ELIGN : Elastic alignment of cryo-EM density maps

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Amudha Kumari Duraisamy
aus Salem, India

Jülich, April 2017

aus dem Institute of Complex Systems, Strukturbiochemie (ICS-6)
des Forschungszentrums Jülich

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Jun. Prof. Dr. Gunnar Schröder
Korreferent:	Prof. Dr. Dieter Willbold
Tag der mündlichen Prüfung:	Juli 4, 2017

Declaration of Authorship

I, Amudha Kumari Duraisamy, declare that this thesis 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: _____

Abstract

Knowing the structure of biomolecular complexes, especially proteins, is an important prerequisite for understanding their function and provides much more information than the genetic sequence alone. Cryo-electron microscopy (cryo-EM) is a powerful technique to study the structure of biomolecular assemblies which are often large, flexible and conformationally heterogeneous.

This thesis describes three methods developed for improving various data processing steps used for structure determination in cryo-EM. All the three are post-processing methods that can be used after 3D reconstruction. The methods are used to enhance the resolution of density maps, sharpen the density maps and to determine the atomic coordinate precision during refinement of a structure against a density map, respectively.

(1) ELIGN: Cryo-imaging a sample in its native physiological environment contains conformational heterogeneity. The conformational heterogeneity in the sample can severely limit the resolution of density reconstructions. The conformational motions lead to global structural changes, however proteins and protein complexes often contain rather rigid domains. Those rigid domains can be used to align and average the density to reach higher resolution. We have developed a new algorithm to elastically align the density of one conformation to another conformation. By well superimposing the individual domains, different density maps can be aligned better, thus allowing for averaging maps of different conformational states. Therefore with ELIGN, all the available images can be combined in a single density map with the prospect of significantly improving the resolution and overcoming the limitation typically imposed by conformational heterogeneity.

(2) VISDEM: With the tremendous growth of the cryo-EM field in the recent years, there is a need for optimal visualization tool. It is needed both for proper interpretation of the density map and also for atomic model building in case of high resolution data. However, the reconstructed density maps always have artefacts,

both due to the electron optics and the data processing techniques used. We have developed a method, that uses knowledge about the general statistics of protein structures, which is used as a restraint to sharpen density maps.

(3) Atomic positional precision: A high resolution density map is needed to determine the structure very accurately. The typically reachable resolution in cryo-EM is still in the low- to intermediate range (3.5–6 Å). Thus the structure determined, even by the refinement of a known model against cryo-EM map will contain errors. Estimating the error is a must to evaluate the quality and correctness of the structure. Here, we have developed a method to estimate the coordinate error for atomic models obtained from such real-space refinement.

Zusammenfassung

Die Bestimmung der Struktur von biomolekularen Komplexen, besonders von Proteinen, ist eine wichtige Vorraussetzung, um deren Funktion zu verstehen und liefert viel mehr Information als deren genetische Sequenz allein. Kryo-Elektronenmikroskopie ist eine leistungsfähige Technik, um die Struktur von biomolekularen Komplexen zu bestimmen. Diese Komplexe sind oft sehr groSS, flexibel und zeigen zum Teil erhebliche konformationelle Heterogenität.

In dieser Arbeit werden drei Methoden beschrieben, die entwickelt wurden um verschiedene Schritte der Datenverarbeitung in der Strukturbestimmung mittels Kryo-Elektronenmikroskopie zu verbessern. Bei allen drei Methoden handelt es sich um Nachbearbeitungsmethoden, die nach der 3D-Rekonstruktion angewendet werden können. Die Methoden verbessern zum einen die Auflösung der Dichtekarten, schärfen die Dichte und können die Präzission der Atomkoordinaten während der Verfeinerung einer Struktur zur Dichtekarte bestimmen.

(1) ELIGN: Aufnahmen einer Proteinprobe in deren nativer physiologischer Umgebung können sehr unterschiedliche Konformationen des Proteins enthalten. Diese heterogenen Konformationen können die Auflösung der Dichterekonstruktion stark limitieren. Bei den Konformationsänderungen handelt es sich oft um globale strukturelle Änderungen von rigiden Domänen. Diese rigiden Domänen können deshalb aligniert und gemittelt werden, um eine Dichte mit höherer Auflösung zu bekommen. Durch das Alignement der einzelnen Domänen wird das Alignement verschiedener Dichten verbessert und man kann deshalb auch Dichten verschiedener Konformationszustände mitteln. ELIGN ermöglicht es deshalb, alle Aufnahmen für eine einzelne Dichte zu verwenden und dadurch die Auflösung zu verbessern.

(2) VISDEM: Durch das enorm gestiegene Interesse an der Elektronenmikroskopie in den letzten Jahren sind optimale Visualisierungstools stark gefragt. Sie werden zum einen gebraucht, um die Dichtekarten richtig zu interpretieren und um atomare Modelle bauen zu können, wenn die Auflösung hoch genug ist. Die rekonstruierten Dichtekarten weisen jedoch aufgrund der Datenverarbeitungsschritte und der Elektronenoptik Artefakte auf. Wir haben eine Methode entwickelt, welche die

allgemeingültige Statistiken zur Struktur von Proteinen nutzt, um Dichtekarten zu schärfen.

(3) Atomkoordinatengenauigkeit: Hochaufgelöste Dichtekarten werden gebraucht um die Struktur eines Proteins sehr genau zu bestimmen. Die durchschnittliche Auflösung der Kryoelektronenmikroskopie ist jedoch noch immer im niedrigen und mittlerem Auflösungsbereich (3.5– 6 Å). Die bestimmten Strukturen enthalten deshalb Fehler. Das gilt auch für die Verfeinerung bekannter Strukturen in Dichtekarten. Um die Qualität und Richtigkeit eines Models zu bestimmen, ist es deshalb nötig diese Fehler abzuschätzen. Wir haben deshalb eine Methode entwickelt, um den Koordinatenfehler von Modellen, die mittels Verfeinerung im Realraum generiert wurden, abzuschätzen.

Acknowledgement

I would like to acknowledge my supervisor Jun. Prof. Gunnar Schröder, for giving me the opportunity to work with him. I have to thank him for his guidance, support and giving me good space and freedom to carry out my research. Appreciably, his doors were always open to carry out friendly discussions and his attitude gave a how-to-be-a-good-advisor/researcher inspiration and perception. I also wish to express my gratitude to Prof. Dieter Willbold for his support during my stay.

I would like to thank all my colleagues, particularly Ms. Tatjana Braun, Dr. Zhe Wang, Ms. Michaela Spiegel, Ms. Carla Schlenk, Ms. Lena Möhlenkamp, Dr. Andre Wildberg, Dr. Kumaran Baskaran and Dr. Benjamin Falkner for the good discussions and great help on many occasions. My acknowledgement also goes to Jun. Prof. Brigit Strodel and her group members for a great time in the department. I would also like to thank both my graduate schools, Biostruct - NRW research school and IHRS - Biosoft, for the scientific as well as the non-scientific lectures and courses offered during this time period. A special thanks to all my gurus, friends and family members.

Contents

List of Figures

List of Tables

Acronyms

XRC	**X**-**R**ay **C**rystallography
cryo-**EM**	cryo- **E**lectron **M**icroscopy
cryo-**ET**	cryo- **E**lectron **T**omography
SPA	**S**ingle **P**article **A**nalysis
SFX	**S**erial **F**emtosecond **X**ray crystallography
XFEL	**X**-ray **F**ree **E**lectron **L**aser
CCD	**C**harged **C**oupled **D**evice
DDD	**D**irect **D**etection **D**evice
MTF	**M**odulation **T**ransfer **F**unction
PSF	**P**oint **S**pread **F**unction
DQE	**D**etective **Q**uantum **E**fficiency
SNR	**S**ignal -to- **N**oise **R**atio
FRC	**F**ourier **R**ing **C**orrelation
FSC	**F**ourier **S**hell **C**orrelation
NMA	**N**ormal **M**ode **A**nalysis
DEN	**D**eformable **E**lastic **N**etwork
ENM	**E**lastic **N**etwork **M**odel
ELIGN	**EL**astic al**ING**ment
MD	**M**olecular **D**ynamics
MDFF	**M**olecular **D**ynamics **F**lexible **F**itting
RMSD	**R**oot **M**ean **S**quare **D**eviation
RMSF	**R**oot **M**ean **S**quare **F**luctuation

Dedicated to Mother Nature.

Chapter 1

Introduction

1.1 Motivation

Proteins are the major macromolecules which engage in almost every process within biological systems. The function of a protein is determined by its structure, which makes protein structure determination of utmost importance for understanding their mechanism of action and ultimately also for developing drugs. The most common structure determination methods available as of today are X-ray crystallography (XRC), nuclear magnetic resonance (NMR) and cryo electron microscopy (cryo-EM). Even though XRC had been proven as an excellent technique to reach high resolution, it has its own limitation that the proteins need to be crystallized. Protein crystallization works only for certain proteins and is often difficult to achieve, in particular, for biomolecular complexes. On the other hand, NMR can deal with proteins that cannot be crystallized but is limited to only small biomolecules. Cryo-EM, a relatively new technique, has the advantage that it can be applied to non-crystalline and large biomolecules in its near-native state. Cryo-EM is therefore a powerful tool to study macromolecular complexes that are often large and flexible and even allows visualizing the conformational heterogeneity. The quality of the structure determination in XRC and cryo-EM depends on the resolution of the density maps obtained from these experiments. To determine individual secondary structures like α-helices and β-sheets, the density map resolution should be better than 8 Å and 6 Å respectively. For side chains to be visible, the resolution should be or higher than 4.5 Å. Hence, the need for high resolution density maps is very vital in the field of structural biology. Oftentimes, this is not possible due to various limitations of the experimental techniques.

Despite the above mentioned advantage of cryo-EM that it fills the gap where XRC and NMR cannot be used, still the number of deposited high resolution density maps is very low. There are various limiting factors like low electron dose,

conformational variation and flexibility of the particles, beam-induced movement, thermal expansion of the grid, detection efficiency, non-optimal image processing, etc., which hinders the structure determination to reach the atomic resolution range in cryo-EM. The main limitation is the sensitivity of biomolecules for electron radiation, which requires to use very low electron doses for imaging to avoid excessive radiation damage. The resulting signal to noise ratio is therefore very low in the cryo-EM images, henceforth many thousands of images have to be aligned and averaged to improve the contrast. The quality of the final reconstructed 3D density map is defined by the quality of the images that goes into the image alignment procedure. Even though the recorded images are of low contrast, in principle, the resolution can be reached to the atomic level, if the images of the structures are aligned accurately [1].

Conformational motions and the ability to adopt a variety of conformations are important properties of proteins and key to performing essential biological functions. Conformational heterogeneity leads to very little or almost no density in the flexible regions and decreases the overall resolution of the reconstructed 3D density maps. By using, simply, all available 2D images in a dataset to reconstruct one single 3D density map, the resolution will not be optimal, since the average is a mixture of various functional states. Even though the field of cryo-EM comprising both single particle analysis (SPA) and cryo-electron tomography (cryo-ET) are best suited to study compositional and conformational heterogeneity, most of the density maps obtained from these experiments are limited in resolution by this conformational flexibility. Improving the resolution of the density maps, thus, requires to account for this structural heterogeneity. Averaging density maps is expected to lead to a better signal-to-noise ratio and therefore to reach higher resolution. However, higher resolution can only be achieved if the conformational differences between the different conformational classes are sufficiently small, or somehow accounted for.

Until now, the only way to handle the problem of conformational heterogeneity is by classifying or sorting the 2D images to similar conformational classes and reconstructing them into separate density maps. Several sorting algorithms have been developed over the years, which use methods like bootstrapping [2], Bayesian statistics [3] and multivariate statistical analysis (MSA) [4–6]. Image classification

is always a trade-off between the number of conformational states (classes) and the number of images per class that is used in the reconstruction of each conformation. Therefore, sorting will lead to a decrease in the number of images used in a reconstruction per class and thereby reduces the theoretically achievable resolution.

In this thesis, I present a new method to combine density maps of different classes. This method enables the flexible alignment of a density map into another density map. The density of one conformation into another conformation are flexibly fitted such that all individual domains are well superimposed, which means different density maps can be aligned better. The flexible density map alignment effectively reduces the variance between the image classes and allows for merging different classes to benefit from the higher number of images per reconstruction, in order to improve resolution. The elastic deformation idea can be applied either on the 2D images or on the 3D densities. If done on the 2D images, all the images are elastically bend to fit the projections of the required functional state and thereafter reconstructed into a single density for the specific state. If done on 3D volumes, the images have to be extensively sorted into various classes before reconstruction and the 3D densities are elastically aligned and averaged. As of today there exists no method in the field of electron microscopy to elastically fit a 2D image to another image or 3D density map to another density map.

More efficient and automatic algorithms need to be developed for analysing and processing the data to reach a high-resolution density maps from experimentally collected data. The main work presented in this thesis is on the elastic alignment of cryo-EM density maps. We have developed an efficient algorithm to elastically align and average well sorted 3D density maps of different conformations. The elastic averaging method can overcome the limitation on resolution due to conformation heterogeneity. The method can be used in both single particle analysis and in cryo-ET. Using the algorithm on cryo-ET data would be particularly advantageous because density maps obtained from cryo-ET typically are of low or intermediate resolution. The method was tested by elastically aligning 3D density maps of single particle analysis and cryo-ET. The algorithm bends the density map by treating it like a pseudo-atomic / bead model (point masses) and therefore allowing for a flexible deformation of the density grid to align with the other density

map. The method does not require any prior knowledge of the atomic structure, which makes the technique very versatile.

1.2 Structure determination

Structure determination of a protein, which is the building block of all living creatures, is an important prerequisite in understanding its function. Determining the structure and hence its function helps in drug designing. The first protein structure that was determined was that of the oxygen-storage myoglobin in 1958, at 6 Å resolution [7] however, improved to 2 Å quickly [8] followed by the oxygen-carrier hemoglobin at 5.5 Å [9] by John Kendrew and Max Perutz using X-ray crystallography for which they received the Nobel prize. After that, seven different protein structures were determined in the 1960s, including hen egg lysozyme [10], ribonucleases A and S [11, 12], chymotrpsin [13], papain [14], carboxypeptidase A [15] and subtilsin [16] as cited in [17], which pushed the field forward. As a consequence, the number of protein and nucleic acid structures determined increased steadily over the past five decades. Fig 1.1 shows the statistics of the structures released from 1968 to 2014 along with the milestone-discoveries. However there are only 121,958 structures solved as of today which is only 0.1% of known genetic sequences, which underlines the importance of structural biology.

1.2.1 X-ray Crystallography

X-ray crystallography is a well established technique and more than 89% of the protein structures were solved by this method. The highly concentrated purified sample first has to be crystallized. The crystal is placed in an X-ray beam and the electrons in the crystal scatter the incident X-ray beam. The diffracted X-ray from the crystal produces a specific pattern depending on the lattice structure, which enables to reconstruct the 3D structure of the molecules in the crystal. The electron

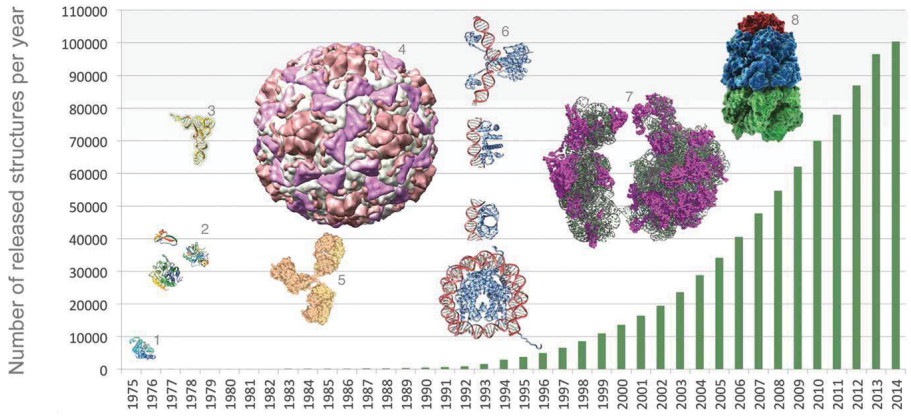

Figure 1.1 The image shows the number of structures per year until 2014 also including the important discoveries. (1) myoglobin, the first structure solved by X-ray crystallography, (2) small enzymes, (3) tRNA, (4) virus, (5) antibodies, (6) protein-DNA complexes, (7) ribosomes and (8) chaperones (taken from wwwPDB.org).

density ρ can be obtained from the diffracted wave using the equation 1.1.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|\exp[2\pi\imath(hx+ky+lz)+\imath\Phi(hkl)] \tag{1.1}$$

In the diffracted pattern the amplitude information, $I = F.F^*$, is recorded whereas the phase information Φ is lost. However, both the information are needed to reconstruct the real space electron density by the Fourier transformation technique. There are many methods available to solve the phase problem like molecular replacement [18], heavy atom labelling etc. Once the 3D electron density map is reconstructed there are number of programs available like Phenix [19] for automatic and Coot [20] for manual model building to determine the structure thereafter. However, the model building will not end in perfect model, therefore refinement [21, 22] has to be done. During the refinement, which is an iterative process, the atomic model is successively improved, which in turn will improve the phase information and enhances the clarity of the electron density map which can be used for better model building, and so on.

The resolving power of a well ordered crystal is very high, since the typical wavelength used is 1 Å. The technique allows for studying small amplitude motions of the protein within the crystal lattice, but not large-scale motions, which are often functionally important. Even though some proteins can be crystallized in different conformations, it is typically not possible to determine all conformations that are accessible to the protein. Moreover, it is often difficult to obtain a well ordered 3D crystal for systems like protein complexes and membrane proteins, which makes it difficult to study these systems by this method. Despite these limitations, X-ray crystallography is, still, the most frequently used method in structural biology.

A new direction in the field of structural study is serial femtosecond crystallography (SFX) [23, 24], which is used to study protein micro- and nano-crystals, which are much easier to obtain than large crystals. Furthermore, imaging of single particles such as viruses [25] and single cells [26] are possible. The field of SFX emerged with the development of the X-ray Free Electron Laser (XFEL) [27–30]. SFX is used for small crystals and its a technique that suitably fills the gap between XRC and serial femtosecond coherent diffractive imaging for non-crystalline samples. It is a destructive technique where the sample is detected very fast by a femtosecond pulse, even before the radiation damage deteriorates the signal.

1.2.2 Electron Microscopy

Electron microscopy (EM) is an important tool in the expanding field of structure determination. The first electron microscope, built by Ernst Ruska in 1933, was capable of resolving to $50nm$; Ruska received the Nobel prize 50 years later, for his contribution. Unlike the optical microscope, where light is focused by optical lenses, electron beams are focused using electromagnetic fields produced by focusing solenoid in an electron microscope. Electrons are affected by the electromagnetic fields and changing the current in the focusing coils affect the motion of the electrons, which can easily be controlled.

There are two kinds of operating modes of an electron microscope: transmission electron microscopy (TEM) and scanning electron microscopy (SEM). If the biology specimen is sufficiently thin, then the electron beam can be transmitted through the specimen and this kind of imaging is called TEM [31, 32]. Whereas, if the sample is too thick to transmit the electrons, in that case the secondary elec-

trons produced from the surface due to the excitation by primary electron beam is imaged. Therefore only the surface properties can be studied on such samples and this technique is known as SEM [33]. TEM needs hundreds of kV accelerating electrons to be fully transmitted by the sample, whereas SEM needs only few tens of kV acceleration. SEM yields a 3D image due to its varying depth in the electron penetration on the surface, unlike TEM which yields 2D projection images of the object. Resolution in EM is limited by spherical and chromatic aberration and development in the field has emerged with aberration correctors [34–36]. In EM, the sample is kept under ultra high vacuum condition leading to boiling of water in the sample. Since, there is a large pressure difference between the sample and the chamber, this could lead to an explosion in the sample. Therefore, the sample has to be prepared by either dehydration, fixation, embedding, or freezing. When the sample is frozen for TEM imaging, it is known as cryo electron microscopy (cryo-EM). The two major techniques used to study the frozen (non-crystalline) sample by TEM are the single particle analysis (SPA) and the cryo-electron tomography (cryo-ET).

1.3 Cryo-Electron Microscopy

Cryo-Electron microscopy, being the method of the year 2015 [37], is the youngest technique used for structure determination. Cryo-EM is the best suited to study structures of large proteins, which cannot be crystallized and that exhibits multiple conformational states. These possibilities make cryo-EM as a powerful technique in structural biology, even though the maximum resolution is, at this moment, not as high when compared to X-ray crystallography method. We can gladly view the the developments in the studies since 1968 to the recent years. The first 3D EM structure of the tail of bacteriophage T4 was obtained by reconstructing a limited set of 2D images from negatively strained electron micrographs [38]. In 2005, the same structure was reconstructed with much higher resolution of 15 Å [39].

With the recent advancements in the detector technology and image processing software techniques, cryo-EM structure determination see a remarkable progress with many data sets at resolutions better than 6 Å and has even reached near-atomic resolution for few highly symmetric complexes, such as virus and very recent ribo-

some data at 2.5 Å [40]. Even the technique has overcome the barriers in imaging the lesser atomic weight complexes, less than 100 kDa. For example, 93 kDa lactate dehydrogenase had been reported to 3.8 Å resolution structure [41]. It is clear that the resolution is reaching atomic accuracy. The rapid developments in the cryo-EM field, can be viewed from the raise in number of entries in PDB database Fig. 1.2. From the graph, we could also infer that the released maps are, on average, in the intermediate resolution range; there are not many very high resolution density maps reported.

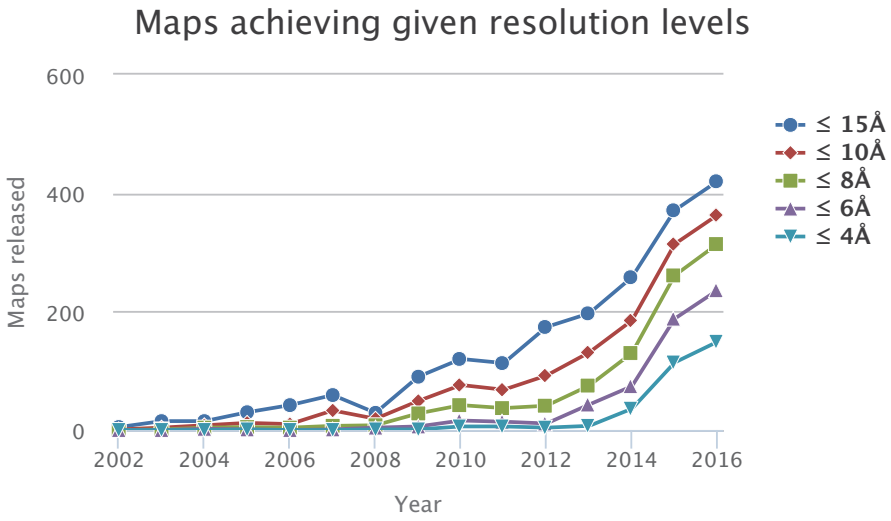

Figure 1.2 Statistics taken from EMBD database (www.ebi.ac.uk) showing the number of released cryo-EM density maps per year at different resolution ranges.

In cryo-EM, the biological sample is placed on an EM grid (classically copper, nickel or gold optionally covered by a thin layer of carbon), it is plunged frozen [42, 43] in liquid ethane. The liquid ethane has a cooling rate of $-10^5 K/\text{sec}$, which can form vitreous ice on the grid. The liquid ethane is surrounded by the liquid nitrogen environment at $-196°C$. Then the sample is imaged under high vacuum. Because the sample is frozen, there is no need for dye or fixatives to hold them in their native state. Plunge freezing a sample, avoids artefacts from fixation, plastic embedding, dehydration, or staining. Only a very small amount of sample is

needed to do cryo-EM imaging. After imaging the sample, the particles are picked from the 2D micrographs either manually or automatically by pattern recognition algorithms. The picked particle images are classified into the different orientations and the set of classified images are averaged to obtain class averages, from which an initial model of the 3D density can be estimated.

1.3.1 Single particle analysis

Single particle analysis is a method, in which the images from the electron microscopy collected by either the negative staining method or the cryo-EM method are processed to get a 3D structure of a molecule or molecular complex. The thickness of the sample has to be less than a few hundred nanometers, which is to avoid multiple scattering of electrons inside the sample. However, the dimension of the sample is still larger than the size of the particle. The sample is imaged using very low electron dose to limit the radiation damage. The covalent bonds break irreversibly, if the biological specimen are exposed to electrons of a typical energy between 100 and 300 keV. In case of breaking of the covalent bond, the distance between the atoms increases. The increase in the atomic distance distorts the molecule and the surrounding ice. The radiation damage causes a beam-induced motion, which is often much larger than the allowed 1 Å, which is the diameter of the hydrogen atom. Due to the beam-induced movement, the micrographs will be blurred and the high-resolution is at stake.

With the older CCD cameras, this blurring could not be avoided, but this has changed after the introduction of direct electron detectors. The direct electron detector can image the sample at a rate of several frames per second, quite the contrary to the CCD cameras which do not have such a fast readout. For the large complexes of MDa weight, the beam-induced movement can be tracked and corrected for every particle separately [44]. The motion correction works well, if the size of the particles is larger than 300 kDa. Because, in the micrographs, the visibility of larger particles is sufficiently good for larger particles [45]. When the complex weight is smaller, the motion correction technique is employed for an ensemble of particles. It is reported that even for small asymmetric proteins the resolution can be improved. By motion correction, they were able to resolve the structure of the γ- secretase complex to 4.5 Å, which is only 170 kDa [46]. For

ideal images, with no particle shift, the particle size can be theoretically as low as 38 kDa [1].

Semi-automatic and fully automatic algorithms are used to pick particles from a large number of 2D micrographs. After identifying the coordinates of the particles, they are aligned and classified. The clustered images are averaged to form 2D class averages, which can be used for sorting images into different states and also for building an initial model. Each particle image is compared to the projections of a reference initial structure into a large number of orientations and the best match is used to assign the corresponding orientation parameters to this particle image. Multiple initial structural references can be used in case of structural heterogeneity. Most popular programs used for image processing of single particle data are RELION [3], SPIDER [47], EMAN2 [48] and XMIPP [49]. If the resolution of the final reconstructed density map is high enough (better than 4.5Å), an atomic model could be built directly from the density map. If the resolution is worse, additional knowledge about the structure (e.g. a known X-ray structure) is required to build or optimize an atomic model that correctly describes the structure corresponding to the map. The final step is to validate and interpret the obtained 3D structure.

