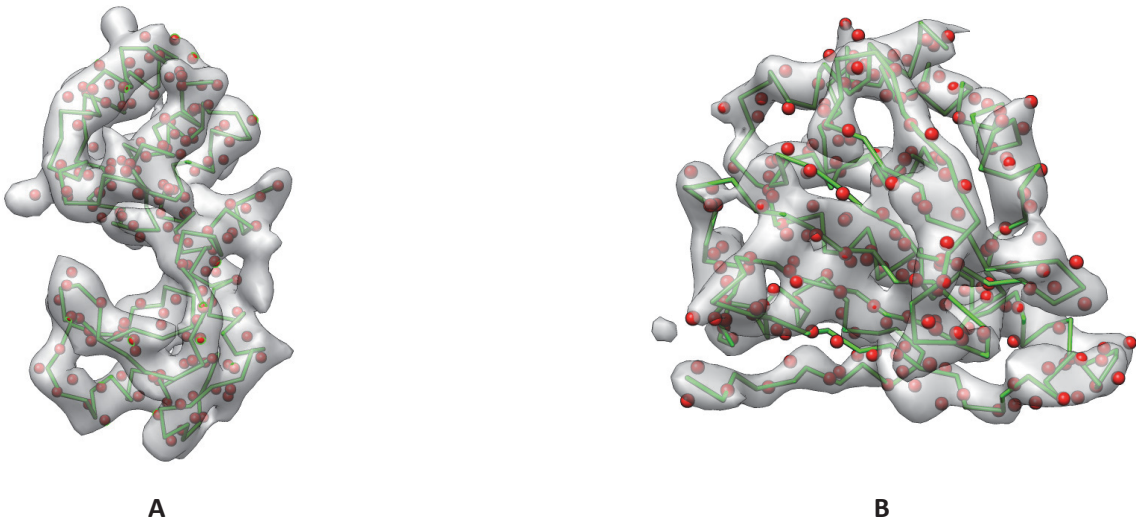

Figure 3.36: Beads model for α and β mixed structures. Panel A shows the bead model (red balls) and the target structure of 201L (green); Panel B shows the bead model (red balls) and the target structure of 3QDD (green).

was less significant than in the previous α-helical and β-sheet tests. Additionally, at 6 Å any misleading side chain density is affecting the tracing less.

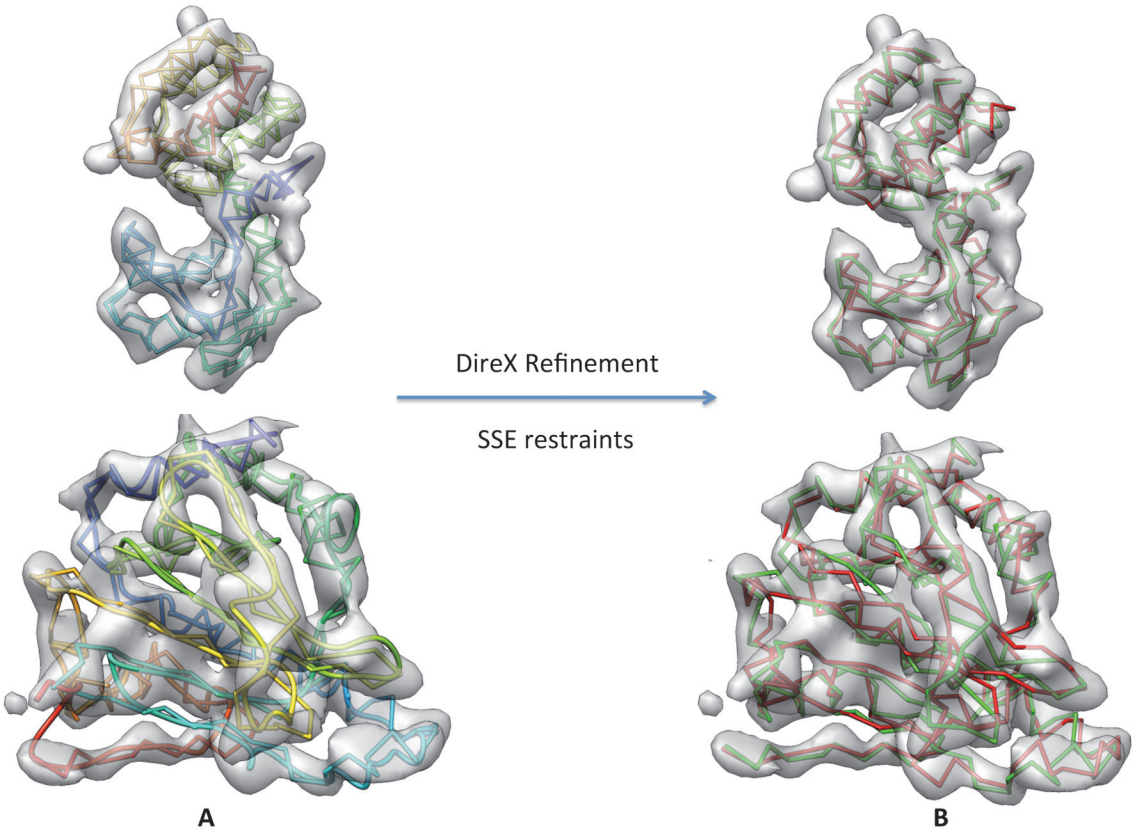

Figure 3.37: 201L and 3QDD tracing and refinement results. Panel A shows the tracing results of both models and colored according to the order of amino acids; Panel B shows the DireX refinement results and green colored the target backbone and red shows the refined model.