1.3.2 Cryo-Electron Tomography

Cryo-electron tomography (cryo-ET) is a three-dimensional imaging technique used to study cellular morphology and structures in their physiological environment. Cryo-ET is similar to magnetic resonance imagining (MRI) and positron emission tomography (PET) used in medical imaging. With the overall advancement both in field of instrumentation and methodology, cryo-ET has become the high resolution imaging technique on cell, tissue, and on membrane scale. Even though the resolution is not comparable to other methods like XRC, the complex does not need to be purified and can be imaged in close-to-native conditions. Cryo-ET is also well suited to study transient and membrane associated complexes. Like in single particle cryo-EM, the sample is plunge frozen to form vitreous ice and 2D projections are recorded using transmission electron microscopy at various tilt angles as in Fig. 1.3 A. The 3D volume is then reconstructed from the tilted projections.

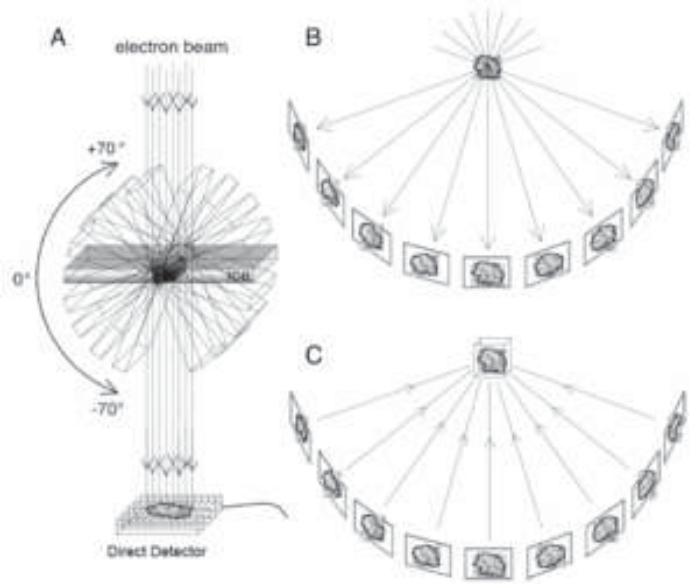

Figure 1.3 Cryo-Electron Tomography (Steven and Belnap, Current Protocols in Protein Science, 2005 ([50])). A) The sample is images at various tilt angles until $\pm 70°$. B) 2D projections at different angles of the sample. C) 2D projections are used for reconstruction of 3D volume.

The sample thickness within 15µm can be plunge frozen and thicker samples are created in amorphous state by high-pressure freezing [51]. In high-pressure freezing, the sample is kept at ∼2000 bar and cooled by spraying liquid nitrogen. The thickness of the sample is limited to 200 nm for an accelerating voltage of 100 kV due to inelastic scattering that electrons would undergo within the sample. Unlike single particle analysis, in cryo-ET the tomograms are recorded at incrementing tilt angles. However, the images cannot be recorded for more than $\pm 70°$ tilt angle. Because the thickness of the sample will become too large, which results in increasing of the effective path length of the electron beam [52]. The tilt angles at which the images cannot be recorded is known as the missing wedge in electron tomography. It would result in a wedge-shaped missing region in the Fourier space where the data has no information. The missing wedge is another severe limitation in electron tomography. It would lead to blurring artefacts and anisotropic resolution in cryo-tomograms. The technique of using a cylindrical specimen holder

like thin walled carbon tubes instead of flat grids are used to overcome the missing wedge problem [53]. But the cylindrical specimen holder which was used for ribosomes and whole bacterial cell has its own limitation that the tubes are fragile and the sample has to be placed at the narrow usable region.

When the sample is too thick then serial section electron tomography can be used. The sample is serially cut to $1 - 2\mu m$ and each section is reconstructed to 3D from single tilt series and finally they are recombined to a single whole volume [54]. The Crowther criterion in equation 1.2 gives the relation for achievable resolution d, for a sample thickness of D and the minimum number of views m needed [55].

$$m = \frac{\pi D}{d} \tag{1.2}$$

As shown in Fig: 1.3.C, the 2D images are back projected by computational methods to form a 3D density map. The images are first aligned to each other, which is usually done using fiducial markers like gold particles. Thereafter their tilt angles and axes are estimated before 3D reconstruction. Many software packages are available to do the reconstruction process [52], such as IMOD [56], SPIDER [57], Bsoft [58], TOM [59] and PyTom [60].

Tomograms reconstructed from a tilt series, usually, have very low resolution. If the same particle can be identified multiple times within the tomogram, then subtomogram averaging is a common method to improve the resolution. The subtomogram averages are the averages of 3D volumes selected from the tomogram to improve the signal-to-noise ratio by iteratively aligning and averaging with a reference model. At every step the highest cross-correlation value between the reference model and the subtomogram is used to evaluate the right rotational and translational alignment. Similar to the single particle analysis, orientations of the particle images are determined at each iteration step by an existing structure, but done in 3D instead of 2D [61]. Due to the missing wedge problem, the average has to be reweighted in Fourier space of the 3D volumes, hence the averaging algorithm is different from those used in single particle cryo-EM, which was developed [62].

1.4 Cryo-EM resolution

A few years ago, cryo-EM density maps were of only low resolution on account of which the field was even called 'blob'-ology. With the advancements in all the different areas, the field is now moving from blob-ology to molecular details [63, 64]. The resolution in cryo-EM is mainly limited by:

- Electron dose
- Quality of detection technique
- Beam optics quality
- Electron beam tilt introducing a phase shift to images
- Estimation of defocus value
- Multiple scattering of electrons within the sample
- Low signal to noise ratio in images

Detectors also play an important role in limiting the achievable high resolution. But the development in instrumentation and methodology have revolutionized the field. Especially, there was a big leap when the charged coupled devices (CCD) and photographic films were replaced with the direct detection devices.

1.4.1 Detectors

Photographic films were used for many years for recording the images and yielded high resolution due to its large number of pixels available [41]. The main drawback with the films is that it has to be developed and scanned and cannot be easily automated, therefore the entire process is not faster, efficient and large scale data collection was problematic. The CCD cameras, which replaced the photographic films, has to convert the electron to photons in the scintillator and then coupled to the sensor through a lens or fiber optics as in Fig: 1.4. Although CCD has fast read-out efficiency, the light scattered within the scintillator leads to signal generation from each electron read by more than one pixel, which, in turn, limits the resolution. If the signal is detected in more than one pixel, a modulation transfer function

(MTF) can be calculated quantitatively [65]. MTF is the Fourier transform of the point spread function (PSF). The PSF describes the blurring effect of a point-object to a minimal size and shape due to scattering. MTF measures the contrast transferred from object to image at each resolution in frequency domain [66]. Direct detection of electrons improves the MTF considerably, this results in enhanced detective quantum efficiency (DQE) of the detector. Due to the multiple scattering of electrons, the output signal-to-noise ratio (SNR) is decreased compared to input SNR, making the DQE at spatial frequency ω lower as shown in equation 1.3. The DQE is affected by both the MTF and the detection noise. Hence, DQE will be lower than the ideal value of 1. This is given by the equation 1.4 in which the noise power spectrum (NPS) represents the noise in the image.

$$\mathrm{DQE}(\omega) = \frac{(\mathrm{SNR}_{\mathrm{out}})^2}{(\mathrm{SNR}_{\mathrm{in}})^2} \tag{1.3}$$

$$\mathrm{DQE}(\omega) = \frac{\mathrm{Dose} \times \mathrm{MTF}(\omega)^2}{\mathrm{NPS}(\omega)} \tag{1.4}$$

The spatial frequency ω ranges from 0 to 0.5 $pixel^{-1}$. The 0.5 $pixel^{-1}$ limit is called the Nyquist frequency. The Nyquist frequency is defined as the minimum rate at which a signal can be sampled without errors, which is twice the highest frequency present in the signal for a pixelated image [65]. The sampling frequency above the Nyquist frequency undergoes sampling error known as aliasing. Cryo-EM data are usually sampled until twice the Nyquist to reduce the aliasing effect and image distortion.

The photographic films are more efficient at high resolution with a DQE $\sim$ 0.3 and the CCD at the low resolution range with DQE $\sim$ 0.1. A high DQE detector combining the advantage of both the films and the CCD has lead to the use of monolithic active pixel sensors (MAPS) [69–71]. These active pixel sensors are called direct detection device (DDD), as the name suggests, it could directly detect the electrons passing through a thin semiconductor membrane and thereafter detected as shown in Fig 1.4, making the DQE higher. Unlike in CCD, image distortion is minimized as there is no coupling needed to transfer the signal for detection and reduced back scattering due to μm thickness allowing to reach higher

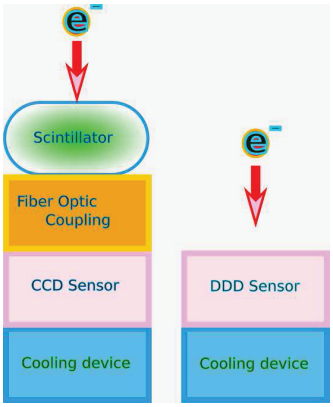

Figure 1.4 The difference between the CCD camera and the direct detection camera techniques. Unlike direct detection cameras, CCD camera uses a scintillator and fiber optics coupling before the electrons are detected by the sensor.

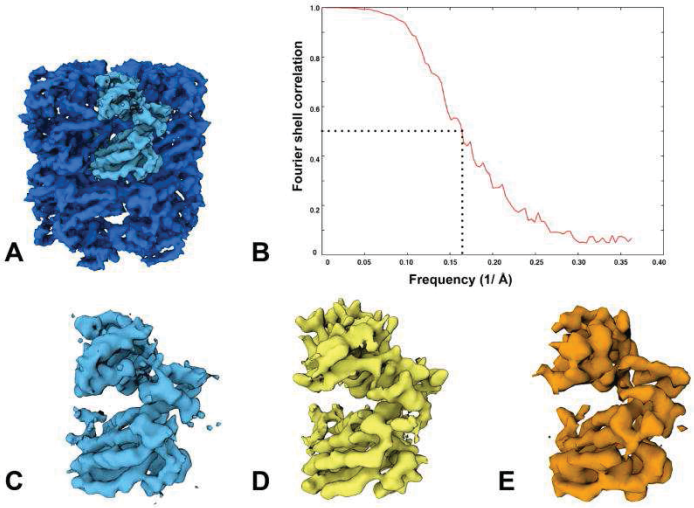

Figure 1.5 Images shows A) GroEL reconstruction of 72,316 particles from direct electron detector. B) Fourier shell correlation curve from the two half maps of the GroEL gives $FSC_{0.5}$ = 6.1 Å. Comparing one of the asymmetric subunits reconstructed from images recorded using C) DDD, D) photographic film from 4 Å map [67] and E) CCD camera from 5.4 Å map [68]. Figure reprinted with permission from A. C. Milazzo et al., Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy, Journal of structural biology, 176: 404-408 (2011).

resolution. From the first structure recorded in 2011- using direct electron detector of GroEL- in Fig. 1.5 [72], the field has taken a big leap that its even known as the resolution revolution [45].

1.4.2 Resolution Criteria

Resolution in optics is defined as the smallest distance between two points in the image that can be distinguished as individual entities. In optics and X-ray crystallography, the resolution depends on the orders of Fourier component of the signal part available for Fourier synthesis [73] and is defined as

$$d = 0.61 \frac{\lambda}{n \sin \alpha} \tag{1.5}$$

where λ is the wavelength of the radiation used, n is the refractive index of the medium (n=1 for air and 1.51 for oil or glass) and α is the biggest scattering angle and $n sin\alpha$ gives the numerical aperture. Imposing the condition that the numerical aperture should be large to capture the first diffraction pattern from the source, makes the value as $\sim$1 for light microscopy and $\sim$0.01 for electron microscopy. So, the resolution limit for light microscopy is 25-42 Å for wavelength of 400-700 nm and for electron microscopy λ depends on the accelerating voltages as

$$\lambda = \sqrt{\frac{h^2}{2\,m\,e\,V}} \tag{1.6}$$

h being the Planck's constant, m and e are the mass and charge of the electron respectively and V is the accelerating voltage. For typical voltage used in experiments, $V = 300$ KeV, the resolution limit should be 1.2 Å [74]. But there are several factors forbidding the resolution to reach close to this limit [75].

The Fourier component of the signal will be concentrated as a peak and the noise will be as a background in electron-crystallography. By subtracting the peak density with the mean of the background noise, would give an idea about the resolution [73]. With non-crystalline nature of cryo-EM sample, the resolution measurement does not hold the same definition. When the structural information is almost near to atomic resolution, the resolution can be fully or partially be as-

sessed by the quality of the structure, specifically, by the quality of the α-helices. But in cryo-EM the resolution in most case is not very close to resolutions allowing to clearly judge the secondary structural information, one has to use the statistical measure to estimate resolution [74]. In cryo-EM, the data set is randomly divided into two half sets of the data. The correlation in the 3D Fourier space between those halfs, defines the resolution of the data. The Fourier shell correlation (FSC), was introduced by 1986 by Harauz and van Heel [76, 77]. It is the correlation of the structure factors of two density maps in the complex Fourier space and it is a 3D extension of 2D Fourier ring correlation (FRC) [78, 79].

$$\mathrm{FSC}(r) = \frac{\sum_{r_i \in r} F_1(r_i).F_2(r_i)^*}{\sqrt{\sum_{r_i \in r}|F_1(r_i)|^2.\sum_{r_i \in r}|F_2(r_i)|^2}} \tag{1.7}$$

F_1 is the complex structure factor of one density map, F_2^* is the complex conjugate of the structure factor of second density map and r_i is the voxel element at radius r.

The structure factors of the two volumes can be written as a sum of the common signal term (S) and and noise terms (N_1 and N_2). Hence, the correlation can be written as:

$$\mathrm{Correlation} = \frac{\sum(S+N_1)(S+N_2)^*}{\sqrt{\sum|S+N_1|^2\sum|S+N_2|^2}} \tag{1.8}$$

If the signal and noise are uncorrelated and have the same signal level, the above equation can also be written as:

$$\mathrm{FSC} = \frac{\sum|S|^2}{\sum|S^2+N^2|} \tag{1.9}$$

When the FSC = 0.5, then S^2 is almost equal to N^2. The FSC = 0.5 criterion works well since it is independent of how the density map is reconstructed, i.e., it could be used when there is a model bias or same model used for both the half maps for reconstruction. This FSC criterion of 0.5 is however an underestimation of the resolution, since the value represents the signal or resolution corresponding to only half the data. Using the whole data information, the noise would reduce to

N/$\sqrt{2}$ [80] making the above equation,

$$\mathrm{FSC_{full}} = \frac{2\,\mathrm{FSC}}{1+\mathrm{FSC}} \tag{1.10}$$

A better criterion for resolution assessment for full data set would be to compare the full data to a perfect noise free density. Hence, one half of the data set can be assumed as full density as in the equation 1.10, which is the average of both half maps and the other half as the noise free density. Substituting the above conditions in the equation 1.8 and having a threshold $C_{ref} = 0.5$ will give a FSC = 0.143.

$$C_{\mathrm{ref}} = \sqrt{\frac{S^2}{S^2 + \frac{N^2}{2}}} = \sqrt{\frac{2\,\mathrm{FSC}}{1+\mathrm{FSC}}} \tag{1.11}$$

In terms of structure factor, the equation 1.11 can be rewritten as below:

$$C_{\mathrm{ref}} = \frac{\sum |F_1||F_{\mathrm{ref}}| \cos(\Delta\phi)}{\sqrt{\sum |F_1|^2 \sum |F_{\mathrm{ref}}|^2}} \tag{1.12}$$

The correlation $\Delta\phi$ between the perfect phase from F_{ref} and experimental map F_1 gives the phase error, which is analogous to the figure-of-merit m in X-ray crystallography [81]. A phase error of 60°, corresponding to $m = 0.5$, represents the data that is good enough for structure modeling [82]. Also, the real space correlation coefficient is equal to the equation 1.12. Hence, the $C_{ref} = 0.5$ is considered as a better choice than C_{ref}=0.8 at FSC = 0.5. The FSC calculated from the two half maps that are reconstructed with the independent models, is known as gold-standard FSC [83], which is typically used to determine the resolution with the FSC = 0.143 criterion.

From the reference [84], the correlation coefficient table 1.1 gives a clear picture of the different FSC criteria.

Table 1.1 Correlation coefficient (Table A1 from [84]

FSC	$\mathrm{FSC_{full}}$	C_{ref}	Phase error(degree)	$S/N_{1/2}$	S/N_{full}
0.50	0.67	0.82	35	1.00	1.41
0.33	0.50	0.71	45	0.71	1.00
0.14	0.25	0.50	60	0.41	0.58

There are two resolution criteria available for cryo-ET data. Analogous to SPA, resolution assessment is done by splitting the tilt series into two data sets and estimating the correlation of the structure factors (FSC) in the Fourier space. The second criterion, known as the noise compensated leave one out (NLOO) is performed on the 2D Fourier ring correlation between one of the original projection image and the reprojected image from the tomogram reconstructed using remaining images. The reprojected tomogram image would have better SNR, hence the noise has to be compensated when calculating the cross correlation. NLOO provides a resolution assessment by tilt angles, hence it is computationally expensive [85]. If the resolution was only noise limited, both the FSC and NLOO would be identical. Very similar to SPA, the resolution is greatly limited due to the conformational heterogeneity. Until now, cryo-ET data is mostly in the low resolution range and only very few higher resolution structures have been published in recent years. The EMBD statistics from August 2016, fig: 1.6 shows the resolution of the average data deposited is around 19.6 Å and highest 14.7 Å and that of subtomogram averages is around 35.1 Å (red curve) and the highest resolution (blue curve) of 3.9 Å. Despite the fact that cryo-ET is the highest resolution imaging technique available to study the whole cellular architecture, it does not easily allow to study the individual components in atomic detail.

Figure 1.6 Statistics taken from EMBD database (https://www.ebi.ac.uk). Cryo-electron tomography resolution trend for tomograms (left) and subtomogram averages (right) having the average (red) and highest (blue) resolution per year in both the plots.

With the great advantage of applying cryo-EM to various biomolecular com-

plexes and with the overall growth of the field, there is a high demand accessing high resolution information in cryo-EM data to get the very detail of a structure, for example in the case of developing new drugs. Because of this need for new methods to improve the resolution of the cryo-EM data, we have developed a new algorithm which could overcome the hurdle due to conformational heterogeneity to reach high resolution.

1.5 Conformational heterogeneity

Cryo-preserving the sample is the best technique to study the conformational flexibility under nearly physiological condition. But analyzing different conformers in the data is a very tedious computational task. Although most of the time these conformers are considered as an inconvenience, studying them helps us to get a deep insight into the pathway of the biological functions. Several methods have been developed to handle the conformational heterogeneity by classifying and sorting the images into different conformational classes. Most common in use are bootstrapping [86], maximum likelihood alignment [87], normal mode analysis [88], brownian trajectories [89], covariance matrix [90], and principle component analysis [91].

In the biological macromolecules, the conformational motions leads to global structural changes, however they can be split into smaller rather rigid domains. Rigid domains are regions where the structure would remain very similar in all the conformations and could be chosen such that the rigid fitting of them leads to perfect fit of the whole domain. Those rigid domains could be aligned and averaged individually and later combined back to get a higher resolution density map. This is analogous to NCS averaging used to improve the phase information in the field of X-ray crystallography [92]. Instead of doing rigid fitting for many rigid domains, one could also flexibly bend or fit the whole density map in one step. The advantage in flexibly bending the whole map is to avoid the step of cutting into smaller rigid domains and later combining all the domains after rigid fitting, to get the full density map that could lead to artefacts if not done carefully.

Rigid fitting or docking is a relatively simple method since it is only a search to match in the six-dimensional translational and rotational space. For rigid fitting a

structure or a density map to a density map, there are fully automated software like Situs [93], Chimera [94], EMfit [95], Modeller [96]. When a previously known atomic structure that belongs to one conformation, has to be fitted to another conformational map, rigid fitting is not appropriate anymore. There emerged the need for flexible fitting or what is commonly known as refinement of structures. Refinement of atomic structures to density maps have started with the development of XRC [97], thus paving the way for the development of many fully automated software packages. Molecular dynamics (MD) simulations have been introduced by Brunger et al. to improve the refinement [98]. Molecular dynamics flexible fitting (MDFF) [99] is a simulation tool combining MD and extra potential derived from EM map. There are also methods combining the normal mode analysis (NMA) [100] with the elastic network model (ENM) [101]. ENM is attracting due to its simplicity, robustness and scalability. In ENM, the complex molecular structure is broken down to nodes and springs. Nodes represent the residues and the springs are the connections between the near by residues. The positions of the nodes are obtained from prior known structures and this reduced model gives the distribution of interaction to determine the structural dynamics. Thus, making it faster and robust when compared to full-fledged atomic force calculations [102]. More recently developed Deformable elastic network (DEN) implemented in DireX [103], CNS [104], and later in Phenix [105] use a prior known reference model in which atom pairs are connected randomly within a fixed interval range. There are two kinds of forces acting against each other. One is the force that restores the structure to the reference model and other that pulls towards the EM map.

Flexible fitting or refinement is very helpful to understand the conformational dynamics of the proteins. Until now, only flexible fitting of atomic structures into density exists and there is no method available to elastically or flexibly fit a 3D density map to another density. Hence, we developed the ELastic alIGN (ELGIN) method which is similar to the elastic image registration technique in medical PET and CT scan, elastic volume reconstruction in serial-section microscopy, where the ultra-thin microscopy sections are elastically aligned to reassemble the volume with minimal artificial deformation [106], and very common in computer graphics-morphing and remote sensing. StructMap developed for visualization of conformation difference by elastic transformation also uses iterative 3D-to-2D elastic align-

ment [107].

1.6 Refinement of atomic models

DireX is the main software used as a basis for developing all the methods in this thesis, particularly the ELIGN algorithm. Molecular dynamics flexible fitting was another program used in the precision measurement of Cartesian coordinates during structure refinement. Other software packages like EMAN, Chimera and VMD were used as tools for filtering, FSC estimation and visualization since they are very well established for these tasks.

1.6.1 DireX

DireX was initially developed for structure refinement to fit atomic structures into density maps at low resolutions in real space. The idea behind the method was to change only those degrees of freedom provided by the experiment and to keep others close to the reference structure. DireX makes use of the geometry based sampling iterative algorithm CONCOORD [108]. DireX uses the deformable elastic network (DEN) restraint potential for refinement.

As described earlier, in the DEN method random atom pairs within an interval range of typically ∼3-15 Å are connected by harmonic distance restraints. There are two kinds of opposing restraints determining the amount of deformation: One is the restraint from the reference structure and the other from the experimental density map. As described in Fig. 1.7, the potential has its minimum at the reference structure at the start. With a proper balance between the experimental data and the restraints used, the potential is moved to fit to the density map. The important parameter while using deformable elastic networks is γ. The γ parameter defines the extent of flexibility for the DEN and acts like a spring constant. If the γ value is 0, then the restraints are not deformable, the DEN potential is identical to a regular elastic network and the output stays very close to the reference structure. In the case of unit γ value, the DEN potential can slowly follow the refined structure and does not use the reference information at all. Some other important parameters used are map strength, distance restraints strength, and the cross-validation cut-off values. The map strength defines the force that acts on the structure to fit the model

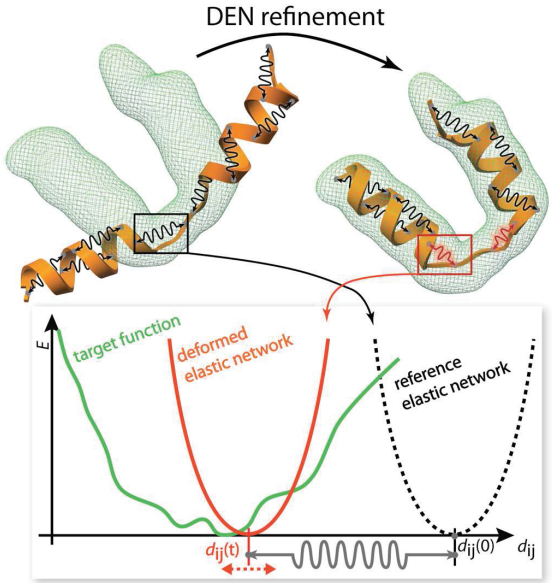

Figure 1.7 Principle of the deformable elastic network (DEN) method. The DEN potential is defined by distance restraints (black springs) between randomly chosen atom pairs. During the refinement the target distances, $d_{ij}(t)$, of the restraints are allowed to follow the motion of the structure as it is refined to fit the target density (green) and in addition are pulled towards the corresponding distances in a reference model $d_{ij}(0)$ (Figure adapted from www.schroderlab.org)

into the density map (the default map strength value is 0.03 in arbitrary units). Distance restraints act like a spring between atom pairs, where the strength of the restraints controls the amount of movement between the defined pairs of atoms. The distance restraint strength value ranges from 0 to 1; a value of 0 means the distance does not change, and a value of 1 means that the distance is not restrained at all (the default value is set to 0.4).

$$E_{\mathrm{DEN}}(t) = k \sum_{\text{pairs } i,j} (d_{ij}(t) - d^0_{ij}(t))^2 \tag{1.13}$$

The DEN potential is defined as in equation 1.7, where k is the spring constant, $d_{ij}(t)$ is the distance between the atom pairs i and j at t^{th} step of the simulation, and $d^0_{ij}(t)$ is the equilibrium distance at step t.

DireX can also use a generic bead model instead of an atomic structure, where the "atoms" are simply point masses without any further geometric constraints. Any density map can be represented by such a bead model, for example by placing point masses onto the grid points of the density grid with a mass that is proportional to the density value at the corresponding grid point. Such a bead model can be used when there is no known structure available.

1.6.2 DireX cross-validation

DireX uses cross-validation, a tool to measure the goodness of fit at every step. Cross-validation is very important to avoid overfitting. Similar to the cross-validation method in X-ray crystallography for refinement of atomic structures, a validation method is also available in DireX. In X-ray crystallography, the R-factor is the measure of the goodness of fit of protein models [109]. The equation 1.14 gives the expression to calculate the R-factor from the amplitudes of structure factors from the experimentally observed diffraction pattern (F_{obs}) and the structure factors calculated from the model (F_{calc}). R-factor is calculated at every step and the final value is considered as a measure of quality.

$$\text{R-factor} = \frac{\sum|F_{\text{obs}} - F_{\text{calc}}|}{\sum|F_{\text{obs}}|} \tag{1.14}$$

R-factors ranges from 0 for a perfect model to 0.6 for a completely wrong model. If the R-factor is typically improved by the structure refinement. The value will be $\sim$ 0.2 for a model refined with 2.5Å data. R-factor value alone cannot validate the model but additional measures such as protein geometry, Ramachandran plot, and rotamer analysis are also needed. The R-factor is however prone to overfitting, it is possible to decrease the R-value in the refinement even though the structure does not improve, since the number of parameters is often larger than the number of experimental observables. To monitor overfitting, the R_{Free} value was introduced by Brünger in 1992 [110], its the same as R-factor but calculated only for randomly chosen structure factors that were not used in the refinement (test set) typically about 5 to 10%. This test set is used only for assessment, the R_{Free} value is more reliable than the R-factor, since its free from any bias from refinement. The rest of the data set (work set) consisting of 90-95%, which defines the R_{Work}

value, is used for the refinement.