CHAPTER 4

Discussion

CryoEM has been widely used as a powerful tool to determine protein structures especially for large complexes. However, because protein complexes often show a significant conformational heterogeneity, the final resolution is limited. As the number of low resolution data is increasing dramatically, interpretation of these is still a considerable problem. Currently, most methods focus on rigidly or flexibly fitting the known crystal structure or homology models in the density maps. Nowadays, the challenge is how to model a protein structure from low-resolution data without any reference structures.

In this thesis work, we developed a method to build protein backbone models without knowing any three dimensional information for the low-resolution density maps. More specifically, the Lin-Kernighan heuristic algorithm, which is used for solving the Euclidean traveling salesman problem, is used as the main optimization algorithm. In addition, several modifications to the LKH algorithm were developed to increase the searching efficiency and a pseudo energy function which acts as a potential restraints is used to bias the searching process. At the end, from an ensemble of generated backbone traces, the best fitting ones are extracted.

To determine out the backbone from scratch in the density map, the most intuitive idea is to identify the characteristic atoms in the density map like Cα-atoms. Here, we proposed an automated way which is based on the sequence compositions to

determine a density threshold that captures the shape of the protein density. With the density threshold, pseudo Cα-atoms are placed on the basis of two criteria: distances and density fit. The placed pseudo Cα-atoms are viewed as a good descriptor of the overall shape of the density and the protein. However, obtaining accurate density threshold that define the proper region for protein volume is a subjective problem. Here we present a method that can automatically determine a density threshold by using the sequence information. The advantage of our method here is that it does not require any visual check and therefore can be used automatically in the program. In addition, by tuning the distance criterion, pseudo Cα-atoms can be flexibly located within a distance range. An optional map correlation refinement process can be used to improve the positions of the pseudo Cα-atoms. Although the pseudo Cα-atoms are not the same each time, they are still always similar and it would not affect the further investigation.

Given a density map and its corresponding pseudo Cα-atoms, how to make proper connections between the atoms is the key question. We proposed a method which is based on LKH that is used for solving the TSP. However, because the backbone tracing problem is not a TSP exactly, we added some modifications to the original algorithm (referred to as MLK), in particular the longest edge identification. Meanwhile, to increase the efficiency of the whole algorithm, Monte Carlo sampling and simulate annealing processes are combined with the LKH. For each of the optimization steps, random restart, structure heating and perturbation are also implemented to improve the efficiency.

A calmodulin structure (1S26:PDBID) which is an α and β mixed structure was chosen as a model to test the whole algorithm. In initial tests implemented with perfect Cα-atoms, the correct backbone structure can be determined easily. During the search for the correct trace, models might be trapped in local minima energy states. By adding the modifications, the starting atom and the ending atom can be identified directly. The MC-SA and other optimization steps dramatically increase the success rate of finding the correct trace. The rational idea behind these methods is to help the structures, which are trapped in local minima, to escape from these minima.

Although the algorithm works well with the perfect Cα-atoms, when it comes to the pseudo Cα-model which means the positions of the atoms are very likely staying at a inaccurate position, the results are not as accurate any more. The correct topology still can be determined but the local geometry is not as good as a normal trace. The reason for this is that the pseudo Cα atoms may not be well distributed. After searching, beads that close to each other between which there should not be any bond are connected. That destroy the local geometry and may even the whole topology.

To address that, we add other three groups of restraints. The first is the structure based secondary structure restraints which are generated by the predicted secondary structure information from the sequence. The second is the sequence based Miyazawa-Jernigan restraint. It is a statistic potential which is defined by the preferences of amino acids to be in contact. The last one is the density based restraint which enforces that connections between each pair of pseudo atoms lies within the density.

By using each of these restraints alone, none of them can yield a correct structure at the end. However, each of them has their own effect. The secondary structure restraints has the effect of localizing the secondary structure like pattern in the protein and the density restraints tend to make connections which stay inside the density. The MJ restraint is quite depending on the sequence assignment. It prefers local connections over longer ones but different from density restraints the connections may intersect with the density. During the optimization, the LK energy is the dominant "force" to make the change, its energy term also plays the main role in the whole energy function. To get a good result, the weighting factor for LK energy is always large and the weight for the density energy is similar, but the SSE energy and the MJ energy factors should have small values. Different restraints have different effects on the models and they bias the searching procedure in different ways. The resulting models therefore depend on the type of restraints used. Even though the restraints improve the tracing result, the running time with all energy terms relatively long. As in each optimization step, the sequence needs to be reassigned twice to the trace one from the starting to the end and also

in the reverse way, because which end of the trace is the first residue is unknown. In cases where the structure is highly symmetric, it would be hard to determine the right assignment of the first residue.