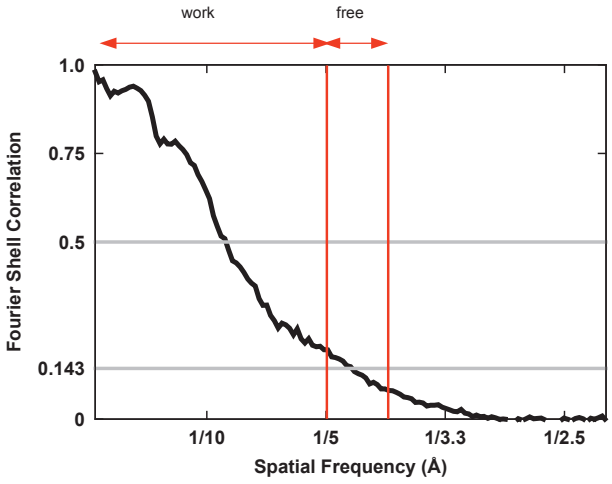

Figure 1.8 The FSC curve along with the work band and free band and the two resolution criteria at 0.143 and 0.5 used in general. The interval between the two red lines makes the cfree interval and the interval with higher frequency from 5Å forms the work range.

A cross-validation method is implemented in DireX [111], which is similar to the procedure used in crystallography, where a random set of structure factors is chosen as the test set. Choosing random set as independent test and work set, is possible only in crystallography and not in cryo-EM. The finite size of the protein inside a box in real space [112, 113], and the image alignment procedure for 3D reconstruction [114] leads to correlations of structure factors in Fourier space. The correlation is bigger between low- than between high-resolution Fourier shells. For the above mentioned reasons, unlike X-ray crystallography, the test set in cryo-EM is not computed for a random set from the data but for a continuous band in a high-resolution Fourier shell. As shown in fig 1.8 the high spatial frequency band which has low signal to noise ratio is used as the free band and used only for validation and the work band is used for refinement. Typically the high resolution end of the free band is the resolution of the density map and the width of the band can be 2Å or 3Å. The free and the work set are more independent as the free band gets wider. The C_{Free} value is the cross-correlation coefficient between the model and

EM density map, where both maps were band-pass filtered with the free band. The C_{Free} value is a good indicator of the refinement quality and should ideally increase during the refinement as long as the model is not overfitted. When the structure fits very well within the density map, then the C_{Free} would converge. The parameter in DireX, C_{Free}, is therefore similar to the R_{Free} value in crystallography.

1.6.3 Molecular dynamics

Molecular dynamics (MD) simulation is defined as study of the atomic and molecular interactions that govern the properties of the physical system. It is possible to do various types of MD simulation depending on the types of interactions and parameters. If the interactions are described quantum mechanically, using Schrödinger equations, then the simulation is called quantum mechanical (QM) simulation. QM is very accurate and expensive computational tool, hence can be used only on small systems like 10-100 atoms on a timescale of 10-100 picoseconds. To study less accurate and larger physical systems (10^4 -10^5 on atoms 10-100 nanoseconds) Newtonian forces are used, it is known as classical molecular mechanics (MM). MM is usually a very fast calculation tool. Mixed quantum and classical interactions (QMMM) can be used for the systems of 10^4-10^5 atoms on 10-100 picosecond timescale.

Force field in MD simulations accounts to the total interaction energy due to degrees of freedom for the atoms, interaction among the atoms, physical conditions such as temperature, pressure and wall force. The force field defines the energy landscape of the system computed from the functional form and the parameters. There are many kind of force fields each optimized using empirical data for different molecular groups. Some force fields includes also additional function such as hydrogen bond, geometrical parameters and anharmonic corrections to a harmonic oscillation. Depending on how the proteins are embedded in the solvent, i.e., including or excluding the water molecules they are known as implicit or explicit solvent respectively. Some of the common force fields used in the field of MD are CHARMM [115] (implicit, explicit, all atom and united atom calculations), AMBER [116] (implicit, explicit and all atom calculations), GROMOS [117] (explicit, vacuum and united atoms) and packages like GROMACS [118] (AMBER, GROMOS) and NAMD [119] (CHARMM, AMBER and GROMOS) use combination

of force fields.

In our method, we use classical molecular mechanics simulations using Nanoscale molecular dynamics (NAMD) force fields. Its a parallel molecular dynamics program written on Charm++ parallel objects but also has the advantage that it can be used with CHARMM, AMBER and X-PLOR. It uses the most common visualization software Visual molecular dynamics (VMD) for setting up the simulation and analysis of the trajectories thereafter. NAMD was developed initially to meet the increasing demand to study the structure of large biomolecules and also to use the growing hardware technology. Hence, NAMD serves as a better computing tool that can be used parallel on thousands of cores for larger complexes. NAMD 2.9 has a desktop version known as MDFF by implementing implicit solvent simulation done by GPU acceleration, "lite" grid forces for faster computing and taking advantage of the shared memory of a machine for optimization.

1.6.4 Molecular dynamics flexible fitting (MDFF)

MDFF is used for flexibly fitting of atomic structures into density maps. MDFF uses NAMD (CHARMM27) force field with dielectric constant of 80 for *in vacuo* simulations. Langevin thermostat is used to maintain the temperature at 300K. MDFF has additional external forces into MD simulation to guide the atoms into the high density regions which corresponds to the energy minima. This force is proportional to the gradient of the EM map. MD force fields take care of the stereochemistry of the structure where as the harmonic restraints are used to preserve the secondary structure and nucleic acids. MDFF total potential energy can be written as sum of three potentials as described in [99]

$$U_{\text{total}} = U_{\text{MD}} + U_{\text{EM}} + U_{\text{SS}} \tag{1.15}$$

here, U_{MD} stands for the MD potential energy, U_{EM} for the potential energy from EM map and U_{SS} for the potential that keeps the wholeness of the secondary structure.

$$f_i^{\text{EM}} = -\frac{\partial}{\partial r_i} U_{\text{EM}}(R) = -w_i \frac{\partial}{\partial r_i} V_{\text{EM}}(r_i) \tag{1.16}$$

where,

$$V_{\mathrm{EM}}(r) = \begin{cases} \xi\left[\frac{\Phi(r)-\Phi_{\mathrm{thr}}}{\Phi_{\mathrm{max}}-\Phi_{\mathrm{thr}}}\right], & \text{if } \Phi(r) \geq \Phi_{\mathrm{thr}} \\ \xi, & \text{if } \Phi(r) \geq \Phi_{\mathrm{thr}} \end{cases} \tag{1.17}$$

Coulomb potential $\Phi(r)$ from the map, Φ_{max} maximum density value in the map, Φ_{thr} is threshold value, below which is mostly solvent density, hence, the data is discarded, ξ is the arbitrary scaling factor, w_i is the atomic mass of ith atom and r_i is the position of the ith atom.

The force f_i^{EM} from the EM map acting on each atom in MDFF is done by grid-steered molecular dynamics [120] which depends on the gradient of the potential at the atom position, making it a local fitting method. This external force can be scaled same for all atoms by the scaling faction ξ. Whereas the weight w_i can be defined on each individual atom separately typically by its atomic mass.

The secondary structure can also be preserved by using harmonic restraints to the set of secondary structure coordinates to the initial structure. Harmonic restraints for the protein are applied by dihedral angles.

$$U_{\mathrm{SS}} = \sum_{\mu} k_{\mu}(X_{\mu} - X_{\mu}^{0})^2 \tag{1.18}$$

X_{μ} stands for the harmonics restraints and X_{μ}^{0} is the equilibrium value from the initial structure.

We have used the standard MDFF refinement procedure in our method. As an initial preparation in the MDFF protocol known as the `mdff setup` function, various input files for both the atomic structure and the density map has to be prepared. Firstly, the structure is rigidly docked into the EM map. This can be done using any software which can fit an atomic structure to the density map. Situs is used here, it gives the best few structures and the first best fitted structure is taken as the starting structure. Since MDFF uses NAMD, a psf and pdb file with the contains the information about the atomic connectivity and charge is created in VMD. Also per atom scaling factor for the whole pdb is needed which is generated by `mdff gridpdb` function. To avoid overfitting secondary structures, cis/trans peptide configuration and chirality restraints can be given as extra informations by the extrabonds feature in NAMD. Girdforces function in NAMD can be used to

define an external potential on the 3D grid. Hence, the density map has to be converted to the potential U_{EM} for using it in MDFF by `mdff griddx` command. The potential can be scaled to the need by a scaling function which is fed as an additional *gscale* parameter. The parameter *gscale* defines the strength of the density map force on the atomic structure used for the refinement. Higher *gscale* means a higher force on the atomic structure to fit into the map. Even though *gscale* should be chosen according to the system, typically, a lower *gscale* value around 0.3 is used. The time for the MD simulation can be defined with the parameter *numsteps* and given always in terms of nanoseconds. In our case the simulation was done for 500ps time which is good enough for the refinement to converge. With all the above mentioned initially prepared files the first configuration NAMD file is created which is used as input for the second NAMD configuration file. In the second step, only energy minimization is done with a much higher scaling factor (*gscale* =10) than used in the first step with *minsteps* parameter. The MDFF setup steps will give output (.namd) files. NAMD should be run with the `namd2` command with these file as an input. The output of the `namd2` run would be a trajectory of the refinement. NAMD has the option of running parallely on many processors with the $+p$ option. Parameter optimization is done by refining the structure with various *gscale* values.

1.6.5 Root mean square fluctuation (RMSF)

The root mean square fluctuation (RMSF) is used for estimating the positional precision in real space refinement described in chapter 4. The root mean square deviation (RMSD) is used to determine only the divergence of a protein structure relative to a reference structure over time. RMSD is the average over all atoms with respect to a specific time.

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i(t) - r_i^{\mathrm{ref}})^2} \tag{1.19}$$

Equation used for calculating RMSD value, where r_i is the position of atom i

and r_i^{ref} is the position of the same atom in the reference structure.

$$\mathrm{RMSF}(i) = \sqrt{\frac{1}{T}\sum_{t_i=1}^{T}(r_i(t_j) - r_i^{\mathrm{ref}})^2} \tag{1.20}$$

Whereas, the RMSF given by the above equation is the average over time T. Thus, RMSF gives the measure of atomic movement for each residue averaged over time. RMSF is useful to determine the most flexible part of the protein which otherwise cannot be done with RMSD. RMSF is also know as "RMSD per residue" since it is the value corresponding to the RMSD time averaged for each residue. Hence, the mean RMSF can be used to calculate the average atomic movement in the ensemble of protein structures. RMSF can be easily calculated for the selected ensemble of protein conformation by the readily available programs.

Bibliography

[1] R. Henderson, "The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules," *Quarterly Reviews of Biophysics*, vol. 28, p. 171, may 1995.

[2] P. A. Penczek, J. Frank, and C. M. Spahn, "A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation," *Journal of Structural Biology*, vol. 154, no. 2, pp. 184–194, 2006.

[3] S. H. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519–530, 2012.

[4] E. V. Orlova and H. R. Saibil, "Methods for Three-Dimensional Reconstruction of Heterogeneous Assemblies," in *Methods in Enzymology*, vol. 482, pp. 321–341, Elsevier, 2010.

[5] M. van Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan, "Single-particle electron cryo-microscopy: towards atomic resolution," *Quarterly Reviews of Biophysics*, vol. 33, no. 4, pp. 307–369, 2000.

[6] S. A. Villarreal and P. L. Stewart, "CryoEM and image sorting for flexible protein/DNA complexes," *Journal of Structural Biology*, vol. 187, no. 1, pp. 76–83, 2014.

[7] J. C. KENDREW, G. BODO, H. M. DINTZIS, R. G. PARRISH, H. WYCKOFF, and D. C. PHILLIPS, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.," *Nature*, vol. 181, pp. 662–6, mar 1958.

[8] J. C. KENDREW, R. E. DICKERSON, B. E. STRANDBERG, R. G. HART, D. R. DAVIES, D. C. PHILLIPS, and V. C. SHORE, "Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution.," *Nature*, vol. 185, pp. 422–7, feb 1960.

[9] M. F. PERUTZ, M. G. ROSSMANN, A. F. CULLIS, H. MUIRHEAD, G. WILL, and A. C. NORTH, "Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis.," *Nature*, vol. 185, pp. 416–22, feb 1960.

[10] C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips, and V. R. Sarma, "Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution.," *Nature*, vol. 206, pp. 757–61, may 1965.

[11] G. Kartha, J. Bello, and D. Harker, "Tertiary structure of ribonuclease.," *Nature*, vol. 213, pp. 862–5, mar 1967.

[12] H. W. Wyckoff, K. D. Hardman, N. M. Allewell, T. Inagami, L. N. Johnson, and F. M. Richards, "The structure of ribonuclease-S at 3.5 A resolution.," *The Journal of biological chemistry*, vol. 242, pp. 3984–8, sep 1967.

[13] B. W. Matthews, P. B. Sigler, R. Henderson, and D. M. Blow, "Three-dimensional structure of tosyl-alpha-chymotrypsin.," *Nature*, vol. 214, pp. 652–6, may 1967.

[14] J. Drenth, J. N. Jansonius, R. Koekoek, H. M. Swen, and B. G. Wolthers, "Structure of papain.," *Nature*, vol. 218, pp. 929–32, jun 1968.

[15] W. N. Lipscomb, J. A. Hartsuck, F. A. Quiocho, and G. N. Reeke, "The structure of carboxypeptidase A. IX. The x-ray diffraction results in the light of the chemical sequence.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 64, pp. 28–35, sep 1969.

[16] C. S. Wright, R. A. Alden, and J. Kraut, "Structure of subtilisin BPN' at 2.5 angström resolution.," *Nature*, vol. 221, pp. 235–42, jan 1969.

[17] Y. Shi, "A Glimpse of Structural Biology through X-Ray Crystallography," *Cell*, vol. 159, pp. 995–1014, nov 2014.

[18] M. G. Rossmann, "The molecular replacement method.," *Acta crystallographica. Section A, Foundations of crystallography*, vol. 46 (Pt 2), pp. 73–82, feb 1990.

[19] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, N. Echols, J. J. Headd, L.-W. Hung, S. Jain, G. J. Kapral, R. W. Grosse Kunstleve, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "The Phenix software for automated determination of macromolecular structures.," *Methods (San Diego, Calif.)*, vol. 55, pp. 94–106, sep 2011.

[20] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of Coot," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 486–501, apr 2010.

[21] P. V. Afonine, M. Mustyakimov, R. W. Grosse-Kunstleve, N. W. Moriarty, P. Langan, and P. D. Adams, "Joint X-ray and neutron refinement with phenix.refine," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 1153–1163, nov 2010.

[22] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin, "REFMAC 5 for the refinement of macromolecular crystal structures," *Acta Crystallographica Section D Biological Crystallography*, vol. 67, pp. 355–367, apr 2011.

[23] W. Liu, D. Wacker, C. Wang, E. Abola, and V. Cherezov, "Femtosecond crystallography of membrane proteins in the lipidic cubic phase.," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 369, p. 20130314, jul 2014.

[24] X. E. Zhou, X. Gao, A. Barty, Y. Kang, Y. He, W. Liu, A. Ishchenko, T. A. White, O. Yefanov, G. W. Han, Q. Xu, P. W. de Waal, K. M. Suino-Powell, S. Boutet, G. J. Williams, M. Wang, D. Li, M. Caffrey, H. N. Chapman, J. C. H. Spence, P. Fromme, U. Weierstall, R. C. Stevens, V. Cherezov, K. Melcher, and H. E. Xu, "X-ray laser diffraction for structure determination of the rhodopsin-arrestin complex.," *Scientific data*, vol. 3, p. 160021, 2016.

[25] M. M. Seibert, T. Ekeberg, F. R. N. C. Maia, M. Svenda, J. Andreasson, O. Jönsson, D. Odić, B. Iwan, A. Rocker, D. Westphal, M. Hantke, D. P. De-

Ponte, A. Barty, J. Schulz, L. Gumprecht, N. Coppola, A. Aquila, M. Liang, T. A. White, A. Martin, C. Caleman, S. Stern, C. Abergel, V. Seltzer, J.-M. Claverie, C. Bostedt, J. D. Bozek, S. Boutet, A. A. Miahnahri, M. Messerschmidt, J. Krzywinski, G. Williams, K. O. Hodgson, M. J. Bogan, C. Y. Hampton, R. G. Sierra, D. Starodub, I. Andersson, S. Bajt, M. Barthelmess, J. C. H. Spence, P. Fromme, U. Weierstall, R. Kirian, M. Hunter, R. B. Doak, S. Marchesini, S. P. Hau-Riege, M. Frank, R. L. Shoeman, L. Lomb, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, C. Schmidt, L. Foucar, N. Kimmel, P. Holl, B. Rudek, B. Erk, A. Hömke, C. Reich, D. Pietschner, G. Weidenspointner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, I. Schlichting, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. Kühnel, R. Andritschke, C.-D. Schröter, F. Krasniqi, M. Bott, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, H. N. Chapman, and J. Hajdu, "Single mimivirus particles intercepted and imaged with an X-ray laser.," *Nature*, vol. 470, pp. 78–81, feb 2011.

[26] M. Marvin Seibert, S. Boutet, M. Svenda, T. Ekeberg, F. R. N. C. Maia, M. J. Bogan, N. Tîmneanu, A. Barty, S. Hau-Riege, C. Caleman, M. Frank, H. Benner, J. Y. Lee, S. Marchesini, J. W. Shaevitz, D. A. Fletcher, S. Bajt, I. Andersson, H. N. Chapman, and J. Hajdu, "Femtosecond diffractive imaging of biological cells," *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 43, p. 194015, oct 2010.

[27] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu, "Potential for biomolecular imaging with femtosecond X-ray pulses.," *Nature*, vol. 406, pp. 752–7, aug 2000.

[28] C. Caleman, G. Huldt, F. R. N. C. Maia, C. Ortiz, F. G. Parak, J. Hajdu, D. van der Spoel, H. N. Chapman, and N. Timneanu, "On the feasibility of nanocrystal imaging using intense and ultrashort X-ray pulses.," *ACS nano*, vol. 5, pp. 139–46, jan 2011.

[29] J. C. H. Spence, U. Weierstall, and H. N. Chapman, "X-ray lasers for structural and dynamic biology," *Reports on Progress in Physics*, vol. 75,

p. 102601, oct 2012.

[30] R. Neutze and K. Moffat, "Time-resolved structural studies at synchrotrons and X-ray free electron lasers: opportunities and challenges.," *Current opinion in structural biology*, vol. 22, pp. 651–9, oct 2012.

[31] L. Reimer, *Transmission Electron Microscopy*, vol. 36 of *Springer Series in Optical Sciences*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1984.

[32] R. C. Burghardt, R. Droleskey, R. C. Burghardt, and R. Droleskey, "Transmission Electron Microscopy," in *Current Protocols in Microbiology*, pp. 2B.1.1–2B.1.39, Hoboken, NJ, USA: John Wiley & Sons, Inc., dec 2006.

[33] K. Vernon-Parry, "Scanning electron microscopy: an introduction," *III-Vs Review*, vol. 13, no. 4, pp. 40–44, 2000.

[34] M. Haider, S. Uhlemann, E. Schwan, H. Rose, B. Kabius, and K. Urban, "Electron microscopy image enhanced," *Nature*, vol. 392, pp. 768–769, apr 1998.

[35] O. L. Krivanek, M. F. Chisholm, V. Nicolosi, T. J. Pennycook, G. J. Corbin, N. Dellby, M. F. Murfitt, C. S. Own, Z. S. Szilagyi, M. P. Oxley, S. T. Pantelides, and S. J. Pennycook, "Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy," *Nature*, vol. 464, pp. 571–574, mar 2010.

[36] K. Takayanagi, S. Kim, S. Lee, Y. Oshima, T. Tanaka, Y. Tanishiro, H. Sawada, F. Hosokawa, T. Tomita, T. Kaneyama, and Y. Kondo, "Electron microscopy at a sub-50 pm resolution.," *Journal of electron microscopy*, vol. 60 Suppl 1, no. suppl 1, pp. S239–44, 2011.

[37] "Method of the Year 2015," *Nature Methods*, vol. 13, pp. 1–1, dec 2015.

[38] D. J. DE ROSIER and A. KLUG, "Reconstruction of Three Dimensional Structures from Electron Micrographs," *Nature*, vol. 217, pp. 130–134, jan 1968.

[39] V. A. Kostyuchenko, P. R. Chipman, P. G. Leiman, F. Arisaka, V. V. Mesyanzhinov, and M. G. Rossmann, "The tail structure of bacteriophage t4 and its mechanism of contraction," *Nat Struct Mol Biol*, vol. 12, pp. 810–813, 09 2005.

[40] Z. Liu, C. Gutierrez-Vargas, J. Wei, R. A. Grassucci, N. Espina, S. Madison-Antenucci, L. Tong, and J. Frank, "Determination of the ribosome structure to a resolution of 2.5 Å by single-particle cryo-EM," *Protein Science*, oct 2016.

[41] A. Merk, A. Bartesaghi, S. Banerjee, V. Falconieri, P. Rao, M. I. Davis, R. Pragani, M. B. Boxer, L. A. Earl, J. L. S. Milne, and S. Subramaniam, "Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery.," *Cell*, vol. 165, pp. 1698–707, jun 2016.

[42] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall, "Cryo-electron microscopy of viruses," *Nature*, vol. 308, pp. 32–36, mar 1984.

[43] J. Dubochet, "The Physics of Rapid Cooling and Its Implications for Cryoimmobilization of Cells," *Methods in Cell Biology*, vol. 79, pp. 7–21, 2007.

[44] S. H. Scheres, "Beam-induced motion correction for sub-megadalton cryo-EM particles.," *eLife*, vol. 3, p. e03665, aug 2014.

[45] W. Kuehlbrandt, "Cryo-EM enters a new era," *eLife*, vol. 3, p. e01963, aug 2014.

[46] P. Lu, X.-c. Bai, D. Ma, T. Xie, C. Yan, L. Sun, G. Yang, Y. Zhao, R. Zhou, S. H. W. Scheres, and Y. Shi, "Three-dimensional structure of human γ-secretase.," *Nature*, vol. 512, pp. 166–70, aug 2014.

[47] J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith, "SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields," *Journal of Structural Biology*, vol. 116, pp. 190–199, jan 1996.

[48] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "EMAN2: an extensible image processing suite for electron microscopy.," *Journal of structural biology*, vol. 157, pp. 38–46, jan 2007.

[49] J. de la Rosa-Trevín, J. Otón, R. Marabini, A. Zaldívar, J. Vargas, J. Carazo, and C. Sorzano, "Xmipp 3.0: An improved software suite for image processing in electron microscopy," *Journal of Structural Biology*, vol. 184, pp. 321–328, nov 2013.

[50] A. Steven, D. Belnap, A. Steven, and D. Belnap, "Electron Microscopy and Image Processing: An Essential Tool for Structural Analysis of Macromolecules," in *Current Protocols in Protein Science*, pp. 17.2.1–17.2.39, Hoboken, NJ, USA: John Wiley & Sons, Inc., nov 2005.

[51] J. C. Gilkey and L. A. Staehelin, "Advances in ultrarapid freezing for the preservation of cellular ultrastructure," *Journal of Electron Microscopy Technique*, vol. 3, no. 2, pp. 177–210, 1986.

[52] L. Gan and G. J. Jensen, "Electron tomography of cells," *Quarterly Reviews of Biophysics*, vol. 45, pp. 27–56, feb 2012.

[53] C. M. Palmer and J. Löwe, "A cylindrical specimen holder for electron cryotomography.," *Ultramicroscopy*, vol. 137, pp. 20–9, feb 2014.

[54] G. E. Soto, S. J. Young, M. E. Martone, T. J. Deerinck, S. Lamont, B. O. Carragher, K. Hama, and M. H. Ellisman, "Serial section electron tomography: a method for three-dimensional reconstruction of large structures.," *NeuroImage*, vol. 1, pp. 230–43, jun 1994.

[55] R. A. Crowther, D. J. DeRosier, and A. Klug, "The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 317, no. 1530, 1970.

[56] D. N. Mastronarde, "Dual-axis tomography: an approach with alignment methods that preserve resolution.," *Journal of structural biology*, vol. 120, pp. 343–52, dec 1997.

[57] T. R. Shaikh, H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith, and J. Frank, "SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs," *Nature Protocols*, vol. 3, pp. 1941–1974, dec 2008.

[58] J. B. Heymann, G. Cardone, D. C. Winkler, and A. C. Steven, "Computational resources for cryo-electron tomography in Bsoft.," *Journal of structural biology*, vol. 161, pp. 232–42, mar 2008.

[59] S. Nickell, F. Förster, A. Linaroudis, W. D. Net, F. Beck, R. Hegerl, W. Baumeister, and J. M. Plitzko, "TOM software toolbox: acquisition and analysis for electron tomography," *Journal of Structural Biology*, vol. 149, pp. 227–234, mar 2005.

[60] T. Hrabe, Y. Chen, S. Pfeffer, L. Kuhn Cuellar, A.-V. Mangold, and F. Förster, "PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis," *Journal of Structural Biology*, vol. 178, pp. 177–188, may 2012.

[61] V. Lučič, A. Rigort, and W. Baumeister, "Cryo-electron tomography: the challenge of doing structural biology in situ.," *The Journal of cell biology*, vol. 202, pp. 407–19, aug 2013.

[62] F. Forster, O. Medalia, N. Zauberman, W. Baumeister, and D. Fass, "Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 4729–4734, mar 2005.

[63] H. Zhu and P. Zhu, "No longer 'blob-ology': Cryo-EM is getting into molecular details.," *Science China. Life sciences*, vol. 58, pp. 1154–6, nov 2015.

[64] M. T. J. Smith and J. L. Rubinstein, "Beyond blob-ology," *Science*, vol. 345, no. 6197, 2014.

[65] R. N. Clough, G. Moldovan, and A. I. Kirkland, "Direct Detectors for Electron Microscopy," *Journal of Physics: Conference Series*, vol. 522, p. 012046, jun 2014.

[66] R. S. Ruskin, Z. Yu, and N. Grigorieff, "Quantitative characterization of electron detectors for transmission electron microscopy.," *Journal of structural biology*, vol. 184, pp. 385–93, dec 2013.

[67] S. J. Ludtke, M. L. Baker, D.-H. Chen, J.-L. Song, D. T. Chuang, and W. Chiu, "De novo backbone trace of GroEL from single particle electron cryomicroscopy.," *Structure (London, England : 1993)*, vol. 16, pp. 441–8, mar 2008.

[68] S. M. Stagg, G. C. Lander, J. Quispe, N. R. Voss, A. Cheng, H. Bradlow, S. Bradlow, B. Carragher, and C. S. Potter, "A test-bed for optimizing high-resolution single particle reconstructions.," *Journal of structural biology*, vol. 163, pp. 29–39, jul 2008.

[69] R. Turchetta, J. Berst, B. Casadei, G. Claus, C. Colledani, W. Dulinski, Y. Hu, D. Husson, J. Le Normand, J. Riester, G. Deptuch, U. Goerlach, S. Higueret, and M. Winter, "A monolithic active pixel sensor for charged particle tracking and imaging using standard VLSI CMOS technology," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 458, no. 3, pp. 677–689, 2001.

[70] A.-C. Milazzo, P. Leblanc, F. Duttweiler, L. Jin, J. C. Bouwer, S. Peltier, M. Ellisman, F. Bieser, H. S. Matis, H. Wieman, P. Denes, S. Kleinfelder, and N.-H. Xuong, "Active pixel sensor array as a detector for electron microscopy," *Ultramicroscopy*, vol. 104, no. 2, pp. 152–159, 2005.

[71] A. Faruqi, R. Henderson, M. Pryddetch, P. Allport, and A. Evans, "Direct single electron detection with a CMOS detector for electron microscopy," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 546, no. 1, pp. 170–175, 2005.

[72] A.-C. Milazzo, A. Cheng, A. Moeller, D. Lyumkis, E. Jacovetty, J. Polukas, M. H. Ellisman, N.-H. Xuong, B. Carragher, and C. S. Potter, "Initial evalua-

tion of a direct detection device detector for single particle cryo-electron microscopy.," *Journal of structural biology*, vol. 176, pp. 404–408, dec 2011.

[73] H. Y. Liao and J. Frank, "Definition and estimation of resolution in single-particle reconstructions.," *Structure (London, England : 1993)*, vol. 18, pp. 768–75, jul 2010.

[74] P. A. Penczek, "Resolution measures in molecular electron microscopy.," *Methods in enzymology*, vol. 482, pp. 73–100, 2010.

[75] X. Zhang and Z. Hong Zhou, "Limiting factors in atomic resolution cryo electron microscopy: No simple tricks," *Journal of Structural Biology*, vol. 175, no. 3, pp. 253–263, 2011.

[76] H. G. M and van Heel, "Direct 3D reconstruction from projections with initially unknown angles.," in *Pattern Recognition in Practice II* (K. L. Gelsema ES, ed.), pp. 279–288, Amsterdam: North-Holland Pubishing, 1986.

[77] G. Harauz and M. van Heel, "Exact filters for general geometry three dimensional reconstruction," 1986.

[78] W. O. Saxton and W. Baumeister, "The correlation averaging of a regularly arranged bacterial cell envelope protein.," *Journal of microscopy*, vol. 127, pp. 127–38, aug 1982.

[79] M. van Heel and M. Stöffler-Meilicke, "Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope.," *The EMBO journal*, vol. 4, pp. 2389–95, sep 1985.

[80] N. Grigorieff, "Resolution measurement in structures derived from single - Technische Informationsbibliothek (TIB)," *Acta Crystallographica Section D Biological Crystallography*, vol. 56, pp. 1270–1277, 2000.

[81] D. M. Blow and F. H. C. Crick, "The Treatment of Errors in the Isomorphous Replacement Method," *Acta Cryst. Phys. Rev. Z. KristaUogr. Aeta Cryst*, vol. 12, no. 1953, pp. 794–801, 1959.

[82] V. Y. Lunin, M. M. Woolfson, and IUCr, "Mean phase error and the map-correlation coefficient," *Acta Crystallographica Section D Biological Crystallography*, vol. 49, pp. 530–533, nov 1993.

[83] R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schröder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, and C. L. Lawson, "Outcome of the first electron microscopy validation task force meeting.," *Structure (London, England : 1993)*, vol. 20, pp. 205–14, feb 2012.

[84] P. B. Rosenthal and R. Henderson, "Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy," *Journal of Molecular Biology*, vol. 333, pp. 721–745, oct 2003.

[85] C. A. DIEBOLDER, A. J. KOSTER, and R. I. KONING, "Pushing the resolution limits in cryo electron tomography of biological structures," *Journal of Microscopy*, vol. 248, pp. 1–5, oct 2012.

[86] P. A. Penczek, C. Yang, J. Frank, and C. M. T. Spahn, "Estimation of variance in single-particle reconstruction using the bootstrap technique.," *Journal of structural biology*, vol. 154, pp. 168–83, may 2006.

[87] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, "Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization.," *Nature methods*, vol. 4, pp. 27–9, jan 2007.

[88] Q. Jin, C. O. S. Sorzano, J. M. de la Rosa-Trevín, J. R. Bilbao-Castro, R. Núñez-Ramírez, O. Llorca, F. Tama, and S. Jonić, "Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes.," *Structure (London, England : 1993)*, vol. 22, pp. 496–506, mar 2014.

[89] A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dinman, J. Frank, and A. Ourmazd, "Trajectories of the ribosome as a Brownian nanomachine.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 17492–7, dec 2014.

[90] J. Anden, E. Katsevich, and A. Singer, "Covariance estimation using conjugate gradient for 3D classification in CRYO-EM," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 200–204, IEEE, apr 2015.

[91] H. D. Tagare, A. Kucukelbir, F. J. Sigworth, H. Wang, and M. Rao, "Directly reconstructing principal components of heterogeneous particles from cryo-EM images.," *Journal of structural biology*, vol. 191, pp. 245–62, aug 2015.

[92] G. Bricogne, "Methods and programs for direct-space exploitation of geometric redundancies," *Acta Crystallographica Section A*, vol. 32, pp. 832–847, sep 1976.

[93] W. Wriggers, R. A. Milligan, and J. McCammon, "Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy," *Journal of Structural Biology*, vol. 125, pp. 185–195, apr 1999.

[94] 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.," *Journal of computational chemistry*, vol. 25, pp. 1605–12, oct 2004.

[95] M. G. Rossmann, "Fitting atomic models into electron-microscopy maps.," *Acta crystallographica. Section D, Biological crystallography*, vol. 56, pp. 1341–9, oct 2000.

[96] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints.," *Journal of molecular biology*, vol. 234, pp. 779–815, dec 1993.

[97] A. Jack and M. Levitt, "Refinement of large structures by simultaneous minimization of energy and R factor," *Acta Crystallographica Section A*, vol. 34, pp. 931–935, nov 1978.

[98] A. T. Brünger, J. Kuriyan, and M. Karplus, "Crystallographic R factor refinement by molecular dynamics.," *Science (New York, N.Y.)*, vol. 235, pp. 458–60, jan 1987.

[99] 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 (London, England : 1993)*, vol. 16, pp. 673–83, may 2008.

[100] M. M. Tirion, "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis," *Physical Review Letters*, vol. 77, pp. 1905–1908, aug 1996.

[101] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," 1997.

[102] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, "Global Dynamics of Proteins: Bridging Between Structure and Function," *Annual Review of Biophysics*, vol. 39, pp. 23–42, apr 2010.

[103] G. F. Schröder, A. T. Brunger, and M. Levitt, "Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution," *Structure*, vol. 15, pp. 1630–1641, dec 2007.

[104] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren, "Crystallography & NMR system: A new software suite for macromolecular structure determination.," *Acta crystallographica. Section D, Biological crystallography*, vol. 54, pp. 905–21, sep 1998.

[105] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy,

N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "PHENIX : a comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 213–221, feb 2010.

[106] S. Saalfeld, R. Fetter, A. Cardona, and P. Tomancak, "Elastic volume reconstruction from series of ultra-thin microscopy sections," *Nature Methods*, vol. 9, pp. 717–720, jun 2012.

[107] C. O. Sanchez Sorzano, A. L. Alvarez-Cabrera, M. Kazemi, J. M. Carazo, and S. Jonić, "StructMap: Elastic Distance Analysis of Electron Microscopy Maps for Studying Conformational Changes.," *Biophysical journal*, vol. 110, pp. 1753–65, apr 2016.

[108] B. L. de Groot, D. M. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. Berendsen, "Prediction of protein conformational freedom from distance constraints.," *Proteins*, vol. 29, pp. 240–51, oct 1997.

[109] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, "Stereochemical quality of protein structure coordinates," *Proteins: Structure, Function, and Genetics*, vol. 12, pp. 345–364, apr 1992.

[110] A. T. Brünger, "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures," *Nature*, vol. 355, pp. 472–475, jan 1992.

[111] B. Falkner and G. F. Schroder, "Cross-validation in cryo-EM-based structural modeling," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 8930–8935, may 2013.

[112] S. Yang, X. Yu, V. E. Galkin, and E. H. Egelman, "Issues of resolution and polymorphism in single-particle reconstruction.," *Journal of structural biology*, vol. 144, no. 1-2, pp. 162–71, 2003.

[113] D. Sousa and N. Grigorieff, "Ab initio resolution measurement for single particle structures," *Journal of Structural Biology*, vol. 157, no. 1, pp. 201–210, 2007.

[114] A. Stewart and N. Grigorieff, "Noise bias in the refinement of structures derived from single particles.," *Ultramicroscopy*, vol. 102, pp. 67–84, dec 2004.

[115] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, pp. 187–217, jan 1983.

[116] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, "An all atom force field for simulations of proteins and nucleic acids," *Journal of Computational Chemistry*, vol. 7, pp. 230–252, apr 1986.

[117] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren, "The GROMOS Biomolecular Simulation Program Package," *The Journal of Physical Chemistry A*, vol. 103, pp. 3596–3607, may 1999.

[118] H. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, pp. 43–56, sep 1995.

[119] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kale, R. D. Skeel, and K. Schulten, "NAMD: a Parallel, Object-Oriented Molecular Dynamics Program," *International Journal of High Performance Computing Applications*, vol. 10, pp. 251–268, dec 1996.

[120] B. Isralewitz, M. Gao, and K. Schulten, "Steered molecular dynamics and mechanical functions of proteins," *Current Opinion in Structural Biology*, vol. 11, pp. 224–230, apr 2001.

Results

This thesis aims to develop various methods needed for better structure determination of the cryo-EM data. Three individual algorithms were developed in this thesis and are presented below:

(1) **ELIGN: Elastic alignment of Cryo-EM density maps**
Biomolecular complexes can be studied under near-native physiological conditions using cryo-EM. Most of the density maps obtained from cryo-EM experiments are limited in resolution by conformational heterogeneity. Improving the resolution of the density maps, thus, requires to account for the structural heterogeneity. The resolution can often be improved when sorting the images into classes of similar conformations and reconstructing each class separately. However, sorting leads to a lower number of images per reconstruction, which again lowers the quality of the reconstruction. To address this problem, density maps of such different conformational classes need to be averaged. Averaging can be expected to lead to better signal-to-noise ratio and therefore to higher resolution. However, averaging density maps of different conformations needs to allow for elastic alignment of a density map into another density map. By elastically fitting the density of one conformation into another conformation, such that individual domains are well superimposed, different density maps can be aligned better. The density grid will be treated as a semi-flexible body which enables flexible alignment and averaging during the density alignment. The flexible density map alignment effectively reduces the variance between the image classes and allows for merging different classes to benefit from the higher number of images per reconstruction, with the aim to improve the resolution.

The manuscript is close to being submitted. My contribution: Complete implementation of the method, analysis of all data and writing of the manuscript.

(2) **VISDEM: Improving the Visualisation of Cryo-EM Density Reconstructions**

Appropriate visualization is important for the proper interpretation of a density map. VISDEM is a sharpening algorithm which uses the general statistical packing density information of proteins. This helps to build a pseudo atomic model from which the constraints in both the real space - density histogram and the Fourier space - radial structure factor are estimated. VISDEM is not model biased, because only the density reconstruction and statistical information of proteins are used to build the pseudo-atomic model. The result shows that the method yields sharpened maps that are better than the common B-factor sharpening approach. The results are also as good as sharpening using the correct structure factor and density histogram from a crystal structure.

This work was published in *Journal of Structural Biology , vol. 191, pp. 207 - 213 (2015)*. My contribution: Its an equal contribution by myself and Michaela Spiegel. Both performed calculations and analysis of all the data and also wrote the manuscript together.

(3) **Estimating positional precision in real space refinement**

There are several approaches available for estimating the precision of atomic structures for refinement against structure factors from X-ray crystallography. With the enormous growth of the cryo-EM field in the recent years, there is a need for determining both accuracy and precision of atomic models built from cryo-EM data. In particular, when the data is in the low- to intermediate-resolution range, the atomic models obtained from the refinement are prone to errors. Thus, a new method was developed to estimate the positional precision in real space refinement. The method uses the cross-validation value available in the software DireX to find the best ensemble to calculate the coordinate error.

The manuscript is in preparation. My contribution: Complete execution, analysis of all simulations and writing of the manuscript. The manuscript is 80% finished.

The following three chapters in this thesis include the above three manuscripts either in published or in preparation formats.

Chapter 2

ELIGN: Elastic alignment of Cryo-EM density maps

A. K. Duraisamy[a,1], G. F. Schröder[a,b 1]

[a] Institute of Complex Systems (ICS-6), Structural Biochemistry, Forschungszentrum Jülich, 52428 Jülich, Germany

[b]Physics Department, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[1]Corresponding author : E-mail address: gu.schroeder@fz-juelich.de (G.F. Schröder)

author contributions: G.F.S. designed research; A.K.D. and G.F.S. performed research; and A.K.D and G. F. S. wrote the paper.

author declaration: The authors declare no conflict of interest.

Keywords
Cryo-EM, Single particle analysis, Cryo-ET, ELIGN, Rigid averaging, Bead model

Significance statement
In cryo-electron microscopy, the conformational heterogeneity is a boon as well as a bane because on the one hand side, it severely limits the resolution, but on the other hand side it provides deep inside into the functions of proteins and protein assemblies. Nevertheless, higher resolution can be achieved, if the conformational differences between the different classes are accounted for. We have developed a new method to elastically align 3D density maps to another by flexibly deforming the density grid. Our method, thus, accounts for global conformational differences and allows for averaging all image data and thereby overcomes the resolution barrier imposed by the heterogeneity. With our method, classifying the images further is beneficial since it can be recombined later with elastic alignment.

ABSTRACT
Cryo-electron microscopy (Cryo-EM) is a technique to study the structure of large biomolecular assemblies, which are often flexible and conformationally heterogeneous. Conformational heterogeneity severely limits the resolution of density reconstructions in both single-particle analysis as well as subtomogram averaging. The resolution is typically improved by sorting the images (or volumes) into classes of similar conformations and reconstructing each class separately. However, sorting leads to a lower number of images per class, which again lowers the quality of the reconstructions. We present a method to average the density maps of such different classes (conformations), which further improves the resolution due to an improved signal-to-noise ratio. The method treats a density map as a semi-flexible body and allows for global deformations such that conformationally different structures can be well superimposed. Here, we demonstrate the attainable resolution improvement with the simulated data as well as experimental single-particle and tomographic data.

2.1 Introduction

Single particle cryo-EM is a popular technique to study the structure of large biomolecules. However, in the single particle cryo-EM method, the resolution is limited by the signal-to-noise ratio, due to the low dose of electron radiation, as the biomolecules are very sensitive to electron radiation. Therefore, the high-resolution (i.e. high signal-to-noise ratio) can only be reached by aligning and averaging a sufficiently large number of particle images during the 3D reconstruction process [1]. Similarly, in cryo-electron tomography, the resolution of specific particles identified in tomograms can often be improved by subtomogram alignment and averaging. The alignment of the images (or volumes) is typically achieved by rigid global transformations.

A major challenge for both single particle analysis and tomography is the conformational heterogeneity: If the particles are in different conformations the resolution reached after averaging over these particles will necessarily be limited by the heterogeneity. Improving the resolution of the density maps, thus, requires to account for the structural heterogeneity. The typical remedy is to sort images into different classes of similar conformations and to reconstruct each class separately. Image classification is therefore an important step in all cryo-EM image processing protocols, for which many methods have been developed [2, 3].

However, sorting the images leads to a lower number of images per class. Thus, the resolution improvement due to the purification of conformations within the classes is countered by the reduction in image statistics per class. Ideally, one should make use of all (or at least most) of the image data for an optimal signal-to-noise ratio. It is obviously not useful to rigidly average the individual classes, because the advantage of classification would be lost. Fortunately, conformational variance in proteins (and more generally in biomolecular complexes) is not uniformly distributed, but to a large extent the result of global collective motions, which are mainly responsible for the large-scale conformational heterogeneity. In addition, these collective motions are often governed by only a few degrees of freedom, for example, hinge regions between relatively rigid but mobile domains [4–6].

Globally different conformations, thus, can be locally very similar within these

domains. It can therefore be expected that local alignment and averaging of these domains improve the signal-to-noise ratio and consequently also the local resolution. An obvious choice is to treat the density map as an elastic body. Alignment of entire density maps of different conformations can be achieved by treating the density map as an elastic body, which has been shown before [7]. Resolution can of course only be improved, if the maps are sufficiently similar on the local level. If a conformational change leads to different side-chain packing, side-chain rotamers etc., then the local resolution in these regions cannot be improved.

We have developed a method to elastically align the entire density map similar to elastic image registration in the field of medical imaging-PET/CT scan [8, 9], computer graphics-morphing and warping [10, 11], virtual reality [12], remote sensing [13] and elastic volume reconstruction in the field of serial-section microscopy [14]. Already existing elastic deformation methods in cryo-EM such as 3DEM-Loupe [15] computes normal modes on density maps by defining springs between voxels. The iterative 3D-to-2D alignment method known as HEMNMA [16, 17], first predicts the possible motions by normal mode analysis (NMA) and then computes amplitudes of the conformational ensemble. It therefore relies on relevant motions to be sampled by NMA. Recently, StructMap [7] was developed based on the iterative 3D-to-2D alignment method. StructMap gives a graphic visualization of difference of maps based on elastic transformation for conformational modeling and performs a qualitative analysis on the flexibility of the EM maps using elastic deformation of the bead model, but is not used for resolution improvement.

Here, we present a method called ELIGN (**EL**astic al**IGN**ment) which enables to elastically align a density map of one conformation to a density map of another conformation and to optimally superimpose all local regions (domains) at the same time. The density grid is treated as a semi-flexible body during the density alignment and the elastically aligned maps are averaged to enhance the resolution. We demonstrate the resolution improvement by the ELIGN method on three datasets: 1) a test case with simulated data of the glutamate dehydrogenase in different conformations, 2) experimental single-particle data of the E-coli ribosome complex, and 3) experimental tomography data of the ribosome-OST-TRAP complex.

2.2 Results

The ELIGN algorithm was developed to improve the resolution of the cryo-EM density maps by elastically aligning and averaging the reconstructed 3D maps after extensive sorting of the images into different classes. Presently, ELIGN is applied as a post-processing technique used after the 3D reconstruction of volumes from 2D images. In this method, the density map that has to be elastically aligned is called the start density map. The density map that the start map has to be fitted to, is called the target density map. The output of ELIGN will be a resolution improved density map which combines all the images effectively into one single density map. In other words, the map resolution will be improved for the same conformation as that of the target density map, assuming that the local domains in the more than two different conformations are identical.

2.2.1 ELIGN

The main steps of the ELIGN method are:

1. Filter start density map to low resolution.
2. Translate density map to model of point masses on density grid (referred to as 'pseudo-atomic' or 'bead model') and define harmonic distance restraints between neighboring point masses.
3. Flexibly fit bead model to filtered target density map.
4. Replace low-resolution density values in the elastically aligned map with values from original high resolution density map at corresponding voxels.
5. Normalize and FSC weight the elastically aligned high resolution density map.
6. Average elastically aligned density and target density.

To elastically fit a start density map to a target density map, the start density map has to be transformed to a form, that allows for easily describing map deformations. For this, the density grid points (voxels) are replaced by pseudo-atoms (or

beads) with a mass proportional to the corresponding density values of the voxels. For deciding where to place the beads, a density threshold in the start density is chosen such that the beads cover the region of the molecule. No beads are placed where there is negligible or low density. The start density is filtered to a resolution of typically 2–3 times lower than the resolution of the start density map. The purpose of filtering the map is to define a bigger enclosed volume for placing the beads. If one would use the volume defined by the unfiltered start map, the volume mask around the atomic structure could be too tight. A tight mask would lead to artefacts of the density on the surface of the particle. After creating a bead model with a specific threshold, a few layers of beads are typically added to further extend the outer region of the bead model to ensure the bead model covers the entire region of the molecule.

The bead model is refined or elastically fitted with the program DireX against the target density map. DireX does not require the atomic models to be real molecular structures, but it is sufficiently generic to allow for refining bead models. Instead of modifying the mass, all pseudo-atoms are defined as carbon atoms where the occupancy is set equal to the density value. The pseudo-atomic model will hold the density values from the unfiltered map, hence, the cross-correlation between the unfiltered map and the map created from the bead model should be very high of about 0.99. Now the density map is in the same format, that can be elastically fitted in the same way as that of an atomic structure is refined against a density map. The bead model is very large when compared to the corresponding atomic structure. If the bead model is extremely large, then coarser grid are used for faster alignments. The pseudo-atom sampling can be coarser by placing beads only on every other grid point. The coarser bead model has only half the size compared to the full model and makes the flexible fitting approximately faster by a factor of two.

DireX is used to elastically fit the whole or coarser bead model of the start density to the target density. While placing the beads on the voxels, the distance restraints for the specific beads are also created by looking for the possible nearest neighbors located at specific distances defined by the voxel size of the density map. The beads are restrained to each other by connecting each bead to its immediate nearest neighbors and its second nearest neighbors. For example, a bead in the middle of the biomolecular complex would have 18 restraints and lesser in

the corners depending on the shape and density threshold used. A distant restraint acts like a spring between two beads, that can be changed from very flexible to very rigid connections. In DireX, the force which acts on an atom to move it into a density map is proportional to the density value in its proximity. Hence, during elastic fitting not all the beads experience the same amount of force but the force depends on the density value of the target map. The map force pulls the beads towards the target map and makes the bead to move away from the original grid position whereas the distant restraints take care that they are moving smoothly together without messing the fine structure that may lead to loss in high resolution information. During the elastic fitting, the noise dominated information in the high spatial frequency region or free band is used only for cross-validation and not for fitting [18]. The cross-validation helps to assess the quality of the fit and helps to avoid overfitting. If the elastic fitting is done with coarser bead mode,l then the output should be interpolated to get the whole bead model. From the elastically fitted whole bead model, a density map with the available highest resolution was computed using DireX. The final high resolution elastically bent density is normalized and also FSC-weighted by its own full map FSC function and averaged with the target density. Finally the structure factor from the model map computed from an available high resolution atomic structure is applied to the averaged density. If the two maps that are to be aligned have a different resolution, then we prefer to elastically deform the lower resolution map, since the deformation can lead to artefacts which will predominantly deteriorate the high resolution signal. If several density maps are aligned, the target density map to which all other maps are aligned should be the conformation with the improved resolution.

To assess the improvement of averaging the elastically aligned maps, the output of ELIGN is compared to the following density maps:

1. The target density. Although the target density map was reconstructed with only part of the images, this is the density map with the best density information for the specific conformation which can be obtained from any sorting method.

2. The rigidly averaged density maps, which are simply the rigidly fitted and averaged FSC weighted density maps of all conformations.

3. For the simulated dataset: The optimal averaged density, which is an average of the same number of density maps for the identical conformation with the same noise level (i.e. FSC curve) but different random number seed values. The optimal averaged density would be the best map obtained while averaging certain number of maps. For single particle data: The optimal maps is reconstructed using all-images which will be similar to the rigid averaging of all possible different conformations by assigning equal weightage on all the images.

4. In case of cryo-ET data, two half maps of the same data are compared. The half maps will be very similar, if it contains signal; the half maps will be different, if it has random noise.

2.2.2 Simulated Data: Glutamate dehydrogenase

The six near atomic resolution structures ranging from 3.2 Å to 3.6 Å for glutamate dehydrogenase (GDH), a metabolic enzyme of 336-kDa [19] was used to test the method. Since, six densities can be created from the structures in the same resolution range it is easy to show a clear improvement due to elastic averaging compared to the rigid averaging. The glutamate dehydrogenase structures in both the unliganded state (PDB-3JCZ) and in complex with GTP (PDB ID: 3JD0), complex with coenzyme NADH - closed and open conformation (PDB IDs 3JD1 and 3JD2, respectively) and complex with NADH and GTP- open and closed conformations (PDB-3JD3 and PDB-3JD4) are used to create six simulated density maps with a resolution of 4.3 Å. The simulated density in Fig. 2.1(a) for the PDB-3JCZ unliganded state, is the chosen target conformation to show the improvement in resolution. The other five density maps are the start maps which were elastically aligned and averaged with the target map.