Even though the traces often have the correct topology, their local structure is typically not correct. Therefore structure refinement is needed. From the secondary structure prediction, a list of secondary structure distance restraints was generated and used as input for DireX refinement. With DireX refinement, the backbone traces were further improved and moved closer to the target structures.

Further more, 6 test cases were chosen with different second structure content: α-helix only, β-sheet only and α and β mixed groups as defined by CATH classification. Their corresponding density maps were simulated at 8, 4.5 and 6 Å resolution, respectively. For all the six targets, the correct backbone traces were found. In the α-helical tests, even small gaps can be correctly connected by our method. The most problematic test cases were those from the β-sheet group. The resolution to see well separated β-strands is around 4.5 Å. At such resolution, some beads might be placed and refined into side chain density. But beads located at the side chain density can easily end up being connected with beads belonging to the neighbouring β-strands, which causes intersecting connections and spoil the topology. Similarly, high density (large) side chains may pull beads from the main-chain density into the side chain area during the refinement.

After refinement, all the traces get better in secondary structure and move closer to the target structure. However, the final structures are slightly shifted with respect to the target in all test cases. As the distances restraints used for refinement are rather rigid, there can be easily a shift after the refinement if the beads do not register at the right positions.

Except for the calmodulin tests, all the other examples were tested with noise-free density maps. The logical next step we need to do is to test our method on noisy density maps or real experimental data. In addition, we could improve our bead generation method by better placing the beads, for example by identifying suppressing the side chain density, especially for the β-sheet structures.

By now, our optimal models were picked by calculating the RMSD to the known target structure. Although the total energy has a correlation with the RMSD value, sometimes the lowest energy structure does not have the smallest RMSD or is even far away from the best RMSD structure. Therefore, our hybrid energy value can not be used as a golden standard to evaluate the trace. A better method is required to validate the traces and to efficiently pick the optimal trace from the ensemble of generated traces.

After picking the best generated Cα backbone trace, the next step is to complete the atomic model by adding side-chains to the backbone trace. At low- or intermediate resolution (worse than 4 Å) this task is still quite challenging and beyond the scope of this thesis. If the quality of the backbone structure is very high, side chains can be placed quite reliably. However, backbone traces obtained from low-resolution density maps, as discussed in this thesis, are typically relatively inaccurate, which makes side-chain placement a big challenge.

APPENDIX A

Reference Distances of Cα Atoms

Table A.1: Reference Structure Distances of Cα atoms in α-helix

α-helix										
1	0.00	3.75	5.36	5.02	6.11	8.53	9.75	10.43	12.18	14.09
2	15.15	16.32	18.19	19.77	20.85	22.33	24.13	25.49	26.70	28.36
3	30.01	31.27	32.65	34.35	35.84	37.11	38.64	40.30	41.67	43.06
4	44.64	46.20	47.52	48.97	50.61	52.07	53.41	54.95	56.55	57.93
5	59.33	60.93	62.45	63.81	65.29	66.89	68.33	69.72	71.27	72.82

β-sheet					
1	0.00	3.75	6.47	9.89	12.94
2	16.28	19.40	22.72	25.87	29.17
3	32.34	35.62	38.81	42.09	45.28
4	48.55	51.74	55.01	58.21	61.48

APPENDIX B

Trilinear Interpolation

Trilinear interpolation is a method based on linear interpolation but extended to 3D data. In the three dimensional space, trilinear interpolation approximates the value of an intermediate point (x, y, z) which stay inside the local cubic. It is frequently used in computer graphics, numerical analysis especially when it has a volumertric dataset.

The value at each vertex can be denoted by $V_{000}, V_{001}, V_{010} \ldots V_{111}$. The value at a certain position (x, y, z) within the cube will be denoted V_{xyz} and can be calculated by:

$$\begin{aligned} V_{xyz} = {} & V_{000}(1-x)(1-y)(1-z) \\ & + V_{100}x(1-y)(1-z) \\ & + V_{010}(1-x)y(1-z) \\ & + V_{001}(1-x)(1-y)z \\ & + V_{101}x(1-y)z \\ & + V_{011}(1-x)yz \\ & + V_{110}xy(1-z) \\ & + V_{111}xyz \end{aligned}$$

APPENDIX C

Calmodulin Sequence in Fasta Format

>Calmodulin.pdb (#0) chain D/5-147

```
TEEQIAEFKE  AFSLFDKDGD  GTITTKELGT  VMRSLGQNPT
EAELQDMINE  VDADGNGTID  FPEFLTMMAR  KMKDTDSEEE
IREAFRVFDK  DGNGYISAAE  LRHVMTNLGE  KLTDEEVDEM
IREADIDGDG  QVNYEEFVQM  MTA
```

APPENDIX D

SSE Prediction of Calmodulin

Our SSE information is give by an ASCII text file, with the number of lines equal to the number of amino acids. Each line defines in which kind of secondary structure the correspondent atom is. The possible value are H (helix), E (beta strand), C (coil) and N (not specified)
Here to make it readable, we simplify the line by line form into one line format.