The outer periphery of the complex, two left and two right regions are more flexible than the central core region. The flexible regions would be interesting to consider for testing ELIGN method since they cannot be improved by rigidly averaging the density maps. The rigid core region is shown in Fig. 2.1(b) and one of the four flexible peripheral regions as in Fig. 2.1(c). The density maps were filtered to 10 Å and the coarser bead models were created with a density threshold

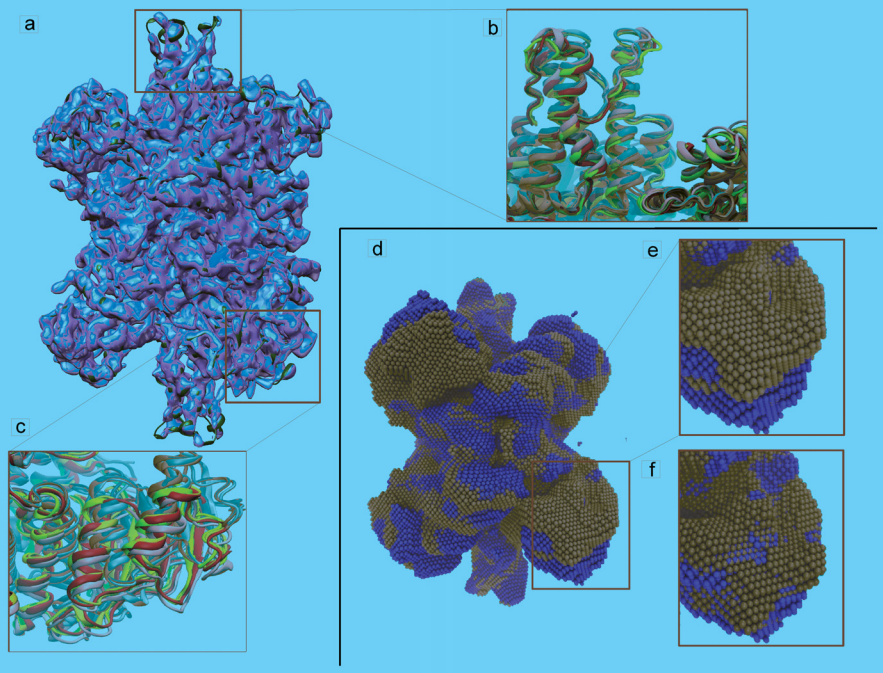

Figure 2.1 (a) Simulated target density with the atomic structure for the PDB-3JCZ. (b) Showing all the six atomic structures (target in green) in the core region which is not as flexible as the four peripheral regions. (c) One of the peripheral regions where the six conformations are more different. (d) Bead models for both the target in blue and one of the start density in grey. A part of the bead models enlarged in one of the flexible peripheral region (e) before and (f) after the elastic fitting.

of 0.84σ. The bead model shown in Fig. 2.1(d), shows the difference between the coarser target bead models in blue, and coarser start density, in grey. A part of the coarser bead models are enlarged to clearly show the difference in the flexible region before (Fig. 2.1(e)) and after (Fig. 2.1(f)) the elastic fitting. After the elastic fitting, the fitted coarser bead model is interpolated to the whole bead model and a density is created for this bead model including all the high-resolution information up to 2 Å. The final high resolution elastically fit density map is then normalized and FSC-weighted before averaging all the start densities and the target.

Fig. 2.2 shows the target map (blue) created for PDB-3JCZ and rigidly averages (grey) and the ELIGN (orange) of all the six start density maps. The output of

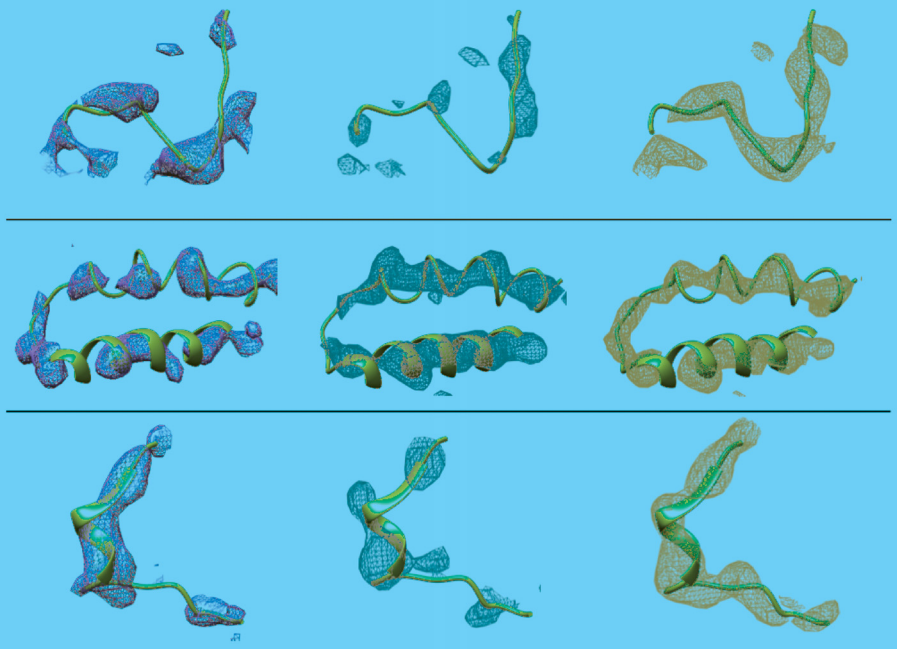

Figure 2.2 Showing details of the density map improvement due to ELIGN in different part of glutamate dehydrogenase. Shown are (a) the target map (blue), the rigidly averaged map (grey), and the output of the ELIGN (orange) with the PDB-3JCZ in the green trace. The output of the ELIGN (orange) clearly has better densities than the other density maps. The regions shown are from the outer periphery. Due to the flexibility of these regions, it is possible only with ELIGN to see an enhanced resolution whereas the rigid averaging will worsen the resolution. The threshold is chosen to give similar volume for all the maps.

ELIGN clearly shows an improvement in the resolution, when compared to the rigid average and the target map. All the figures in Fig. 2.2 are chosen from the outer regions of the enzyme, where the rigid averaging will not help to enhance the resolution, since they are the most flexible regions.

FSC curves were calculated between the model map created from the PDB-3JCZ at 2 Å and the ELIGN map, rigidly averaged, target map and optimal map. The FSC curve of the ELIGN (orange) show an improvement in resolution, when compared to rigid averaging (black) and target map (blue). The optimal FSC curve is the best improvement that one could get by averaging six maps at the specific

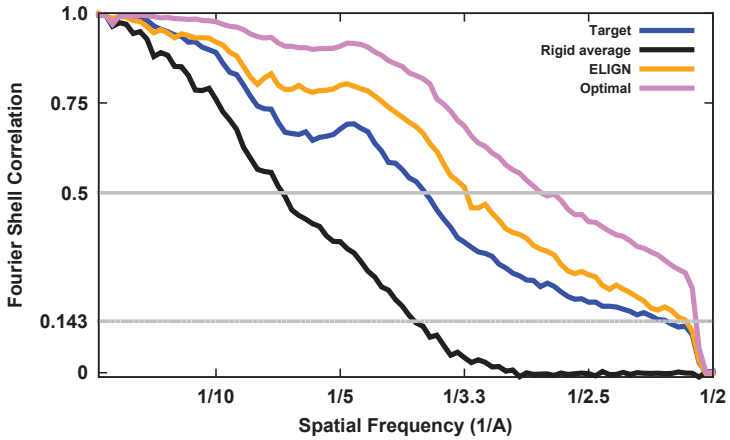

Figure 2.3 Various FSC curve for the simulated GDH density maps corresponding to target, rigid average, ELIGN and optimal averages. ELIGN shows an improvement in resolution when compared to the rigid averaging.

resolution. The target map FSC at 0.143 criterion is 2.17 Å, rigid fitting is at 3.6 Å, ELIGN is at 2.09 Å and the optimal FSC is at 2.05 Å.

2.2.3 Single-particle data: Ribosomal L1 stalk

The method was also tested on the five different density maps of the E-coli ribosome from the experimental data. The images were sorted for the first five eigenvectors belonging to the first principal motion by using principal component analysis technique on the L1 stalk which is one of the most flexible parts in the ribosome (done by Michaela Spiegel, results will be published elsewhere). The five density maps are different mostly in the regions functionally connected to the L1 stalk and the other regions are almost the same. Due to this, the output of the ELIGN shows a major improvement in resolution in the L1 stalk and its connected regions. The whole ribosome density maps were all in the resolution range from 6 Å, to 6.5 Å, but due to its high flexibility the L1 stalk has even lower resolution. The map with the higher resolution for the L1 stalk is chosen as the target map. Even though for the target map, roughly one-tenth of all images were used for the 3D reconstruction, this density map has the best density information in the L1 stalk which can be obtained from any sorting methods. Showing an improvement in resolution

in the L1 stalk region with ELIGN when compared to target density should also be interesting. Also, in this case the target or the highest resolution L1 density is the map created from the first eigenvector, hence it is in the middle. The density maps of second and third eigenvectors have two conformations on the left and two conformations on the right. Therefore, the middle conformation was chosen as the target map to prove that the method can work for elastically fitting the density maps from any direction. Since the sorting was based on principle component analysis, the model for sorting were slightly different so the two half maps are different from the usual half maps. The normal procedure of applying ELIGN on two half maps separately and averaging the ELIGN output of the two half maps will not give the best output for this dataset. Hence, one of the half maps belonging to the target conformation is chosen as the target map and all the other nine density maps are the start densities that can be elastically fitted. The elastically fitted maps are then normalized, FSC weighted and then averaged with the target density.

In Fig. 2.4(a) the target density with the atomic structure in green for the ribosome and insets Fig. 2.4(b) and Fig. 2.4(c) show one set of half maps for all the five conformations and thus include the target density and four start densities which are FSC filtered for better visibility. The inset Fig. 2.4(b) shows the L1 stalk from the five densities having target in the middle (purple). The figure shows clearly that the densities are mostly different only in the L1 region and the other regions are almost the same due to the sorting method based on the L1 motion. The Fig. 2.4(c) does not have a large motion as the L1 stalk but there is a small motion which cannot be perfectly fitted to each other rigidly, but rather was done only by ELIGN and the improvement for these regions is also shown in the Fig. 2.6. The start maps are filtered to 10 Å and coarser bead models are created with a density threshold of 1.57 σ. The density maps are very different in the L1 stalk and this can be seen in Fig. 2.4(d) having the coarser bead model for the target (purple) and one of the start densities (pink). For the case of ribosome, the coarser bead model was created from start map filtered to 10 Å. It has approximately 730,000 beads and the whole bead model is approximately 1.4 million beads. The coarser bead models enlarged on the L1 stalk before (Fig. 2.4(e)) and after (Fig. 2.4(f)) the elastic fitting. After interpolating the coarser bead model to the whole bead model, final elastically fit density map was created and normalized. The FSC weighting is done with full

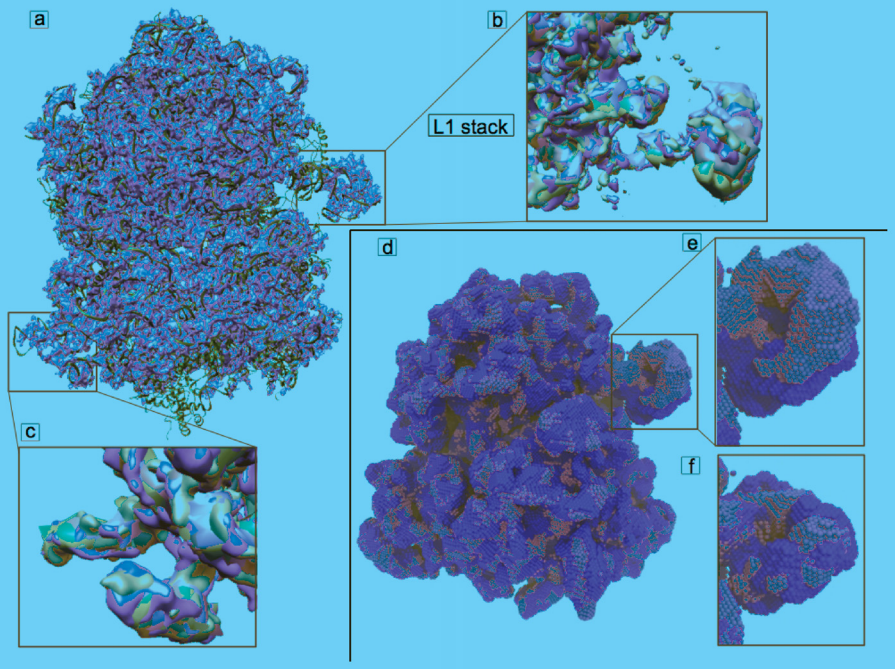

Figure 2.4 (a) The target density in purple and the atomic structure of the ribosome in green with the insets showing one set of half maps which includes the target density (purple) in the middle and other four start densities which were fsc filtered. The inset (b) clearly shows that the densities varies largely in the L1 stalk region since the data is sorted for the L1 motion. The inset (c) shows other region which is not very flexible as the L1 stalk but there is a small motion that cannot be achieved by rigid fitting which makes it also an interesting region for the ELIGN method. (d) shows the coarser bead models for the target in blue and one of the start density in pink. L1 stalk is enlarged to show the movement (e) before and (f) after elastic fitting.

map FSC function plus a 10% estimated correction, since the half maps are not very similar like the usual half maps due to the principal component analysis sorting technique. The sorting technique uses different reference structures for the two half maps.

The Fig. 2.5 shows the resolution improvement due to ELIGN in the L1 stalk when compared to all the other density maps. The purple map is the target map and has only 20% of the images used for the reconstruction but this would be the map with the best available information in the L1 stalk, since it is sorted on L1

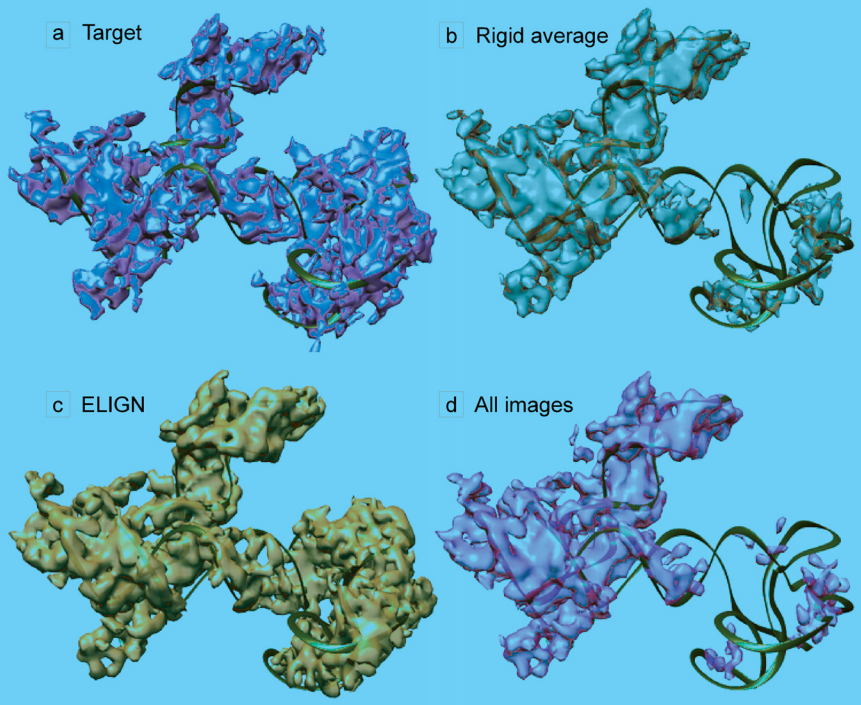

Figure 2.5 Showing details of the density map improvement with the ELIGN when compared to all other method in the L1 stalk of the ribosome which has the worst resolution due to its flexibility with the atomic structure (green) . a) The target density (purple) for which roughly 10% of the images were used for the reconstruction and which should have the best density due to sorting done for the correct conformation. b) Ten density maps that are FSC-weighted and rigidly averaged (grey). c) The output of the ELIGN (orange) for nine start maps and the target map. d) The density map (pink) reconstructed with all images for the target conformation. Comparing all density maps shows clearly an improvement in resolution with the ELIGN (orange) density. The density threshold in each map is chosen to enclose the same volume.

movement. The grey is the rigidly averaged density map and the pink is the density map which was reconstructed using all-images. The all-images reconstructed map will not be better in the flexible regions than rigid averaging since the images are weighted in the latter case. Also all-images reconstructed map will not be better than the target map due to correct sorting of images going into the reconstruction

of target conformation. Since the number of images and weighting factor for the images are same in both the ELIGN and the rigidly averaged densities, only the flexible regions will have an improvement in the ELIGN when compared to the rigid averaging. The non-flexible regions will have the same resolution in both ELIGN and rigid average. We can also see in the orange map, the ELIGN output the RNA base pairs are getting its shape and there is more density in the L1 stalk. The density map is better than rigid fitting not only in the very flexible regions but can also be in the parts with very small motion and it is shown in Fig: 2.6. The FSC curves (Fig: 2.6.e) were always calculated between the model map calculated at 3 Å from the atomic structure and the output from different methods. The purple, gray, pink and orange solid line curves are the FSC between the model map and the target map, rigidly averaged map, all-images reconstructed map and the ELIGN map respectively for the whole ribosome and the dashed line curves are the FSC for the cropped L1 stalk region. The FSC curves show a clear improvement due to ELIGN and measure as 3.27 Å when compared to the target 4.5 Å and rigid averaging and all-images 3.57 Å.

2.2.4 Tomography data: OST and TRAP

ELIGN was also tested on subtomogram averages of experimental cryo-electron tomography data. The data is a clipped region from the mammalian oligosaccharyl-transferase complex in native endoplasmic reticulum protein translocon [20, 21]. The clipped region contains only the translocon associated protein complex (TRAP) and the oligosaccharyl-transferase (OST) complex. We got three different sets of classes of the subtomograms with different signal-to-noise ratio. The same images are sorted and the subtomogram averages are classified to 170 classes, 30 classes and 10 conformational classes. The FSC computed from the half maps gives the resolution of the data, $\sim$16 Å. Since the number of class averages are different, the target map in each set of classes is not exactly the same. The target map has the best class weightage in each set.

The start maps, shown in supporting information Fig: 2.9, do not have too large conformational motion as in the case of single particle data. The bead models are created using a density cut-off at 1.5 σ for 40 Å filtered map and holding density values from the original maps. The bead models also have additional three layers

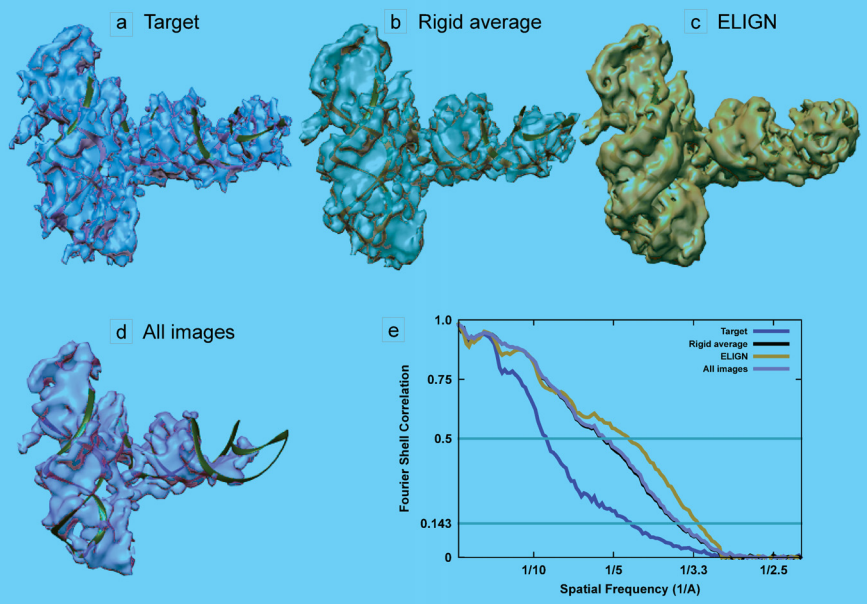

Figure 2.6 Density map improvement with ELIGN in comparison to all other method showing residues of chain a of the E-coli ribosome (PDB ID 5AFI) shown in Fig: 2.4.c that is not as flexible as the L1 stalk. a) The blue is the target map, grey is the FSC weighted rigidly averaged map and the orange is the output of the ELIGN. The pink is the all images reconstructed density map. FSC curve showing the improvement both for the whole ribosome and only in the L1 stalk region using ELIGN. The density thresholds for all maps is chosen to enclose the same volume.

of beads on the outer regions, so have more volume around the protein complex. The number of beads in the bead models are approximately 200,000. Due to the fact that, the whole bead models are not as large as in the case of ribosome data, these are directly used for flexible fitting rather than using coarser bead model and later interpolating. After elastically fitting the start maps to the target, weighted averages are computed using class weights. In addition, weighted averages were computed from rigidly fitted density maps. The averaged maps are masked with a mask map filtered to 100 Å . Since no atomic model exists for the TRAP and OST, the structure factor from the deposited EMD-3071 is applied on the masked maps. In all the three datasets, the resolution was improved to 12 Å by using ELIGN. This resolution is still lesser than 8 Å needed to view individual α-helix and 6 Å for

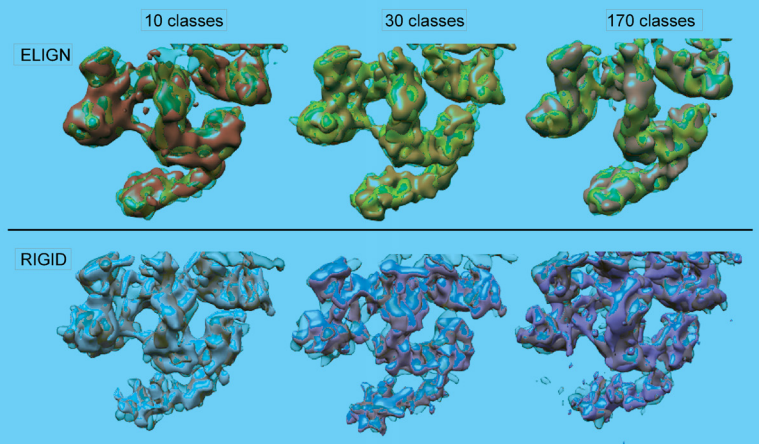

Figure 2.7 Half maps from ELIGN and rigid averages in 10, 30 and 170 classes. All the even maps (green) are overlapped with the corresponding odd maps (different shades of orange) from ELIGN. The even maps (grey) are overlapped with corresponding odd maps (different shades of blue) from rigid averaging.

β-sheets. Hence, half maps were compared to see the similarity in the case of both ELIGN and rigid fitting. In this case, half maps are not created by splitting the 2D images into two sets, but rather by averaging half of the sub-tomogram densities in each class. The half maps from ELIGN (Fig: 2.7-ELIGN) are more similar in all three cases. The green maps are even maps and corresponding different shades of orange are odd maps from different classes. The rigid averages (Fig: 2.7-RIGID) even maps (grey) and corresponding odd maps (different shades of blue) are not as similar as the ELIGN maps.

Not only the similarities in the half maps are better in ELIGN, also the FSC curves in Fig: 2.8 computed between half maps of ELIGN (solid lines) show an increase in resolution compared with rigid averages (dashed lines). ELIGN should work the best for 170 classes (red solid curve) than others datasets (orange and black solid curves), as they should contain images from single conformation. But ELIGN leads to same improvement in the resolution to 12 Å in all the three cases. This could mean that the data contains only 10 different conformations and 30

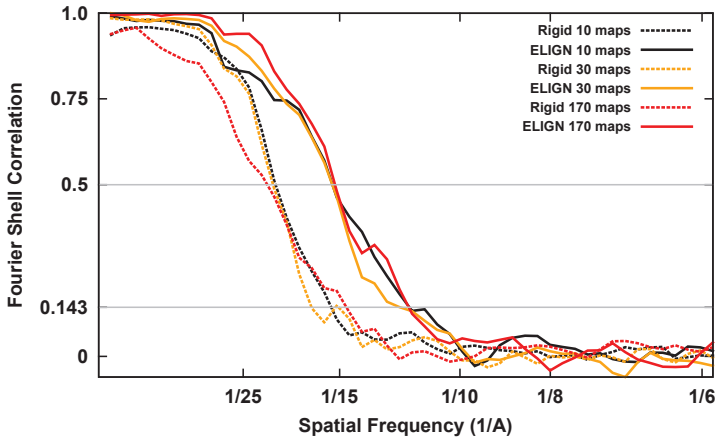

Figure 2.8 FSC curves computed between even and odd maps of rigid averages (dashed lines) and ELIGN (solid lines) for 10 classes (black), 30 classes (orange) and 170 classes(red).

classes and 170 classes are more classifications of the 10 conformations or the improvement is limited by the information in the data.

2.3 Discussion

ELIGN is a method for elastically aligning density maps of different conformations by elastic deformation with the goal of enhancing the resolution. Instead of rigidly averaging density maps of many rigid domains, ELIGN flexibly bends the whole density map in one step. The advantage in flexibly bending the whole map is to avoid the step of cutting the map into smaller rigid domains and later combining all the domains after rigid fitting to obtain the full density map, which leads to artefacts. Also the alignment error depends on the size of the rigid domains. The smaller the rigid domains, the larger the alignment errors. Thus, fitting the whole map reduces the alignment errors. ELIGN is a robust method since it is not limited by the number of start density maps, the direction of bending of the density maps or the difference in conformational changes. Also there is no need to have prior knowledge of the atomic structure for different conformations, since only the density maps and the pseudo atomic model created from the maps are used. With

the ELIGN method, it could be useful to use more class averages than normally and then to flexibly average all the 3D maps to reach significantly higher resolution. The method is needed, when there is strong conformational heterogeneity in the sample, this would normally limit the resolution. With ELIGN, this resolution limit is now defined by the similarity of individual domains and not by the similarity of the entire structure. ELIGN is a good tool to use for intermediate- or low-resolution data, especially for cryo-tomograms, which typically contain a few hundred low-resolution sub-tomogram densities. The averaging of similar domains is analogous to NCS averaging, which is used to improve the phase information in the field of X-ray crystallography [22]. Therefore, ELIGN is also applicable to averaging electron density maps from X-ray diffraction experiments and in particular to crystal averaging. As an extension of ELIGN, now used on 3D volumes, one could also apply the method directly on the 2D class averages and do a 3D reconstruction with all the elastically fitted images to reach high resolution 3D density.

Supporting Information (SI)

Preparation of Density Map The density map is filtered to a resolution 2-3 times lower than the original resolution, which defines the volume to place the beads inside. The beads are placed at every voxel and carries the density information from the unfiltered map. Every alternating bead is omitted unless the data is too small which is not computationally expensive. While creating the bead model, either whole or coarser bead model, the distance restraints used while fitting is also generated.

Refinement of bead models with DireX We describe in detail the DireX elastic bending in this section. During the elastic fitting by DireX, the bead mode is switched on. Distance restraint strength is the force which acts like a spring between the connected beads. The value 0 corresponds to no restraint which will allow the spring to stretch or compress to any length and 1 keeps the connected beads at fixed length which is defined by the apix of the density. Even though the parameters have to be optimized, the distance restraint which is named as the experimental distance restraint strength could be used around 0.4. The beads also experience forces from the target density known as map strength. The map strength force is an important parameter that should be optimized and could be started from

the value 0.01. The more the map strength value, the force is stronger which makes the beads to move faster, thus making the fitting converge faster than using low map strength value. Using a very high map strength greater then 0.3 is not advisable since it will make the bead model very unstable. The fitting parameters has to be optimized, specially the forces on the beads like the map strength and the experimental distance restraint strength. The parameter, compute-map using occupancy, is set to yes to compute the output density with the same density value from start map after each step.