SSE:

```
CHHHHHHHHHHHCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCHHHHHHHHHH
HHCCCCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHHHHCCCCCCCCCEEEHHH
HHHHHHHCCCCCCHHHHHHHHHHHHCCCCCCEEEHHHHHHHHHCC
```

APPENDIX E

Secondary Structure Restraints File Template

With the predicated secondary structure result, the distance between each pair of atoms within the secondary structure elements will be wrote to a secondary structure restraint file. The file contents are exemplified as the following template:

566
1 2 3.80 1.0 1.0
2 3 3.80 1.0 1.0
3 4 3.80 1.0 1.0
4 5 3.80 1.0 1.0
5 6 3.80 1.0 1.0
6 7 3.80 1.0 1.0
7 8 3.80 1.0 1.0

First line gives the number of restraints in total of this file.
column 1: first atom id
column 2: second atom id
column 3: defined distance in Å
column 4: weight of restraint, that will be multplied with a strength factor
column 5: No meaning yet, but have to be present

APPENDIX F

Sequences for Further Tests

α-helical group:

>1AEP

NIAEAVQQLN HTIVNAAHEL HETLGLPTPD EALNLLTEQA
NAFKTKIAEV TTSLKQEAEK HQGSVAEQLN AFARNLNNSI
HDAATSLNLQ DQLNSLQSAL TNVGHQWQDI ATKTQASAQE
AWAPVQSALQ EAAEKTKEAA ANLQNSIQSA VQK

>4GOW

QLTEEQIAEF KEAFSLFDKD GDGTITTKEL GTVMRSLGQN
PTEAELQDMI NEVDADGNGT IDFPEFLTMM ARKMKDSEEE
IREAFRVFDK DGNGYISAAE LRHVMTNLGE KLTDEEVDEM
IREADIDGDG QVNYEEFVQM MT

β-sheet group:

>1DC9

DNLKLTITQE GNKFTVKESS NFRNIDNVFE LGVDFAYSLA
DGTELTGTWT MEGNKLVGKF KRVDNGKELI AVREISGNEL
IQTYTYEGVE AKRIFKKE

>3EMM

GTWRGQGEGE YPTIPSFRYG EEIRFSHSGK PVIAYTQKTW
KLESGAPMHA ESGYFRPRPD GSIEVVIAQS TGLVEVQKGT
YNVDEQSIKL KSDLVGNASK VKEISREFEL VDGKLSYVVR
MSTTTNPLQP HLKAILDKL

α-helix and β-sheet mixed group:

>201L

MNIFEMLRID EGLRLKIYKD TEGYYTIGIG HLLTKSPSLN
AAKSELDKAI GRNTNGVITK DEAEKLFNQD VDAAVRGILR
NAKLKPVYDS LDAVRRAALI NMVFQMGETG VAGFTNSLRM
LQQKRWDEAA VNLAKSRWYN QTPNRAKRVI TTFRT

>3QDD

QAEIAQLMSL IINTFYSNKE IFLRELISNS SDALDKIRYE
SLTDPSKLDS GKELHINLIP NKQDRTLTIV DTGIGMTKAD
LINNLGTIAK SGTKAFMEAL QAGADISMIG QFGVGFYSAY
LVAEKVTVIT KHNDDEQYAW ESSAGGSFTV RTDTGEPMGR
GTKVILHLKE DQTEYLEERR IKEIVKKHSQ FIGYPITLFV
EKE

APPENDIX G

SSE Prediction Results for Further Tests

The format is the same as we introduced for the calmodulin. To make it readable, we compressed multiple lines into one line.

α-helical group:

1AEP:

```
NHHHHHHHHHHHHHHHHHHHHCCCCCCCCNTTHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHCCCCCHHHHHHHHHHHHHHHHCCCCCCNNCCHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHC
CCC
```

4GOW:

```
NNNHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHHHCCNNNNHHHHHHHH
HHHNCCNCCNEEHHHHHHHHHHCNNNNNHHHHHHHHHHNCCNCCNENHHH
HHHHHHHHCNNNNHHHHHHHHHHNCCNCCCENHHHHHHHHN
```

β-sheet group:

1DC9:

```
NNNEEEEEEECCEEEEEEECCCCEEEEEEECCCCEEEECCCCNEEEEEEE
EECCEEEEEEEECCCNNEEEEEEEEECCEEEEEEECCEEEEEEENN
```

3EMM:

```
NEEEEEEEEECCEEEEEEEEEEEENNCCCNEEEEEEEENCCCNNEEEE
EEEEEEENCCCEEEEEEEECCCNEEEEEEEEECCCCEEEEEEEEECNCC
EEEEEEEEEECCEEEEEEEEECCCNCEEEEEEEEEN
```

α-helix and β-sheet mixed group:

201L:

```
NNHHHHHHHHHNNEEEEEENCCNNEEEECCEEEECNCNHHHHHHHHHNNN
NCNNCCENNHHHHHHHHHHHHHHHHHHHHNCCCHHHHHHCNHHHHHHHH
HHHHHHHHHHHHCNHHHHHHHHCCNHHHHHHHHCCCHHHHHCHHHHHHH
HHHHN
```

3QDD:

```
NHHHHHHHHHHHHCNNCNCCHHHHHHHHHHHHHHHHHHHHHHNHHHCC
NNNNNEEEEEECCCCEEEEEENCNNNNHHHHHHHHHHHHHHHHHHHHHH
HCCNNHHHHHHCNHHHHHHHCEEEEEEEEENCCCNEEEEEENCCCNNEE
EENNCCNNCCEEEEEEEENHHHHHHHCHHHHHHHHHHNCCNCCNEEENN
NNN
```

Bibliography

[1] M Magrane and U Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 72:bar009, 2011.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N.Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[3] Stretton AO Finch JT, Klug A. A simple extended-cavity diode laser. *Journal of Molecular Biology*, 10:570–575, Dec 1964.

[4] Wuthrich.K. Wagner.G. Dynamic model of globular protein conformations based on NMR studies in solution. *Nature*, 275(5677):247–248, Sep 1978.

[5] Ernst.R.R Wüthricha.K, Nagayamaa.K. Emerging techniques two-dimensional nmr spectroscopy. *Trends in Biochemical Sciences*, 4(8):N178–N181, Aug 1979.

[6] J. Frank. A study of heavy-light atom discrimination in bright-field electron microscopy using the computer.

[7] B. V. Prasad, J. W. Burns, E. Marietta, M. K. Estes, and W. Chiu. Localization of VP4 neutralization sites in rotavirus by three-dimensional cryo-electron microscopy. *Nature*, 343(6257):476–479, Feb 1990.

[8] Sim G.A Cheung K.K. Aflatoxin g1: Direct determination of the structure by the method of isomorphous replacement. *Nature*, 201:1185–1188, Mar 1964.

[9] Vallee B.L Ulmer D.D. Anomalous rotatory dispersion of enzyme-chelate complexes. i. alcohol dehydrogenase. *J. Biol. Chem*, 236(12):730–734, Mar 1961.

[10] Blow D.M. Rossmann M.G. The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst*, 15:24–31, Jan 1962.

[11] M. Z. Haque. A computational study of convoluted back projection algorithm. *Comput. Biol. Med*, 21(5):289–294, Jun 1991.

[12] M. Tagari, R. Newman, M. Chagoyen, J. M. Carazo, and K. Henrick. New electron microscopy database and deposition system. *Trends Biochem. Sci*, 27(11):589, Nov 2002.

[13] A. E. Leschziner and E. Nogales. Visualizing flexibility at molecular resolution: analysis of heterogeneity in single-particle electron microscopy reconstructions. *Annu Rev Biophys Biomol Struct*, 36(10):43–62, Jun 2007.

[14] D. A. Jaffray, D. G. Drake, M. Moreau, A. A. Martinez, and J. W. Wong. A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. *Int. J. Radiat. Oncol. Biol. Phys.*, 45(3):773–789, Oct 1999.

[15] M. van Heel and M. Schatz. Fourier shell correlation threshold criteria. *J. Struct. Biol.*, 151(3):250–262, Sep 2005.

[16] M. Van Heel G. Harauz. Exact filters for general geometry three dimensional reconstruction. *Optik*, 73(4):146–156, Dec 1986.

[17] W. Jiang and S. J. Ludtke. Electron cryomicroscopy of single particles at subnanometer resolution. *Curr. Opin. Struct. Biol.*, 15(5):571–577, Oct 2005.

[18] W. Chiu, M. L. Baker, W. Jiang, M. Dougherty, and M. F. Schmid. Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure*, 13(3):363–372, Mar 2005.

[19] J.Frank J.Fu, H. Gao. Unsupervised classification of single particles by cluster tracking in multi-dimensional space. *Journal of structural biology*, 157 (1):226–239, Jan 2007.

[20] J. H. Cate, M. M. Yusupov, G. Z. Yusupova, T. N. Earnest, and H. F. Noller. X-ray crystal structures of 70S ribosome functional complexes. *Science*, 285 (5436):2095–2104, Sep 1999.

[21] N. Eswar, D. Eramian, B. Webb, M. Y. Shen, and A. Sali. Protein structure modeling with MODELLER. *Methods Mol. Biol.*, 426:145–159, 2008.

[22] M. G. Rossmann, R. Bernal, and S. V. Pletnev. Combining electron microscopic with x-ray crystallographic structures. *J. Struct. Biol.*, 136(3): 190–200, Dec 2001.

[23] M. G. Rossmann. Fitting atomic models into electron-microscopy maps. *Acta Crystallogr. D Biol. Crystallogr.*, 56(Pt 10):1341–1349, Oct 2000.

[24] W. Wriggers. Using Situs for the integration of multi-resolution structures. *Biophys Rev*, 2(1):21–27, Feb 2010.

[25] W. Wriggers, R. A. Milligan, K. Schulten, and J. A. McCammon. Self-organizing neural networks bridge the biomolecular resolution gap. *J. Mol. Biol.*, 284(5):1247–1254, Dec 1998.

[26] T. Kawabata. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys. J.*, 95 (10):4643–4658, Nov 2008.

[27] K. Lasker, M. Topf, A. Sali, and H. J. Wolfson. Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J. Mol. Biol*, 388(1):180–194, Apr 2009.

[28] K. Lasker, A. Sali, and H. J. Wolfson. Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins*, 78(15):3205–3211, Nov 2010.

[29] L. G. Trabuco, E. Villa, K. Mitra, J. Frank, and K. Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16(5):673–683, May 2008.

[30] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, 49(2):174–180, Oct 2009.

[31] Z. Chen, E. Blanc, and M. S. Chapman. Real-space molecular-dynamics structure refinement. *Acta Crystallogr. D Biol. Crystallogr.*, 55(Pt 2):464–468, Feb 1999.

[32] J. Z. Chen, J. Furst, M. S. Chapman, and N. Grigorieff. Low-resolution structure refinement in electron microscopy. *J. Struct. Biol.*, 144(1-2):144–151, 2003.

[33] A. P. Pandurangan and M. Topf. Protein structure fitting and refinement guided by cryo-EM density. *Structure*, 16(2):295–307, Feb 2008.

[34] A. P. Pandurangan and M. Topf. RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. *Bioinformatics*, 28(18):2391–2393, Sep 2012.

[35] F. Tama, O. Miyashita, and C. L. Brooks. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.*, 337(4):985–999, Apr 2004.

[36] F. Tama, O. Miyashita, and C. L. Brooks. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, 147(3):315–326, Sep 2004.

[37] K. Suhre, J. Navaza, and Y. H. Sanejouand. NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density map. *Acta Crystallogr. D Biol. Crystallogr.*, 62 (Pt 9):1098–1100, Sep 2006.

[38] L. Skjaerven, S.M. Hollup, and N. Reuter. Normal mode analysis for proteins. *J. Mol. Struct.*, 898(Pt 9):42–48, Sep 2009.

[39] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3):417–429, Nov 1998.

[40] K. Suhre and Y. H. Sanejouand. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, 32(Web Server issue):W610–614, Jul 2004.

[41] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2 (3):173–181, 1997.

[42] I. Bahar, T. R. Lezon, L. W. Yang, and E. Eyal. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys*, 39:23–42, 2010.

[43] K. Kanou, M. Iwadate, T. Hirata, G. Terashi, H. Umeyama, and M. Takeda-Shitaka. FAMSD: A powerful protein modeling platform that combines alignment methods, homology modeling, 3D structure quality estimation and molecular dynamics. *Chem. Pharm. Bull.*, 57:1335–1342, Dec 2009.

[44] J. A. Velazquez-Muriel and J. M. Carazo. Flexible fitting in 3D-EM with incomplete data on superfamily variability. *J. Struct. Biol.*, 158(2):165–181, May 2007.

[45] A.S. Dore, N. Furnham, O.R. Davies, B.L. Sibanda, D.Y. Chirgadze, S.P. Jackson, L. Pellegrini, and T.L. Blundell. Structure of an Xrcc4-DNA ligase IV yeast ortholog complex reveals a novel BRCT interaction mode. *DNA Repair(Amst)*, 5(3):362–368, Mar 2006.

[46] N. Furnham, A. S. Dore, D. Y. Chirgadze, P. I. de Bakker, M. A. Depristo, and T. L. Blundell. Knowledge-based real-space explorations for low-resolution structure determination. *Structure*, 14(8):1313–1320, Aug 2006.

[47] G. F. Schroder, A. T. Brunger, and M. Levitt. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*, 15(12):1630–1641, Dec 2007.

[48] Z. Wang and G. F. Schroder. Real-space refinement with DireX: from global fitting to side-chain improvements. *Biopolymers*, 97(9):687–697, Sep 2012.

[49] G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc*, 3(7):1171–1179, 2008.