For the cross validation in DireX, C_{free} band values should typically have the resolution of the density as the upper limit and the lower limit could be 2 or 3 Å lesser than the upper limit. Even though there is no strict rule on choosing the C_{free} interval, using a upper limit that is too low than the resolution of the map will not have enough information for validation and will lead to a very low C_{free} value. It is the trend of the C_{free} curve to be checked, that the curve converges, than the actual value of the C_{free} itself. The C_{free} value would increase with the number of steps and would converge once the fitting is done. If the model is over-fitted, the curve would decrease after converging. If there is a huge conformational change, the curve will not just increase once and converge but rather does few oscillations depending on the difference and would finally converge when the model fits to density.

Averaging of Aligned Density Maps In case the coarser bead model was used, after the elastic fitting, the output is interpolated to the whole bend bead model. The standard 3D interpolation algorithms cannot be used due to the fact that the shape of the bead model depends on the threshold chosen and will never be a regular cubic model. The interpolation is done by finding the missing bead by comparing x,y,z coordinates with the start whole bead model and listing the nearest neighbor for the missing bead. Searching for the listed nearest neighbors in the coarser bend model and finding the mean of those in the elastically fitted bead model will give the coordinates values for the missing bead. At least two nearest neighbor is needed to find the mean if not the bead is omitted or not written to the whole bend bead model. There are very few, less than one percent omitted beads during the interpolation. Missing such few beads does not cause any relevant information loss. The density value of the beads is used from the unfiltered whole start bead model. With the

whole bent bead model, high resolution density map can be computed.

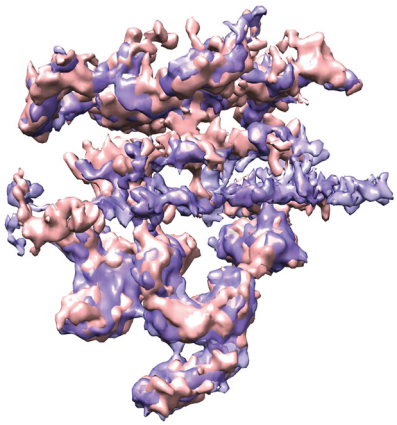

Figure 2.9 Cryo-ET data: One of the start density (pink) and target density (purple).

Cryo-ET data The figure 2.9 gives a better picture of the difference between the two conformations, one of the start conformation in pink and the target map in purple.

2.4 Acknowledgement

We thank Dr. Niels Fischer and Prof. Holger Stark for providing the E.coli ribosome data and Ms. Michaela Spiegel for doing further classification of the E-coli data. We also thank Dr. Stefan Pfeffer and Prof. Friedrich Fröster for sharing the cryo-ET data. We would like to acknowledge the financial support from the International NRW Research School BioStruct, granted by the Ministry of Innovation, Science and Research of the State North Rhine-Westphalia, the Heinrich Heine University of Düsseldorf, and the Entrepreneur Foundation at the Heinrich Heine University of Düsseldorf. Computational support and infrastructure was provided by the "Centre for Information and Media Technology" (ZIM) at the University of Düsseldorf (Germany).

Bibliography

[1] R. Henderson, "The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules," *Quarterly Reviews of Biophysics*, vol. 28, p. 171, may 1995.

[2] M. van Heel and J. Frank, "Use of multivariates statistics in analysing the images of biological macromolecules," *Ultramicroscopy*, vol. 6, pp. 187–194, jan 1981.

[3] N. Bonnet, "Artificial intelligence and pattern recognition techniques in microscope image processing and analysis," in *Advances in Imaging and Electron Physics*, vol. 114, pp. 1–77, 2000.

[4] M. F. Thorpe, M. Lei, A. J. Rader, D. J. Jacobs, and L. A. Kuhn, "Protein flexibility and dynamics using constraint theory.," *Journal of molecular graphics & modelling*, vol. 19, no. 1, pp. 60–9, 2001.

[5] H. Gohlke and M. Thorpe, "A Natural Coarse Graining for Simulating Large Biomolecular Motion," *Biophysical Journal*, vol. 91, pp. 2115–2120, sep 2006.

[6] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *Physical Biology*, vol. 2, pp. S127–S136, nov 2005.

[7] C. O. Sanchez Sorzano, A. L. Alvarez-Cabrera, M. Kazemi, J. M. Carazo, and S. Jonić, "StructMap: Elastic Distance Analysis of Electron Microscopy Maps for Studying Conformational Changes.," *Biophysical journal*, vol. 110, pp. 1753–65, apr 2016.

[8] Y.-C. Tai, K. P. Lin, C. Hoh, S. Huang, and E. Hoffman, "Utilization of 3-d elastic transformation in the registration of chest x-ray ct and whole body pet," in *IEEE Nuclear Science Symposium. Conference Record*, vol. 3, pp. 1903–1907, IEEE, 1996.

[9] C. Cohade and R. L. Wahl, "Applications of positron emission tomography/computed tomography image fusion in clinical positron emission tomography—clinical use, interpretation methods, diagnostic improvements," *Seminars in Nuclear Medicine*, vol. 33, pp. 228–237, july 2003.

[10] T. Beier and S. Neely, *Feature-based image metamorphosis*, vol. 26. New York, New York, USA: ACM Press, 1992.

[11] G. Wolberg and George, *Digital image warping*. IEEE Computer Society Press, 1990.

[12] M. Wierzbicki and T. M. Peters, *Determining Epicardial Surface Motion Using Elastic Registration: Towards Virtual Reality Guidance of Minimally Invasive Cardiac Interventions*. Springer Berlin Heidelberg, 2003.

[13] S. AbdelSayed, D. Ionescu, and D. Goodenough, "Matching and registration method for remote sensing images," in *International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications*, vol. 2, pp. 1029–1031, IEEE, 1995.

[14] S. Saalfeld, R. Fetter, A. Cardona, and P. Tomancak, "Elastic volume reconstruction from series of ultra-thin microscopy sections," *Nature Methods*, vol. 9, pp. 717–720, jun 2012.

[15] R. Nogales-Cadenas, S. Jonic, F. Tama, A. A. Arteni, D. Tabas-Madrid, M. Vázquez, A. Pascual-Montano, and C. O. S. Sorzano, "3DEM Loupe: Analysis of macromolecular dynamics using structures from electron microscopy.," *Nucleic acids research*, vol. 41, pp. W363–7, jul 2013.

[16] Q. Jin, C. O. S. Sorzano, J. M. de la Rosa-Trevín, J. R. Bilbao-Castro, R. Núñez-Ramírez, O. Llorca, F. Tama, and S. Jonić, "Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes.," *Structure (London, England : 1993)*, vol. 22, pp. 496–506, mar 2014.

[17] C. O. S. Sorzano, J. M. de la Rosa-Trevín, F. Tama, and S. Jonić, "Hybrid Electron Microscopy Normal Mode Analysis graphical interface and protocol.," *Journal of structural biology*, vol. 188, pp. 134–41, nov 2014.

[18] B. Falkner and G. F. Schroder, "Cross-validation in cryo-EM-based structural modeling," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 8930–8935, may 2013.

[19] M. J. Borgnia, S. Banerjee, A. Merk, D. Matthies, A. Bartesaghi, P. Rao, J. Pierson, L. A. Earl, V. Falconieri, S. Subramaniam, and J. L. S. Milne, "Using Cryo-EM to Map Small Ligands on Dynamic Metabolic Enzymes: Studies with Glutamate Dehydrogenase.," *Molecular pharmacology*, vol. 89, pp. 645–51, jun 2016.

[20] S. Pfeffer, J. Dudek, M. Gogala, S. Schorr, J. Linxweiler, S. Lang, T. Becker, R. Beckmann, R. Zimmermann, and F. Förster, "Structure of the mammalian oligosaccharyl-transferase complex in the native ER protein translocon," *Nature Communications*, vol. 5, pp. 869–878, jan 2014.

[21] S. Pfeffer, L. Burbaum, P. Unverdorben, M. Pech, Y. Chen, R. Zimmermann, R. Beckmann, and F. Förster, "Structure of the native Sec61 protein-conducting channel," *Nature Communications*, vol. 6, p. 8403, sep 2015.

[22] G. Bricogne, "Methods and programs for direct-space exploitation of geometric redundancies," *Acta Crystallographica Section A*, vol. 32, pp. 832–847, sep 1976.

Chapter 3

Improving the Visualisation of Cryo-EM Density Reconstructions

M. Spiegel[a,1], A. K. Duraisamy[a,1] , G. F. Schröder[a,b]
[a] Institute of Complex Systems (ICS-6), Structural Biochemistry, Forschungszentrum Jülich, 52428 Jülich, Germany
[b]Physics Department, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

DOI link to the publication
http://dx.doi.org/10.1016/j.jsb.2015.06.007

---

[1] Contributed equally
Corresponding author : E-mail address: gu.schroeder@fz-juelich.de (G.F. Schröder)

ABSTRACT

Cryo-electron microscopy yields 3D density maps of macromolecules from single-particle images, tomograms, or 2D crystals. An optimal visualisation of the density map is important for its proper interpretation. We have developed a method to improve the visualisation of density maps by using general statistical information about proteins for the sharpening process. In particular, the packing density of atoms is highly similar between different proteins, which allows for building a pseudo-atomic model to approximate the true mass distribution. From this model the radial structure factor and density value histogram are estimated and applied as constraints to the 3D reconstruction in reciprocal- and real-space, respectively. Interestingly, similar improvements are obtained when using the correct radial structure factor and density value histogram from a crystal structure. Thus, the estimated pseudo-atomic model yields a sufficiently accurate mass distribution to optimally sharpen a density map.

Abbreviations

VISDEM, Visualisation improvement by structure factor and density histogram correction of cryo-EM maps; cRDF, cumulative radial distribution function; FSC, Fourier shell correlation; FAS, fatty acid synthase; TRPV1, Transient receptor potential channel V1

Keywords

Cryo electron microscopy, B-factor sharpening, Density histogram matching

3.1 Introduction

Cryo-electron microscopy (cryo-EM) is a powerful technique to determine the structure of large macromolecules. In cryo-EM, a three-dimensional density map is reconstructed from a series of single-particle 2D images, tomograms or sub-tomogram averages, helical reconstructions or 2D crystals. An optimal visualisation of the reconstructed density map is important for its interpretation, and, if the resolution of the density map is high enough, for atomic model building. However,

density reconstructions often suffer from distortions or artifacts which arise from both electron optics as well as the density reconstruction process.

For example the contrast transfer function modulates the amplitudes of the structure factor, and in particular the envelope function of the transfer function dampens higher-resolution features in the density map. To improve the visualisation of the reconstruction, typically B-factor sharpening [1, 2] is used to amplify high-resolution features in the density map. Since B-factor sharpening only affects the radial structure factor amplitudes, i. e., all structure factors within the same resolution shell are scaled by the same factor, this method improves the visualisation but by definition not the resolution. A more advanced regularized deconvolution technique has recently been developed to address this problem [3].

Another density artifact that originates from the reconstruction process is a blurring of the particle periphery due to uncertainties in the angular assignment of particle image orientations. This leads to a density that is increasingly reduced and smeared out the farther away it is located from the center of the particle. Recently, a spherical deconvolution algorithm has been developed to reduce this blurring [4]. But as for B-factor sharpening, the resolution as determined by the Fourier Shell Correlation (FSC) is not affected by this deconvolution approach. A similar correction can be done also on 2D class-averages [5].

Density modification procedures such as for example solvent flattening [6] (or solvent flipping [7]) are often used to improve phases of electron density maps obtained from X-ray diffraction experiments [8]. For solvent flattening a mask needs to be determined, which defines the region that is covered by the molecule (or by the solvent). In addition the expected histogram of density values is often used as a constraint on the density map [9]. At high-resolution, where the density shows atomicity of the molecular structure, this histogram is rather shape-independent and depends mainly on the resolution (Fourier cut-off) of the density map. Therefore, in X-ray crystallography typically a generic resolution-dependent density histogram is applied.

Here we present a technique for improving the visualisation of cryo-EM density maps by applying two constraints: 1) in Fourier space on the radial structure factor, and 2) in real-space on the density value histogram. The estimates of the radial structure factor and the density histogram are obtained from an approxi-

mate model of the mass distribution within the reconstructed particle density. This approximate model is built using only the density reconstruction and statistical information on proteins calculated from the Protein Data Bank (PDB); no further knowledge about the protein structure is required. Our approach is therefore free of any model bias.

We tested our approach, called VISDEM (**V**isualisation **I**mprovement by **S**tructure factor and **D**ensity histogram correction of cryo-**EM** maps), on three different experimental density maps of fatty acid synthase, GroEL, and the transient receptor potential channel TRPV1 which have been determined at different resolutions of 18 Å, 8.9 Å, and 3.3 Å, respectively. For all of these cases either a docked X-ray structure or (for the highest resolution case) an atomic model built from the EM density was available for evaluating our results.

3.2 Results

3D density reconstruction procedures typically do not take into account the fact that one knows that the particle is, for example, a protein. However, this knowledge can provide additional information which yields restraints on the density map. The main idea of our method is to use information on the mass distribution of average protein structures to improve the visualization of cryo-EM density maps of proteins. In particular we are using knowledge of the average atomic density and average composition of elements in proteins. This provides constraints on both the radial structure factor (in reciprocal space) and the distribution of density values (in real space). For this, we need to create an approximate (pseudo-atomic) model of the mass distribution, which we refer to as the bead model. This bead model is supposed to be a very rough approximation to the true protein structure. The beads represent only the non-hydrogen atoms; adding the weakly scattering hydrogen atoms does not change the results in a noticeable way. The shape of the particle is given by the density map and the beads are placed randomly into the density map, i. e., the beads are placed in regions where the density is above a certain threshold. The question is only which density threshold defines the correct protein boundary or volume?

As the atomic density does not vary much within a protein and is very similar

in all proteins [10], the number of beads to be placed is given by the volume. However, the volume of the particle is defined by the density threshold. At high resolution (3–5 Å) the volume does not depend strongly on the density threshold and is therefore relatively well defined. With such an estimate of the volume the number of atoms can also be estimated using the known average atomic density. At lower resolution (worse than 5 Å) the volume depends more strongly on the density threshold. In that case the molecular volume cannot easily be determined and one needs to assume that the number of atoms is known.

As will be shown below, the exact placement of the beads in the volume is not critical for determining the radial structure factor and the density value histogram. But since the structure factor and the distribution of density values are not independent, one cannot set one without changing the other. We therefore modify the density by iteratively applying the constraints in reciprocal- and real-space, such that eventually a density is obtained which has a radial structure factor and distribution of density values close to the expected ones. The VISDEM method consists of the following steps:

1. Estimate volume of the particle with known or estimated number of atoms.
2. Generate a pseudo-atomic model (or "bead model").
3. Apply radial structure factor from bead model.
4. Apply density value histogram from bead model.
5. Repeat steps 3 and 4 one more time.

These steps are explained in more detail in the following.

3.2.1 Estimating the Volume

It is important to use a number of beads that is close to the actual number of atoms in the structure. Often the protein sequence is known which directly yields the number of atoms. The beads are then placed randomly at positions where the density is above a given threshold. This threshold determines the volume covered by the bead model and therefore the average bead density. To decide which density

threshold is best we compare the obtained bead distribution with the average atom density of protein structures in the Protein Data Bank (PDB). To do this, the cumulative radial distribution function (cRDF) of all atoms is calculated. The cRDF of atoms in protein structures is very similar at small distances for all proteins and then converges to the total number of atoms for larger distances. The average cRDF value for all non hydrogen atoms at a distance of 5 Å is 16.9 ± 0.85 calculated for 140 protein structures randomly chosen from the PDB. The best threshold value is the one for which the value of the cRDF function at 5 Å is closest to 16.9.

3.2.2 Estimating the Number of Atoms

Oftentimes an EM reconstruction does not show equally strong density for all atoms. Due to their high flexibility some loop regions or even entire protein domains might not be visible. We therefore propose a method to estimate the number of atoms that are visible in an EM reconstruction. At fixed density map threshold, we generate models with different numbers of beads. These bead models are then refined with the program DireX [11] into the EM density. The number of beads that yields the best cRDF is then chosen as the best estimate. This approach works only reliably for resolutions better than about 5 Å, since at lower resolution the boundaries of the protein are not well defined and the refined bead model will be blurred far outside the protein surface, which yields a cRDF curve that cannot be compared with the average cRDF obtained from PDB structures. The blurring of the refined bead model will then also lead to a wrong structure factor and would not improve the map sharpening.

The structure factor and the density value histogram strongly depend on the types of the scattering atom. We therefore randomly assign atom types to the beads in the pseudo-atomic model such that its composition is the same as the average composition observed for proteins in the PDB, i. e., 62.2% carbon, 17.2% nitrogen, 20.1% oxygen, and 0.5% sulfur. The obtained bead model is then an approximation of the real protein structure in terms of shape, mass distribution and elemental composition. If the sequence of the amino acids in the protein is known, one could directly use the correct composition of elements, but in practice this does not make a noticeable difference.

3.2.3 Matching Radial Structure Factor and Density Histogram

In the third step of the VISDEM protocol the radial structure factor of the EM reconstruction is matched to that of the bead model. First, a density map at high-resolution (typically two times the grid spacing) is created from the bead model with the program DireX, which uses theoretical electron scattering factors approximated by a sum of five Gaussian functions [12]. Then the radial structure factor is computed from the bead density map and applied to the original EM reconstruction. For these operations we use the program *e2proc3d.py* as part of the EMAN2 software [13].

In the fourth step, the real-space density value histogram of the structure-factor corrected EM reconstruction is matched to that of the bead model. Since the density value histogram is very different for different resolutions it is important to filter the bead density map to the final resolution at which the EM reconstruction will be examined. To match the density histogram of one map A to another map B, the density values in both maps are first sorted. Then the density values of map A are assigned in the same order to the corresponding grid points of map B. The new density map B' contains the same density values as map A (and therefore has the identical density histogram), but they are assigned to possibly different grid points. This procedure is implemented in the program *dx_matchhist* which is part of the DireX package (http://www.simtk.org/home/direx).

Steps 3 and 4 of the VISDEM protocol are repeated one more time, since the structure factor has to be corrected after applying the density value histogram. We did not observe consistent improvement when performing more than this one iteration.

It should be noted that no explicit mask is applied; the density map is affected only by the radial structure factor and density histogram constraints estimated from the bead model. While the bead model itself is created by defining a density threshold, which divides the density map into particle and solvent regions, the density values in the solvent region are not set to zero (as would be done in masking). One could in principle also apply a mask (e. g. using the density threshold) and set all density values outside the mask to zero, which would further shift all FSC curves to higher resolutions, but here we wanted to demonstrate only the effect of the VISDEM sharpening procedure without the additional effect of masking.

In the case that the atomic structure is already known, e. g., by X-ray crystallography, this known structure can be used to estimate both the radial structure factor and the density value histogram to be used in steps 2 and 3 of the VISDEM protocol. We refer to this as the ideal-VISDEM sharpened map, which we show in the application examples below for comparison to the regular bead model based VISDEM protocol.

3.3 Application Examples at Different Resolutions

The method was applied to three published test cases of different resolutions taken from the EMDataBank (EMDB): The density map of Mycobacterium tuberculosis fatty acid synthase multienzyme complex (EMDB-2538) with a resolution of 18 Å [14], the chaperonin GroEL/ES density map (EMDB-2325) with a resolution of 8.9 Å [15] and the transient receptor potential channel V1 (TRPV1) (EMDB-5778) with a resolution of 3.3 Å [16] were chosen as examples. Since there is no general rule on how EMDB deposited maps are processed, sometimes the maps are not optimally filtered or sharpened. Comparing the VISDEM sharpened maps with the original EMDB maps may thus exaggerate the improvement. We therefore show optimally B-factor sharpened EMDB density maps instead for comparison with the VISDEM sharpened maps. The B-factor sharpening was done by simply multiplying the Fourier components with $\exp(-B\,\mathbf{s}^2)$ using a negative value for B; no further weighting was applied.

In the first example, the density map of fatty acid synthase (FAS) was used to test VISDEM. For FAS, the cRDF was calculated from a bead model generated from the FAS density map at different threshold values. The number of beads was set to 125,670, which corresponds to the expected number of non-hydrogen atoms. The optimal density threshold was then found by comparing the cRDF of the bead model to an average cRDF obtained from a random selection of PDB structures (cf. Fig. 3.1(b)). The cRDF curves for different density threshold values were calculated and the optimal value of 3.0 yielded a cRDF curve (red) closest to 16.9 at 5 Å. This optimal density threshold corresponds to a volume of 2760 nm^3.

With the known molecular weight of 1.98 MDa and assuming an average partial specific volume of 0.714 ml/g [10] the expected volume is 2350 nm^3.

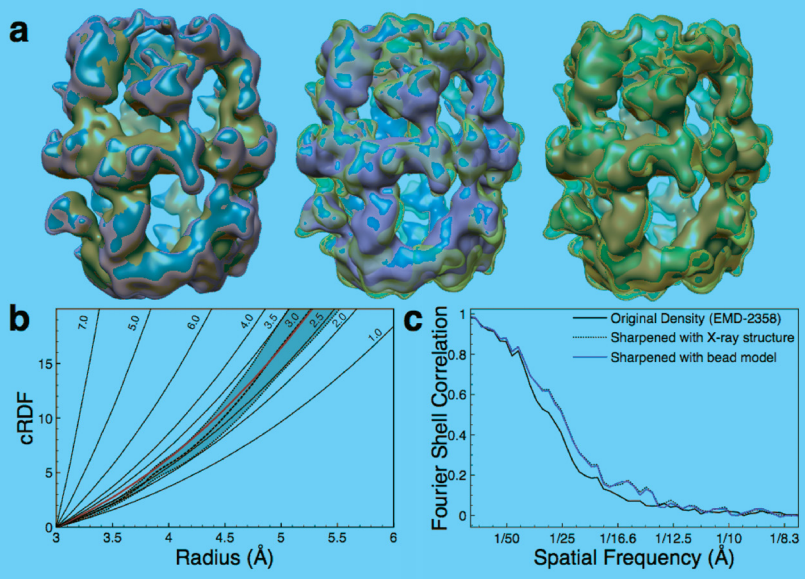

Figure 3.1 VISDEM was tested on the density map of the fatty acid synthase (FAS) protein from the EMDataBank (EMDB-2358) with a resolution of 20 Å. (a) The density map calculated from the X-ray structure (transparent green) is compared to the B-factor sharpened EMDB density map (orange) and the VISDEM sharpened map (blue). Density thresholds are chosen to enclose the same volume in all three maps. (b) The cumulative radial distribution function (cRDF) is plotted against the radius. The cRDF curves (black lines) of the bead models were calculated for different density threshold values. A density threshold of 3.0 (red cRDF curve) yields the best agreement of the bead model cRDF curve with the cRDF curve averaged over 140 randomly chosen PDB structures (black dashed line) at a radius of 5 Å. The $\pm 1\sigma$ cRDF curves are shown as dotted lines. (c) The Fourier shell correlation (FSC) was calculated between the deposited PDB structure and the original EMDB density map (black line), the VISDEM sharpened map using the bead model (blue line), and the ideal-VISDEM sharpened map (dotted line). The improvement in the FSC is the same for the bead model as for the X-ray structure sharpened VISDEM maps.

The FSC was calculated to evaluate the similarity between the atomic PDB model and the differently sharpened density maps. For this a 3 Ådensity map was computed from the atomic PDB model. Figure 3.1(c) shows the improvement of the FSC curve of the VISDEM sharpened map (blue curve) compared to the

FSC curve obtained for the original EMDB deposited density map. Interestingly, the FSC curve for the VISDEM sharpened map is almost identical to the ideal-VISDEM sharpened map, where the radial structure factor and density value histogram were computed from the PDB model instead of from the bead model. The VISDEM sharpened map (cf. Fig. 3.1(a), blue) shows more pronounced features on the periphery of the structure whereas the density of the B-factor sharpened map (orange) falls off more quickly towards the outside. The same trend can also be seen when comparing the sharpened maps with a density map computed from the PDB structure filtered to 18 Å(cf. Fig. 3.1(a), green).

As a second example GroEL is shown in Fig. 3.2. To calculate the bead model a threshold value of 0.9 was determined which yields a value closest to 16.9 at 5 Åin the cRDF. We used this value for further calculations even though a slightly better result could be achieved with a different threshold value of 1.1. As in the example for FAS, FSC curves were calculated for GroEL between a density calculated from the X-ray structure (PDB ID 3zpz) and the original EMDB map (black), the VISDEM-sharpened map (blue) and the ideal-VISDEM sharpened map (black dotted curve) which are shown in Fig. 3.2(c). Again, the FSC curves indicate better agreement of the VISDEM sharpened maps with the X-ray structure. The differently sharpened density maps of GroEL/ES are shown in Fig. 3.2(a). The outer regions in the VISDEM sharpened density map (blue) have overall stronger density and are on a more similar scale as the central regions of the particle when compared to the B-factor sharpened map (orange). Also visually, the VISDEM sharpened map shows more similarity to the density map computed from the X-ray structure (green).