[50] A. D. Ferguson, B. M. McKeever, S. Xu, D. Wisniewski, D. K. Miller, T. T. Yamin, R. H. Spencer, L. Chu, F. Ujjainwalla, B. R. Cunningham, J. F. Evans, and J. W. Becker. Crystal structure of inhibitor-bound human 5-lipoxygenase-activating protein. *Science*, 317(5837):510–512, Jul 2007.

[51] J. A. Velazquez-Muriel and J. M. Carazo. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.*, 62(Pt 9):1002–1011, Sep 2006.

[52] T. C. Terwilliger. SOLVE and RESOLVE: automated structure solution and density modification. *Meth. Enzymol.*, 374:22–37, 2003.

[53] T. R. Ioerger and J. C. Sacchettini. TEXTAL system: artificial intelligence techniques for automated protein model building. *Meth. Enzymol*, 374:244–270, 2003.

[54] M. L. Baker, T. Ju, and W. Chiu. Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, 15(1):7–19, Jan 2007.

[55] D. Si, S. Ji, K. A. Nasr, and J. He. A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps. *Biopolymers*, 97(9):698–708, Sep 2012.

[56] D. Si and J. He. Beta-sheet detection and representation from me- dium resolution cryo-EM density maps. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 764, 2013.

[57] M. Rusu and W. Wriggers. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. *J. Struct. Biol.*, 177(2):410–419, Feb 2012.

[58] K. Al Nasr, W. Sun, and J. He. Structure prediction for the helical skeletons detected from the low resolution protein density map. *BMC Bioinformatics*, 11 Supp1 1, S44 2010.

[59] A. Biswas, D. Si, K. Al Nasr, D. Ranjan, M. Zubair, and J. He. Improved efficiency in cryo-EM secondary structure topology determination from inaccurate data. *J Bioinform Comput Biol*, 10(3):1232006, Jun 2012.

[60] S. Lindert, N. Alexander, N. Wotzel, M. Karaka?, P. L. Stewart, and J. Meiler. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure*, 20(3):464–478, Mar 2012.

[61] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovi?, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Meth. Enzymol.*, 2487:545–574, 2011.

[62] M. L. Baker, S. S. Abeysinghe, S. Schuh, R. A. Coleman, A. Abrams, M. P. Marsh, C. F. Hryc, T. Ruths, W. Chiu, and T. Ju. Modeling protein structure at near atomic resolutions with Gorgon. *J. Struct. Biol.*, 174(2):360–373, May 2011.

[63] M.R. Baker, I. Rees, S.J. Ludtke, W. Chiu, and M. L. Baker. Constructing and validating initial C-alpha models from subnanometer resolution density maps with pathwalking. *Structure*, 20(3):450–463, Mar 2012.

[64] S. J. Ludtke, P. R. Baldwin, and W. Chiu. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.*, 128(1): 82–97, Dec 1999.

[65] Shen Lin and Brian W. Kernighan. An effective heuristic algorithm for the travelling-salesman problem. *Operations Research*, 21(2):498–516, 1973.

[66] S.Lin. Computer solutions of the traveling-salesman problems. *BSTJ*, 44: 2245–2269, 1965.

[67] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[68] S. Kirkpatrick; C. D. Gelatt; M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[69] H.W.Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[70] H.Emmons;K.Matur. Lot sizing in a no-wait flow shop. *Operations Research Letters*, 17(4):159–164, 1995.

[71] L.J. Hubert;F.B. Baker. Applications of combinatorial programming to data analysis: the traveling salesman and related problem. *Psychomerika*, 43(1): 81–91, 1978.

[72] R.G. Bland;D.F. Shallcross. Large travelling salesman problems arising from experiments in x-ray crystallography: A preliminary report on computation. *Operations Research Letters*, 8(3):125–128, 1989.

[73] C.Korostensky;G.H.Gonnet. Near optimal multiple sequence alignments using a tsp approach. pages 105–114, 1999.

[74] Gregory Gutin, Abraham Punnen, Alexander Barvinok, Edward Kh. Gimadi, and Anatoliy I. Serdyukov. The traveling salesman problem and its variations. *Kluwer*, 2002.

[75] Gerard Sierksma. Hamiltonicity and the 3-opt procedure for the traveling salesman problem. *Applicationes Mathematicae*, 22(3):351–358, 1994.

[76] A.J Morton K.T Mak. A modified lin-kernighan traveling-salesman heuristic. *Operations Research Letter*, 13(3):127–132, 1993.

[77] S.Kirkpatrick, C.D Gelatt Jr., and M.P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[78] Z. Wang and G. F. Schroder. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4):404–405, Apr 2000.

[79] Kevin Karplus, Christian Barrett, and Richard Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10): 846–856, 1998.

[80] Kevin Karplus, Rachel Karchin, Christian Barrett, Spencer Tu, Melissa Cline, Mark Diekhans, Leslie Grate, Jonathan Casper, and Richard Hughey. What is the value added by human intervention in protein structure prediction? *Proteins: Structure, Function, and Genetics*, 45(S5):86–91, 2001.

[81] G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, Apr 2005.

[82] M. Ouali and R. D. King. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, 9(6):1162–1176, Jun 2000.

[83] R. Adamczak, A. Porollo, and J. Meller. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, 56(4):753–767, Sep 2004.

[84] M. N. Nguyen, K. P. Tan, and M. S. Madhusudhan. CLICK-Topology independent comparison of biomolecular 3d structures. *Nucleic Acids Res.*, 39(Web Server issue):W24–28, 2011.

[85] Y. Harpaz, M. Gerstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2(7):641–649, Jul 1994.

[86] W. Wriggers and S. Birmanns. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol.*, 133(2-3):193–202, 2001.

[87] F. Alber, M. F. Kim, and A. Sali. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure*, 13(3):435–445, Mar 2005.

[88] Andersson K. M. and Hovmoeller S. The average atomic volume and density of proteins. *Z.Kristallogr*, 213:369–373, 1998.

[89] J. Tsai, R. Taylor, C. Chothia, and M. Gerstein. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.*, 290(1):253–266, Jul 1999.

[90] M. L. Quillin and B. W. Matthews. Accurate calculation of the density of proteins. *Acta Crystallogr. D Biol. Crystallogr.*, 56(Pt 7):791–794, Jul 2000.

[91] P. G. Squire and M. E. Himmel. Hydrodynamics and protein hydration. *Arch. Biochem. Biophys.*, 196(1):165–177, Aug 1979.

[92] Craig E.Kundrot Frederic M.Richards. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *PROTEINS*, 220(2):71–84, 1983.

[93] R.L.Jernigan S.Miyazawa. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[94] R.L.Jernigan S.Miyazawa. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3):623–644, 1996.

[95] R.L.Jernigan S.Miyazawa. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, 34(1):49–68, 1999.

[96] Thomas S. Helena R. L., Olivier C. M. Iterated local search. *International Series in Operations Research and Management Science*, 57:320–353, 2003.

[97] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[98] C. Cole, J. D. Barber, and G. J. Barton. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res*, 36(Web Server issue):197–201, Jul 2008.

[99] D Frishman and P Argos. Incorporation of long-distance interactions into a secondary structure prediction algorithm. *Protein Engineering*, 9:133–142, 1996.

[100] G Karypis. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins*, 36(3):575–586, Aug 2006.

[101] I. Sillitoe, A. L. Cuff, B. H. Dessailly, N. L. Dawson, N. Furnham, D. Lee, J. G. Lees, T. E. Lewis, R. A. Studer, R. Rentzsch, C. Yeats, J. M. Thornton, and C. A. Orengo. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, 41(Database issue):D490–498, Jan 2013.

[102] D. R. Breiter, M. R. Kanost, M. M. Benning, G. Wesenberg, J. H. Law, M. A. Wells, I. Rayment, and H. M. Holden. Molecular structure of an apolipoprotein determined at 2.5-A resolution. *Biochemistry*, 30(3):603–608, Jan 1991.

[103] Q. Xu, A. Chang, A. Tolia, and D. L. Minor. Structure of a Ca(2+)/CaM:Kv7.4 (KCNQ4) B-helix complex provides insight into M current modulation. *J. Mol. Biol.*, 425(2):378–394, Jan 2013.

[104] I. J. Ropson, B. C. Yowler, P. M. Dalessio, L. Banaszak, and J. Thompson. Properties and crystal structure of a beta-barrel folding mutant. *Biophys. J.*, 78(3):1551–1560, Mar 2000.

[105] C. M. Bianchetti, G. C. Blouin, E. Bitto, J. S. Olson, and G. N. Phillips. The structure and NO binding properties of the nitrophorin-like heme-binding protein from Arabidopsis thaliana gene locus At1g79260.1. *Proteins*, 78(4): 917–931, Mar 2010.

[106] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, Oct 2004.

[107] Heinz D.W., Baase W.A., Dahlquist F.W., and Matthews B.W. How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. *Nature*, 11(361): 561–564, Feb 1993.

[108] J. Shi, R. Van de Water, K. Hong, R. B. Lamer, K. W. Weichert, C. M. Sandoval, S. R. Kasibhatla, M. F. Boehm, J. Chao, K. Lundgren, N. Timple, R. Lough, G. Ibanez, C. Boykin, F. J. Burrows, M. R. Kehry, T. J. Yun, E. K. Harning, C. Ambrose, J. Thompson, S. A. Bixler, A. Dunah, P. Snodgrass-Belt, J. Arndt, I. J. Enyedy, P. Li, V. S. Hong, A. McKenzie, and M. A. Biamonte. EC144 is a potent inhibitor of the heat shock protein 90. *Nature*, 55(17):7786–7795, Sep 2012.

[109] V. Tugarinov, W. Y. Choy, V. Y. Orekhov, and L. E. Kay. Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc. Natl. Acad. Sci. U.S.A*, 102(3):622–627, Jan 2005.