In the third example, the results for TRPV1 are shown in Fig. 3.3. The outer domains of the tetramer show reduced density which suggests that not all atoms are equally well visible in the density map. Since the resolution of 3.3 Åis very high, the number of (visible) atoms can be estimated directly from the density map. For this, bead models with different numbers of beads were created which were then refined into the density without any restraints between the beads using the program DireX. At this resolution the different refined bead models have a similar clearly defined boundary and thus enclose a similar volume, the only difference is the number of the beads. The best cRDF curve was obtained with 12,500 beads, which

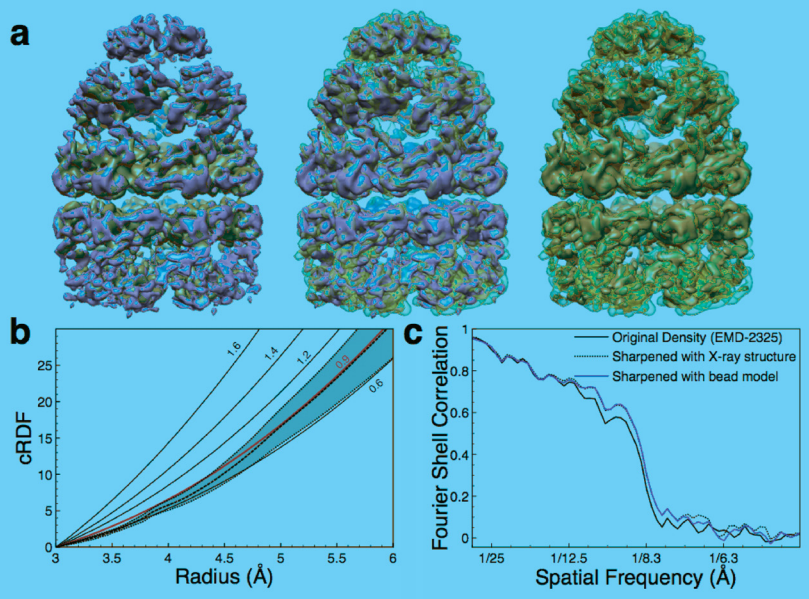

Figure 3.2 Results of the VISDEM sharpening protocol are shown for a GroEL/ES density map (EMDB-2325) with a resolution of 8.9 Å. (a) Density maps of GroEL created for X-ray structure (transparent green), sharpened by VISDEM (blue) and B-factor only (orange). (b) Cumulative radial distribution functions of the bead models (black curves) are shown for different threshold values. The optimal threshold of 0.9 yields a curve (red) closest to the PDB average (black dashed line). (c) FSC curve calculated between X-ray model map, filtered to 7 Åand original map (black), VISDEM bead model sharpened map (blue) and VISDEM X-ray sharpened map (black dotted).

yields a volume of 220 nm^3. Whereas the volume estimated from the molecular weight of 245 kDa is 290 nm^3, which suggests that 25% of the atoms are not clearly visible in the density. These atoms likely belong to flexible domains that are not well resolved in the density.

Interestingly, the FSC curve for the bead model sharpened map indicates better agreement with the PDB model than the PDB model (ideal-VISDEM) sharpened map. The reason is likely the fact that the PDB model includes the ankyrin repeat domains, which are not completely visible in the density, i. e., the model sticks out of the density. The density histogram and radial structure factor computed from this PDB structure therefore do not match well the (smaller) particle shown by the density. Our bead model reflects better the visible density and yields a more

appropriate estimate for the density histogram and structure factor.

The black FSC curve in Fig. 3.3 (c) which we calculated between the original density map and the fitted atomic model differs from the comparison in the original publication [16] as we did not apply a mask for the calculation of the FSC.

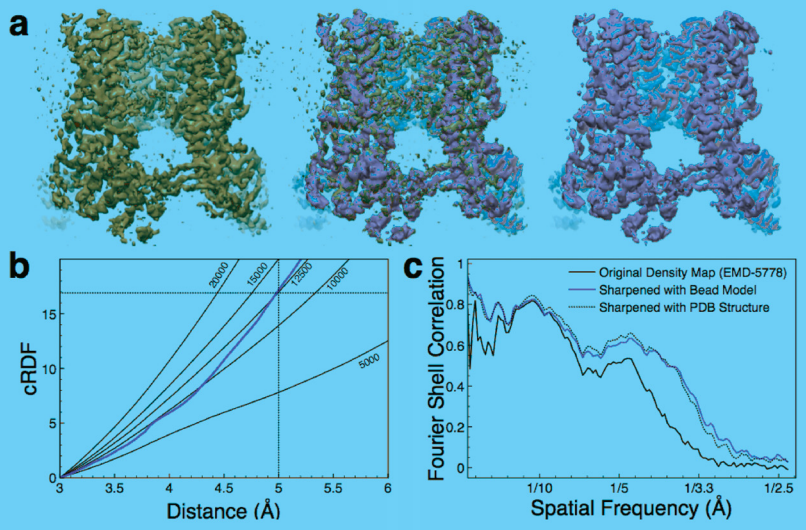

Figure 3.3 Showing results of the VISDEM sharpening protocol for the transient receptor potential channel V1 (TRPV1) (EMDB-5778) with a resolution of 3.3 Å. (a) The density contour level was chosen to yield the same contour for both maps in the central pore region. (b) Cumulative radial distribution function (cRDF) of bead models created with different numbers of beads and then refined into the density map (black curves). The bead model with 12,500 atoms has the same cRDF value as the PDB average (blue) at a radius of 5 Å. (c) FSC curve for TRPV1 .

B-factor sharpening emphasizes high-resolution features such as side-chains, but at the same time it leads to more noise all over, which results in a smaller correlation with the PDB model. Our sharpening shows similar details as the best B-factor sharpened map, but with much less additional noise. It also puts peripheral regions on a more similar scale as interior regions (cf. Fig. 3.4) this leads to better balance of the density over the entire structure.

However, one should be aware of the fact that lower density regions are suppressed, as can be seen by the noisy density around the membrane bound region which likely comes from amphipol molecules which were used in the experiment

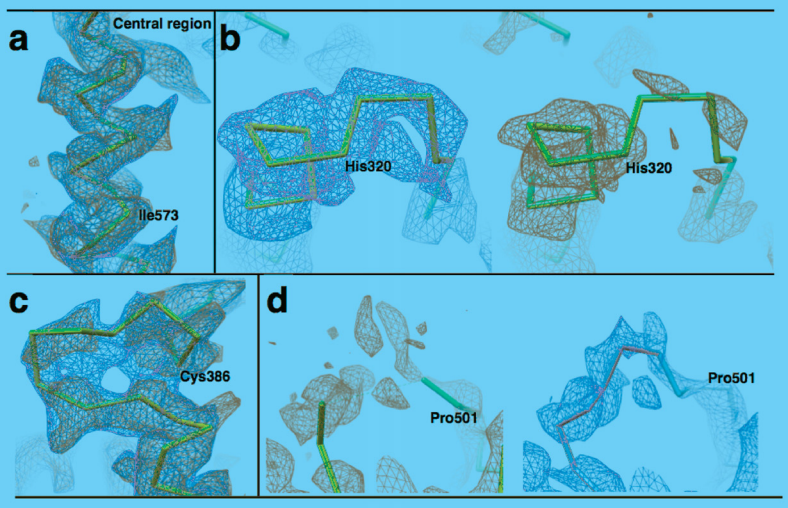

Figure 3.4 Showing details of the density improvement upon VISDEM sharpening for the TRPV1 channel. (a) The threshold was chosen to yield a similar surface for the B-factor sharpened map (orange) and the VISDEM sharpened map (blue) at the central region of the protein. The deposited PDB model is superimposed (green). For this threshold, density regions are shown in (b-d) at regions farther away from the center of the particle, showing stronger and better connected density for the VISDEM sharpened map. (d) The backbone in the loop around Pro501 could not be traced with the original density map, while the VISDEM sharpened map suggests a Ca-trace (blue trace) due to the better connected density, even if the assignment of the amino acid sequence in this loop is still ambiguous.

to stabilize the membrane protein and therefore has physical meaning. The corresponding density is too weak to be filled with beads and therefore does not contribute to the structure factor estimate. The visualisation of such low-density features in the map, thus, does not benefit from the VISDEM sharpening.

To further evaluate our results the cross-correlation coefficient was calculated between the differently sharpened maps and a density calculated from the corresponding PDB structures, which served as the best possible answer. We assume that a better EM density is more similar to this best available (PDB) atomic model

and we use this similarity to assess the quality of the sharpening. The different models that were compared to the PDB structure were the original density map as deposited to the EMDB, an optimally B-factor sharpened density map, the VISDEM and ideal-VISDEM sharpened maps.

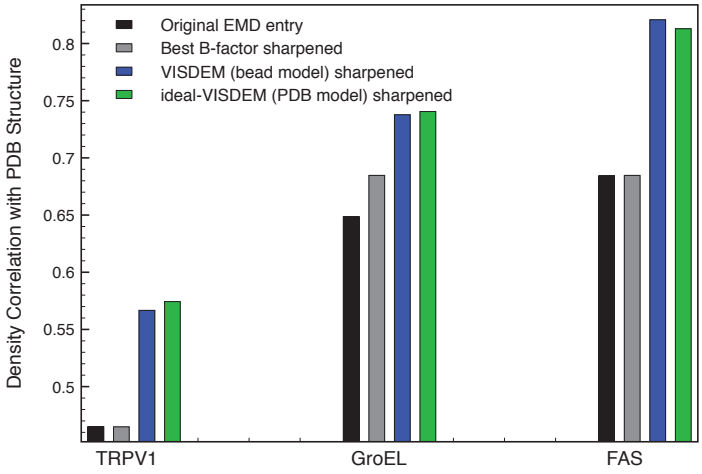

Figure 3.5 The graph shows cross-correlation coefficients for all three test cases which are always calculated to a filtered map of the corresponding PDB structures. The correlation with the original EMDB density map is shown in black, the B-factor sharpened map (-200 Å^2) in gray, the VISDEM sharpened map (bead model) in blue and the ideal-VISDEM sharpened map (PDB structure) in green.

All density maps were filtered to a resolution of 18 Å, 8.9 Å, and 3.3 Å for the FAS, GroEL, and TRPV cases, respectively, before calculating the cross-correlation coefficient. The obtained cross-correlation coefficients are compared in Fig. 3.5. From the graph it is clear that the VISDEM sharpened maps have the highest correlation coefficients when compared to both the original map from EMDB and the best B-factor sharpened map. The VISDEM sharpened maps for the bead model and the PDB model are highly similar, which shows that approximate bead model carries sufficient information to optimally sharpen the density maps and that knowl-

edge of the "real" structure would not even improve the density further.

3.4 Discussion

We have developed the VISDEM method to sharpen cryo-EM density maps by using constraints on the radial structure factor as well as on the density value histogram. Application of this method to three examples shows clear improvement in the density map (Figs. 1-3) when compared to B-factor sharpened maps. Further, the VISDEM sharpened maps are significantly more similar to the known atomic models than the B-factor sharpened maps. This similarity was quantified by computing FSC curves and density cross-correlation coefficients between the sharpened maps and maps computed from the atomic models.

The regular VISDEM method uses a simple bead model to approximate the mass distribution of the particle, whereas the ideal-VISDEM method uses a known atomic model. Interestingly, both approaches yield very similar results, which means that the simple bead model provides sufficient information on the radial structure factor and density histogram, without the risk of any model bias.

The resolution in cryo-EM is defined by the FSC between two 3D reconstructions that were obtained from two independent half-sets of the image data. While B-factor sharpening can significantly improve the visualization of density maps it does not affect FSC curves. The VISDEM method instead does affect FSC curves and the resolution estimate in fact improves when applying the VISDEM procedure to both half maps. However, even when constructing the bead models independently for both half maps, it is not clear how the procedure of placing the beads imposes similar information on both half maps and therefore introduces correlations, which could artificially improve the FSC. We therefore suggest not to use VISDEM sharpened maps for estimating the resolution.

We have shown that the volume (optimal threshold) of a low-resolution density map can be found, if we assume that the number of visible atoms is known. And we have also shown that the number of visible atoms can be estimated from the density, if the resolution of the density map is better than 5 Å.

Density maps of macromolecules containing mixtures of DNA, RNA, and protein cannot as easily be improved by VISDEM, since nucleic acids produce stronger

density than amino acids. The position of the nucleic acids influences the structure factor and density histogram and therefore the placement of nucleic acids in the map needs to be at least approximately known. The same holds true in principle for every deviation from the average atom distribution in proteins, such as for example metalloproteins.

The VISDEM sharpening procedure improves density maps of proteins over a large range of resolutions. Such a sharpened map is a better target map for atomic model building and refinement, as some of the artifacts from imaging and density reconstruction (both in real and reciprocal space) are removed and the sharpened map therefore agrees better to density expected from an atomic model without adding a model bias.

Acknowledgement

This work was inspired by the IEEE - Signal processing cup challenge 2014. Funding was provided by the Deutsche Forschungsgemeinschaft grant SCHR 969/5-1 (to G.F.S.).

Bibliography

[1] J. J. Fernández, D. Luque, J. R. Castón, and J. L. Carrascosa, "Sharpening high resolution information in single particle electron cryomicroscopy.," *Journal of structural biology*, vol. 164, pp. 170–5, oct 2008.

[2] B. DeLaBarre and A. T. Brunger, "Considerations for the refinement of low-resolution crystal structures.," *Acta crystallographica. Section D, Biological crystallography*, vol. 62, pp. 923–32, aug 2006.

[3] M. Hirsch, B. Schölkopf, and M. Habeck, "A blind deconvolution approach for improving the resolution of cryo-EM density maps.," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 18, pp. 335–46, mar 2011.

[4] G. P. Kishchenko and A. Leith, "Spherical deconvolution improves quality of single particle reconstruction.," *Journal of structural biology*, vol. 187, pp. 84–92, jul 2014.

[5] W. Park, D. R. Madden, D. N. Rockmore, and G. S. Chirikjian, "Deblurring of Class-Averaged Images in Single-Particle Electron Microscopy.," *Inverse problems*, vol. 26, pp. 3500521–35005229, mar 2010.

[6] B. C. Wang, "Resolution of phase ambiguity in macromolecular crystallography.," *Methods in enzymology*, vol. 115, pp. 90–112, 1985.

[7] J. P. Abrahams and A. G. Leslie, "Methods used in the structure determination of bovine mitochondrial F1 ATPase.," *Acta crystallographica. Section D, Biological crystallography*, vol. 52, pp. 30–42, jan 1996.

[8] K. Cowtan, "Recent developments in classical density modification," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 470–478, apr 2010.

[9] K. Yong, J. Zhang, and P. Main, "Histogram Matching as a New Density Modification Technique for Phase Refinement and Extension of Protein Molecules," *Acta Cryst*, vol. 46, pp. 41–46, 1990.

[10] J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, "The packing density in proteins: standard radii and volumes," *Journal of Molecular Biology*, vol. 290, pp. 253–266, jul 1999.

[11] G. F. Schröder, A. T. Brunger, and M. Levitt, "Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution," *Structure*, vol. 15, pp. 1630–1641, dec 2007.

[12] L. M. Peng, G. Ren, S. L. Dudarev, and M. J. Whelan, "Robust Parameterization of Elastic and Absorptive Electron Atomic Scattering Factors," *Acta Crystallographica Section A Foundations of Crystallography*, vol. 52, pp. 257–276, mar 1996.

[13] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "EMAN2: an extensible image processing suite for electron microscopy.," *Journal of structural biology*, vol. 157, pp. 38–46, jan 2007.

[14] L. Ciccarelli, S. R. Connell, M. Enderle, D. J. Mills, J. Vonck, and M. Grininger, "Structure and conformational variability of the mycobacterium tuberculosis fatty acid synthase multienzyme complex.," *Structure (London, England : 1993)*, vol. 21, pp. 1251–7, jul 2013.

[15] D.-H. Chen, D. Madan, J. Weaver, Z. Lin, G. F. Schröder, W. Chiu, and H. S. Rye, "Visualizing GroEL/ES in the Act of Encapsulating a Folding Protein," *Cell*, vol. 153, pp. 1354–1365, jun 2013.

[16] M. Liao, E. Cao, D. Julius, and Y. Cheng, "Structure of the TRPV1 ion channel determined by electron cryo-microscopy," *Nature*, vol. 504, pp. 107–112, dec 2013.

Chapter 4

Estimating positional precision in real-space refinement

A. K. Duraisamy[a,1], G. F. Schröder[a,b]
[a] Institute of Complex Systems (ICS-6), Structural Biochemistry, Forschungszentrum Jülich, 52428 Jülich, Germany
[b]Physics Department, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

Corresponding author : E-mail address: gu.schroeder@fz-juelich.de (G.F. Schröder)

4.1 Introduction

The resolution achieved by cryo-EM has made a huge leap due to the resolution revolution in the field. But, on average most of the cryo-EM data is still in the intermediate- or low-resolution range (3.5–5 Å). At such resolutions building atomic models or even refining known models against the density map is very challenging. The most accurate and precise structures can be obtained if the density map starts to show atomic features (better than 2.5 Å). Hence, the atomic models at intermediate- to low-resolution are very much prone to errors. A robust measure of atomic coordinate error is needed to address the issues of trustability of the structure and the level of details that can be reliably interpretated, such as ambiguity in the side-chain rotamers and the presence of hydrogen bonds. Accuracy is a systematic error that is difficult to determine in practice. If the measured values of an observable fall close to the correct value then the measurement is accurate. Whereas, the standard deviation measured values is called the precision of a measurement, irrespective of whether the average value is accurate. The precision describes the random error which is associated with the parameters used and is expressed in terms of standard uncertainties.

Refinement of protein structure is an important step in structure determination. It is an iterative process to get the best agreement between the experimentally observed value and that numerically calculated from the reference structure. Due to the poor ratio between the observed and refinable parameters, prior informations such as bond lengths, angles are added. These informations are treated as extra experimental observables and they are known as restraints. Restraints are different from constraints: Constraints are the rules to be kept while refinement and thus reduces the number of free refinable parameters. Refinement can be done either restrained or unrestrained. The real space structure refinement can be done by softwares like the molecular dynamics flexible fitting (MDFF) [1], DireX [2], iMODFIT [3], Flex-EM [4] and reciprocal space refinement by REFMAC [5]. The most common refinement in EM is the real-space refinement.

In XRC, both the measure of accuracy and the precision of atomic positions during refinement are well established. A statistical quantity R_{free} is used to assess the accuracy of XRC structures [6]. While precision measurement techniques in

XRC includes Luzzati plot [7], estimating error in least-square method [8], residue R-factor [9]. The positional precision is calculated once the refinement has converged to the best structure by taking the least square covariance matrix from the inversion of the full normal equation matrix gives an estimate on the variances and also the covariances for all parameters [10]. The matrix inversion is doable only for very small structures and gets computationally expensive for large complexes. The atomic precision strongly depends on the Debye B-factor [11]. Unlike XRC, the accuracy and precision estimation in cryo-EM is not so common for refinement.

Here, we have developed a method to determine the precision of atomic positions in the Cartesian coordinates while doing a real-space structure refinement against cryo-EM density maps. The variance in the conformational sampling during the real-space refinement is used to calculate the positional precision. The root mean square deviation (RMSD) is calculated between the output refined structures and the target structure. The mean root mean square fluctuation (RMSF) calculated for the ensemble, which has the lowest RMSD will give us the precision estimate of the atomic positions. Usually, the refinement is done because the target structure is not known. Hence, the RMSD cannot be calculated between the refined and the target structure. Thus, one has to exploit the negative correlation that exists between the cross validation (C_{free}) parameter [12] and the RMSD to select the right ensemble. In other words, the best refined structures having the lowest RMSD will have the highest cross correlation C_{free} value. C_{free} is analogous to R_{free} value in XRC. A random set of data is selected to calculate R_{free}. Whereas in cryo-EM, a continuous band of data known as free band is selected to calculate C_{free}. The error on the cross-validation C_{free} value gives an estimate of the range the ensemble has to picked from. Thus, by estimating the C_{free} value and the C_{free} error value, the best ensemble could be chosen for calculating the positional precision.

4.2 Method

The real-space refinement is done here with molecular dynamics flexible fitting software (MDFF). MDFF refinement yields a broad conformational sampling as output. The variance in the conformational sampling is used to calculate the positional error in real-space refinement. Figure 4.1 shows the main steps used in our

method. The first step of the method is the standard procedure used in MDFF to refine an atomic structure against density map. In MDFF the density map is converted into a an additional potential, which can be scaled by a factor (*gscale*). The optimal *gscale* value will give the lowest RMSD value that is calculated between the target and the output structures. Because the target is not known in general, the (C_{free}) value is used to estimate the optimal *gscale* value. C_{free} is a cross-validated cross-correlation value and hence the optimal *gscale* value will give the highest C_{free} values.

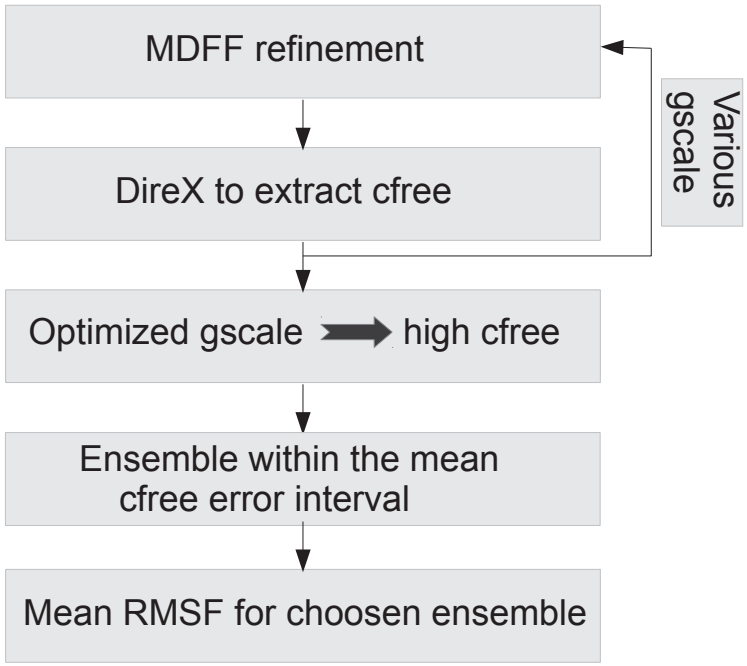

Figure 4.1 The above flow chart explains the workflow to estimate the precision of position of atoms in Cartesian coordinates for a real-space structural refinement.

Once the MDFF refinement is done, the program DireX is used to extract the C_{free} and C_{free} error values. DireX can handle only a PDB-formatted input file and the density map for the fit. Hence, the individual PDB files corresponding to a single time step of the MDFF simulation is extracted from the output trajectory. The cross-validation in DireX has two parameters C_{work} and C_{free} analogous to R_{work}

and R_{free} in XRC. Unlike R_{free}, the absolute C_{free} value does not mean much, since the value of C_{free} depends on the choice of the free band. Thus, only the relative change of C_{free} value is what needs to be considered. Being a cross-correlation value, the C_{free} will increase when the initial structure is moving towards the target, then it will converge and will start to decrease when the structure is overfitted. Unlike in XRC, a few percentages of the data cannot be chosen randomly in cryo-EM because of strong correlations between structure factors. Instead, the structure factors for the free set have to be taken from a continuous band in the high-resolution range. The C_{free} band value is typically selected to be higher than the resolution of the map, even though this high-frequency range is dominated by noise. This range is usually good enough for validating the refinement process and is not for the refinement. Normally, the range from the map resolution up to the lowest frequency information available is used as work band and for the refinement.

After finding the optimal *gscale* value with the highest C_{free} values, the best ensemble of structures from the optimal *gscale* data set is selected for calculating the mean RMSF value. The structures that falls within the interval range having a value equal to the mean C_{free} error is chosen. The interval, defined as the C_{free} error range, has the maximum value corresponding to the maximum C_{free} value obtained. Minimum of the C_{free} error interval is given by difference between maximum C_{free} value and the mean C_{free} error value. The mean RMSF values gives the measure of atomic positional uncertainity in real-space refinement.

4.3 Results

4.3.1 MDFF refinement

As a test example, we have used the adenylate kinase structure (PDB ID: 4AKE), a signal transducing protein containing 214 residues and 3341 atoms. Simulated density maps for this structure were generated at 5 Å and 7 Å. MDFF refinement is a well established protocol and the main interest while testing this method is not in refining an atomic structure to a density map but to create an ensemble of refined structures in the conformational space that is in agreement with the density map. Hence the actual refinement of fitting an atomic structure to density map is

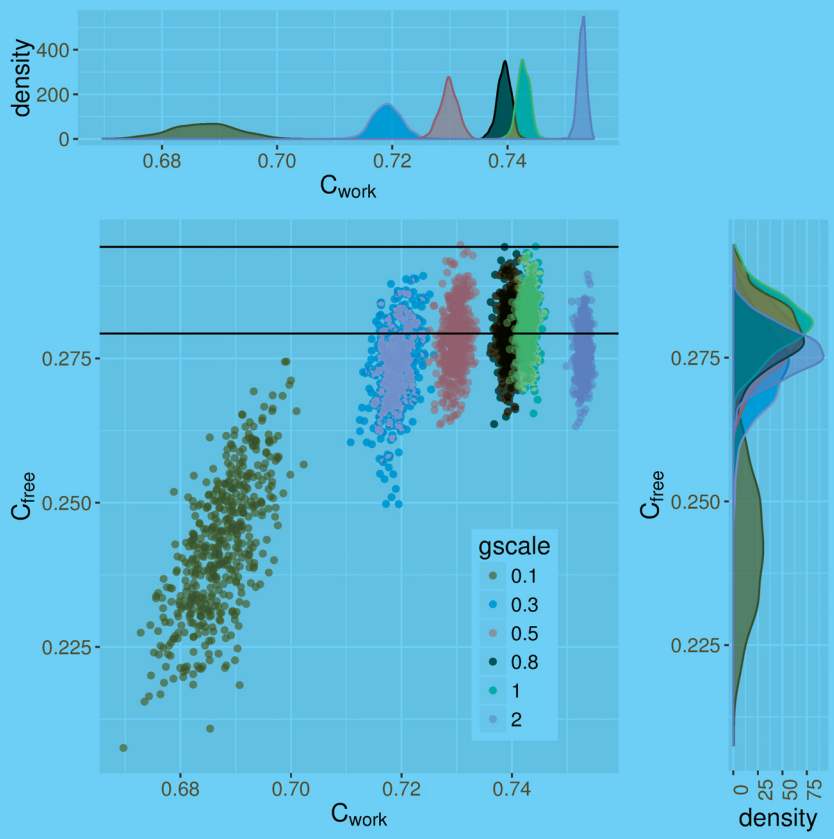

Figure 4.2 Multiple plot of C_{free} vs C_{work} for 5 Å data from MDFF refinement. The top density plot shows the C_{work} density for various *gscale* values. The bottom left is a scattered plot of C_{free} vs C_{work} showing *gscale* value of 1.0 gives the highest $C - free$ values. The two solid black lines in the scattered plot at 0.296 and 0.2819 marks the C_{free} error interval. The bottom right is the density plot for C_{free} values.

not done here. Therefore it is enough to choose the target structure (4AKE) as the start structure for refinement. Hence the usual trend of trajectory and so the change in the C_{free} value during cross validation will not change drastically. Creating an ensemble with MDFF refinement from the same atomic structure against the same density map is good enough to calculate the positional error in real-space refinement.

MDFF refinements for various *gscale* values is done for 500 ps time scale.

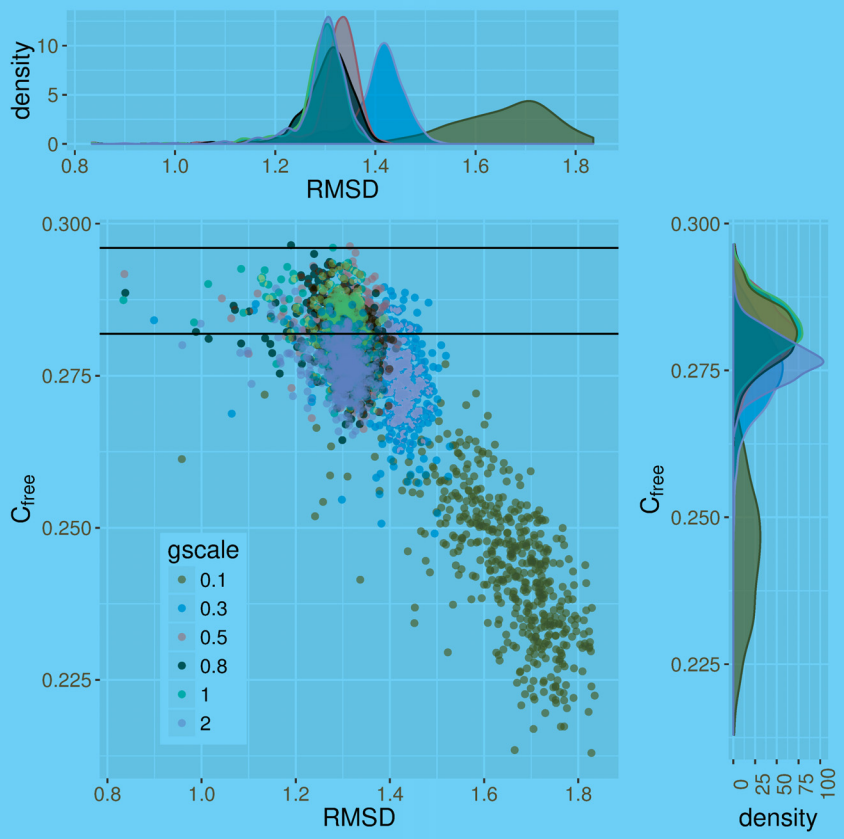

Figure 4.3 Multiple plot of C_{free} vs RMSD for 5 Å data from MDFF refinement. The scatter plot in the bottom left shows the optimized *gscale* value 1.0 has the highest C_{free} values and the lowest RMSD values. The two density plots on the top and the bottom right also clearly shows that the highest C_{free} corresponds to the data having the lowest RMSD. The ensemble corresponding to *gscale* 1.0 within the two solid black lines at 0.296 and 0.2819 defining the C_{free} error interval are those chosen for calculating the mean RMSF value.

The trajectory file from the MDFF is converted to individual PDB files. The output coordinates from the MDFF refinement is then used as input to the DireX to extract the C_{free} values. Just to extract the C_{free} values, a low map strength of 0.01 was used so that there is not so much force on the structure from the density map. The deformable elastic network is also switched off during this step. The optimized *gscale* value will have the highest C_{free} from the plot of C_{work} vs C_{free} as *gscale*

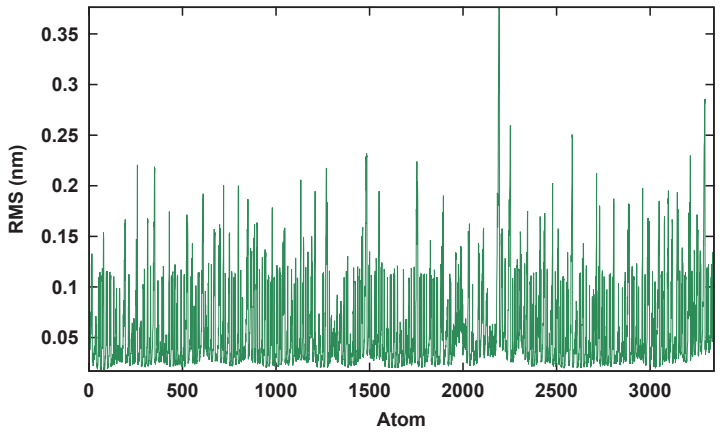

Figure 4.4 The RMS value for each atom with *gscale* 1.0 for 5 Å MDFF data

1.0 in figure 4.2. In our case, the scaling of the density potential - *gscale* value, ranges from 0.1 to 2.0. Once the optimal *gscale* value having the highest C_{free} value is determined, the mean C_{free} error for those output structures is calculated. For the data in figure 4.2 the C_{free} band for 5 Å density map was from 3.5 Å to 4.5 Å. Both from the scattered plot and the C_{free} density plot in figure 4.2 clearly shows the data set in green corresponding to 1.0 is the best *gscale* value. For 5 Å density map, the calculated mean C_{free} error was 0.0142 for *gscale* 1.0. This is also proved from the figure 4.3, that the optimized *gscale* value 1.0 having the highest C_{free} also has the lowest RMSD values. The mean C_{free} error value defines the C_{free} error interval value. The upper limit is the highest C_{free} value which is 0.296 and the lower limit is (0.296 - 0.0142) 0.2819. In figure 4.3, the two solid black horizontal lines in the scattered plot shows the C_{free} error interval to pick the ensembles for the optimized parameter. The *gscale* 1.0 ensemble that falls within the C_{free} error interval is then used for calculating mean RMSF. The RMSF calculated for each atom for *gscale* 1.0 ensemble is in the figure 4.4. The calculated mean RMSF for *gscale* of 1.0 is 0.6138 Å.

The method was also tested on the 7 Å density map. The figure 4.7 corresponds to 7 Å data and is similar to figure 4.3. The C_{free} band for 7 Å density map was from 5.5 Å to 6.5 Å for the figure 4.7. The *gscale* value of 0.8 in black dots is the optimized value. The calculated mean C_{free} error for 7 Å data was 0.0295.

Table 4.1 Positional precision for different resolutions

Resolution	C_{free} range	C_{free} max	C_{free} min	C_{free} error	Mean RMSF
5 Å	3.5-4.5 Å	0.2960	0.2819	0.01415	0.6138 Å
7 Å	5.5-6.5 Å	0.4569	0.4274	0.02954	0.7157 Å

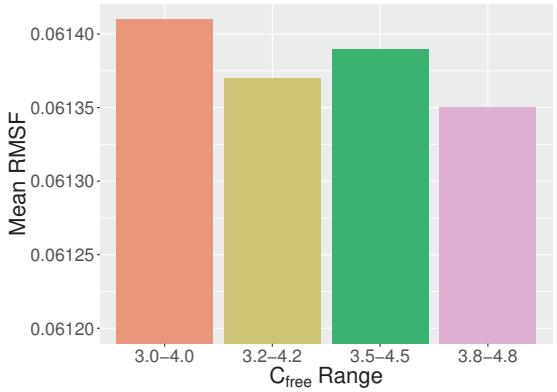

Figure 4.5 The mean RMSF value for different C_{free} band is plotted in the bar plot for 5 Å MDFF refinement data. The mean RMSF values does not change much for any free band value chosen and thus making the method not sensitive to the free band range.

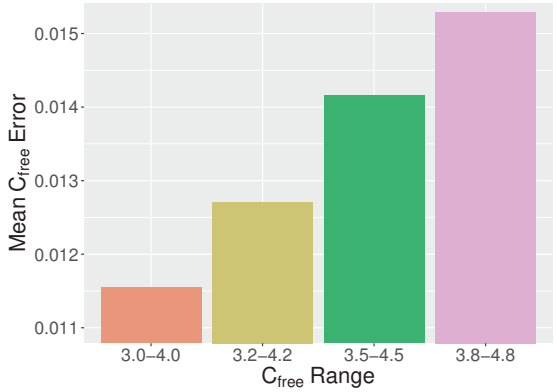

Figure 4.6 The bar plot shows the values of mean C_{free} error for different C_{free} band for 5 Å MDFF refinement data.

The two solid lines representing the C_{free} error interval was between 0.4569 and

0.4274. The mean RMSF for the selected set of best ensembles for *gscale* 0.8 was 0.716 Å. The table 4.1 gives a summary of the various values used and determined while calculating the positional precision for both 5 Å and 7 Å data.

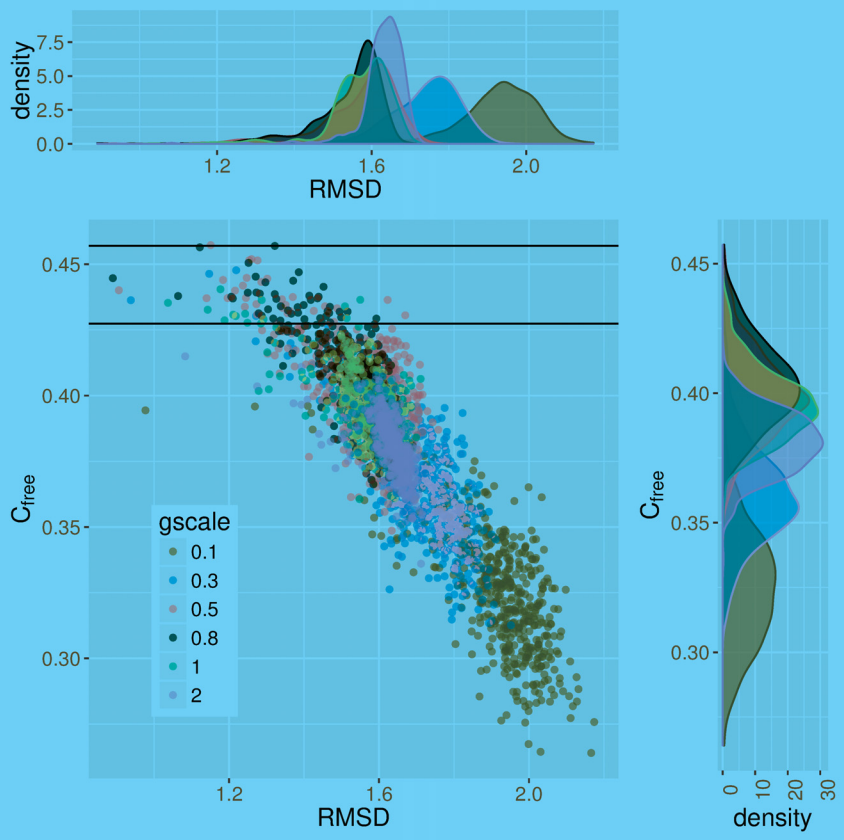

Figure 4.7 Multiple plots of C_{free} vs RMSD from MDFF refinement for 7 Å data. The scattered plot in the bottom left shows the optimized *gscale* value 0.8 in black dots has the highest C_{free} values and the lowest RMSD values. The two density plots on the top and the bottom right also clearly show that the highest C_{free} has to the lowest RMSD. The ensemble corresponding to the *gscale* value of 0.8 within the C_{free} error interval, the two solid black lines at 0.456 and 0.4274, are used for calculating the mean RMSF value.

There is no fixed rule to select the free band range in DireX. Hence, we also would like to determine the dependency of positional precision on the free interval. To do so, we estimate the positional precision by using different free intervals while

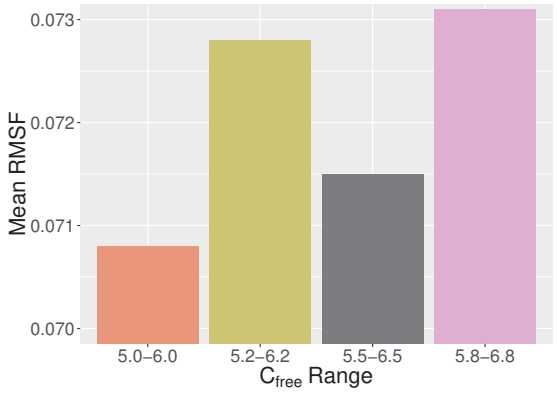

Figure 4.8 The mean RMSF for various $C_{\mathbf{free}}$ band ranges for 7 Å refinement data.

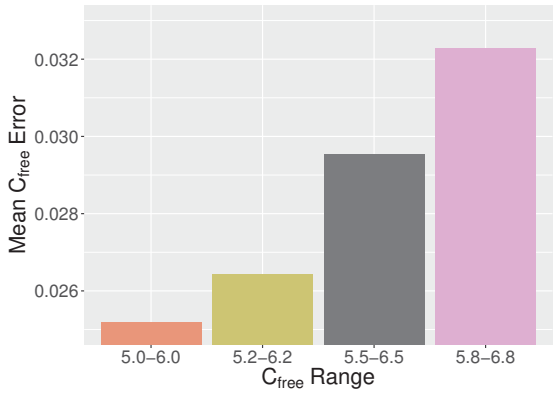

Figure 4.9 The mean $C_{\mathbf{free}}$ error for 7 Å data plotted for various $C_{\mathbf{free}}$ band ranges.

extracting the C_{free} value from the MDFF outputs using DireX. Figure 4.5 and figure 4.8 shows the mean RMSF calculated for different free interval ranges for 5 Å and 7 Å respectively. In both the plots the RMSF does not change much for different interval ranges. So, one can pick an interval range with the high-frequency limit less than the resolution of the map and the low frequency limit 1 Å less than the higher limit. Having a broader interval range will increase the value of C_{free} value at the cost of using less data for the actual refinement. Figure 4.6 shows the mean C_{free} error corresponding to data in figure 4.5 for different free intervals at 5

Å. Figure 4.9 shows the mean C_{free} error for different free intervals corresponding to data in figure 4.8 at 7 Å.

The table 4.2 is just an another representation of the four figures (figure 4.5 - figure 4.9). The table 4.2 gives an overview of the different free band values used and the corresponding mean C_{free} error and mean RMSF calculated for 5 Å and 7 Å data.

Table 4.2 Positional precision for different free bands

Resolution (Å)	free range (Å)	Mean C_{free} error	Mean RMSF (Å)
5	3.0-4.0	0.01155	0.6141
5	3.2-4.2	0.01271	0.6137
5	3.5-4.5	0.01416	0.6139
5	3.8-4.8	0.01529	0.6135
7	5.0-6.0	0.02519	0.7080
7	5.2-6.2	0.02643	0.7281
7	5.5-6.5	0.02954	0.7157
7	5.8-6.8	0.03227	0.7310

4.3.2 Ensembles from DireX refinement

We also performed DireX refinements from the start structure against the density map. Various γ-parameters were used in DireX. The gamma-parameter defines the deformability of the restraints and, thus, to some extent is a weighting factor between the reference model and the experimental data. The gamma-parameter can hold a value between 0 and 1. A small γ value means the start structure will be kept very close to the reference model and a high value means larger deformation from the reference model. An optimal value of γ is found with the highest C_{free} value as in the case of MDFF runs.

The same density map and start structures used for the MDFF refinements were also used for the DireX refinement. The start structure was the target structure (PDB ID: 4AKE) and the refinement was done for 3000 steps. Various γ values were used to check the best C_{free} values. The map strength used was 0.01 and all other parameters are kept to default values. We can see from the figure 4.10 that the variance in the conformational sampling is not as large as in the MDFF output.

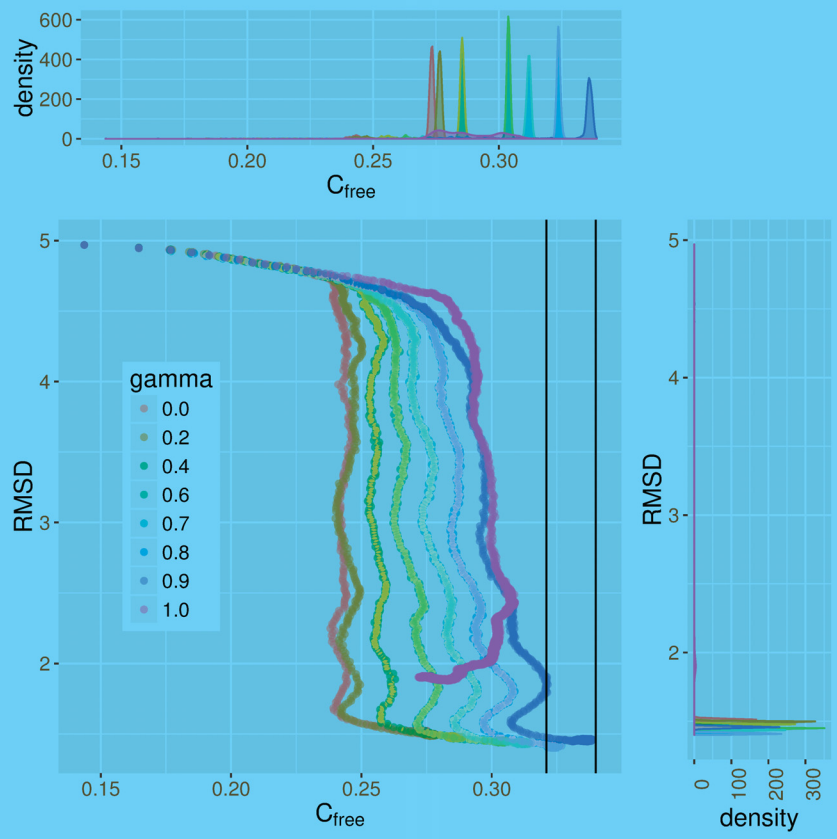

Figure 4.10 The whole plot shows the trend in C_{free} vs RMSD for different γ-values used during the DireX refinement. Top plot is the C_{free} density plot for all gammas showing that the gamma=0.9 gives the highest C_{free} value. Bottom scattered plot shows C_{free} vs RMSD and the best ensemble is selected from γ= 0.9 having highest C_{free} values. The mean C_{free} error value gives the interval range from which the ensemble has to be chosen and it is represented by the two vertical black lines at 0.3213 and 0.3388 for calculating the RMSF value. The bottom right plot shows the RMSD density plot for all γ values, indicating γ = 0.8 having the least RMSD value.

This is because the MDFF method allows for a large conformational sampling space when compared to DireX. The highest γ value of 0.9 yields the highest C_{free} value. Hence the ensemble for this γ value is chosen for calulating the mean RMSF. The mean C_{free} error value for γ 0.9 is 0.0175. The maximum C_{free} value is 0.339, thus making the C_{free} error interval between 0.339 to 0.322. The calculated

mean RMSF for γ = 0.9 within the C_{free} error interval is 0.1186 Å. The mean RMSF value is lower than the MDFF-mean RMSF value since the conformational sampling in DireX is not as spread out as in the case of MDFF. One could see from the density plot for RMSD that the highest C_{free} values does not correspond to the lowest RMSD. The lowest RMSD is for γ = 0.8. Hence, the mean RMSF is calculated for this ensemble which falls within the C_{free} error interval 0.03267 and 0.0309. The calculated mean C_{free} error value is 0.0176 and the mean RMSF is 0.1186 Å for γ = 0.8. Even though the highest C_{free} data does not correspond to the lowest RMSD data, the mean RMSD value remains the same.

4.4 Conclusion

There are different ways to estimate the atomic coordinate precision of the refined crystal structure in XRC, but none are available until now for cryo-EM data. With most of the cryo-EM data still at low- to intermediate resolution, even the refinement of known atomic models is difficult. Thus, atomic models from refinement are prone to errors and to develop a measure for these errors is crucial. In this paper, we have described the newly developed method for estimating the positional precision from real-space refinement of atomic structure against cryo-EM density maps. Our approach uses the cross-validation measure, C_{free}, which is already used during refinement. Here, we have tested the method on 5 Å and 7 Å simulated EM data. The atomic positional precision was computed by estimating the best ensemble with the cross-validation C_{free} value and its statistical error value. The estimated precision was 0.6138 Å for 5 Å density map and 0.7157 Å for 7 Å density map. The method is also robust and not sensitive to the choice of C_{free} band used during refinement.

4.5 Acknowledgement

Computational support and infrastructure was provided by the "Centre for Information and Media Technology" (ZIM) at the University of Düsseldorf (Germany).

Bibliography

[1] 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 (London, England : 1993)*, vol. 16, pp. 673–83, may 2008.

[2] G. F. Schröder, A. T. Brunger, and M. Levitt, "Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution," *Structure*, vol. 15, pp. 1630–1641, dec 2007.

[3] J. R. Lopéz-Blanco and P. Chacón, "iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates.," *Journal of structural biology*, vol. 184, pp. 261–70, nov 2013.

[4] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali, "Protein structure fitting and refinement guided by cryo-EM density.," *Structure (London, England : 1993)*, vol. 16, pp. 295–307, feb 2008.

[5] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, "Refinement of Macromolecular Structures by the Maximum-Likelihood Method," *Acta Crystallographica Section D Biological Crystallography*, vol. 53, pp. 240–255, may 1997.

[6] A. T. Brünger, "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures," *Nature*, vol. 355, pp. 472–475, jan 1992.

[7] V. Luzzati, "Traitement statistique des erreurs dans la determination des structures cristallines," *Acta Crystallographica*, vol. 5, pp. 802–810, nov 1952.

[8] D. W. J. Cruickshank, "ERRORS IN LEAST-SQUARES METHODS," in *Computing Methods in Crystallography*, pp. 112–116, Elsevier, 1965.

[9] T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard, "Improved methods for building protein models in electron density maps and the location of errors in these models," *Acta Crystallographica Section A Foundations of Crystallography*, vol. 47, pp. 110–119, mar 1991.

[10] D. W. J. CruickshankÂš, D. W. J. Cruickshank, and D. Cruickshank, "Remarks about protein structure precision," *Acta Cryst*, vol. 55, pp. 583–601, 1999.

[11] S. Daopin, D. R. Davies, M. P. Schlunegger, and M. G. Grutter, "Comparison of Two Crystal Structures of TGF-fl2: the Accuracy of Refined Protein Structures," *Acta Cryst*, vol. 0, pp. 85–92, 1994.

[12] B. Falkner and G. F. Schroder, "Cross-validation in cryo-EM-based structural modeling," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 8930–8935, may 2013.

[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, pp. 187–217, jan 1983.

[14] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, "An all atom force field for simulations of proteins and nucleic acids," *Journal of Computational Chemistry*, vol. 7, pp. 230–252, apr 1986.

[15] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren, "The GROMOS Biomolecular Simulation Program Package," *The Journal of Physical Chemistry A*, vol. 103, pp. 3596–3607, may 1999.

[16] H. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, pp. 43–56, sep 1995.

[17] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kale, R. D. Skeel, and K. Schulten, "NAMD: a Parallel, Object-Oriented Molecular Dynamics Program," *International Journal of High Performance Computing Applications*, vol. 10, pp. 251–268, dec 1996.

[18] B. Isralewitz, M. Gao, and K. Schulten, "Steered molecular dynamics and mechanical functions of proteins," *Current Opinion in Structural Biology*, vol. 11, pp. 224–230, apr 2001.

Chapter 5

Discussion and Outlook

Cryo-EM data obtained either from single particle analysis, electron tomography, or 2D crystals are often still in the intermediate- to low-resolution range even after the resolution revolution in the field. Due to this fact, there is a great demand for tools that can be used for better interpretation, visualisation, resolution enhancement and henceforth for better structure determination. In this thesis, three different methods have been developed that can be used for improved structure determination from cryo-EM data.

The main project of my thesis is discussed in chapter 2 - ELIGN: **EL**astic al**IGN**ment of Cryo-EM density maps. With the conformational heterogeneity in cryo-EM samples being the main limiting factor to reach optimum resolution, the only option available so far to handle the data is to sort the 2D images into different conformational classes and then to reconstruct 3D volumes for these classes individually. However, sorting the images reduces the number of images used in each 3D volume reconstruction and hence limits the resolution. We developed the ELIGN technique to overcome the resolution limitation due to conformational heterogeneity. The main idea behind this technique is to elastically align all the conformations to average the available data. ELIGN elastically fits all the different conformations to each other and is a good method to overcome the limitation of the resolution by the conformational heterogeneity.

ELIGN converts the density map to a pseudo-atomic model by placing point masses on each voxel of the map, thus allowing for a flexible deformation. The point masses carry the density information corresponding to each voxel. The results presented in section 2.2 shows the output of ELIGN with improved resolution of the cryo-EM data for both single particle analysis and cryo- electron tomography data. Here, we have shown the improvement of resolution for three data sets

and compared the results of ELIGN to the rigid averaging of all maps from different conformations. The enhancement in resolution by using ELIGN is shown both by the FSC curves and images of the density maps. First, ELIGN was shown to have resolution improvement on the simulated density maps of six different conformations of glutamate dehydrogenase. By ELIGN, the density map was improved from 2.17 Å to 2.09 Å whereas the rigid average was making the resolution worse to 3.6 Å. Second, for ten density maps of the E-coli ribosome single particle data an improvement of the resolution from 4.5 Å to 3.27 Å was shown. The resolution improved to only 3.57 Å by rigid averaging. Third, the sub tomogram averages from mammalian oligosaccharyl- transferase complex in native endoplasmic reticulum protein translocon were clipped to just have the translocon associated protein complex (TRAP) and the oligosaccharyl-transferase (OST) complex excluding the ribosome was sorted to three different sets of classes having different signal to noise ratios. The three different classifications (10, 30 and 170 classes) were used to test ELIGN and the improvement for all the classes were almost the same leading from 16 Å to 12 Å. For this data the rigid average of all the density maps was around 15 Å.

ELIGN is proved to be a robust algorithm allowing for deformation of the density map invariant of the direction of the deformation, difference in the conformational changes and without the need for the prior knowledge of the atomic structure. With our new method, it might be beneficial to sort images even further than usually, to a point where the resolution gets limited by the number of images, because after the elastic 3D alignment all data can be merged again, and hence making use of all image data. The application of the method can be extended from 3D density maps to 2D images. The application of the algorithm to 2D images should be really interesting because the elastic alignment on 2D images should be done as part of the 3D reconstruction process unlike the post processing step done in this thesis, which is carried out after the reconstruction. Hence, if the elastic alignment is used on 2D images, one could use all the information in all the images to reconstruct a single resolution-improved 3D map. Elastic fitting on the 2D level will make it possible to obtain class averages with high signal-to-noise ratio and thus a high resolution reconstructed density map.

Moreover, a new method was developed to improve the visualization of cryo-EM density maps. This method, known as VISDEM, is a part of the program DireX, and is described in Chapter 3. Optimal visualization is a must for proper interpretation of the data. Here, general statistical information about proteins, especially the packing density of atoms, is used to sharpen the density maps. A pseudo-atomic model with approximate mass distribution of the protein can be built from the packing density of the atoms, since the packing density is very similar for different proteins. After building the pseudo-atomic model, constraints for both real and reciprocal space reconstruction are determined from the pseudo-atomic model. The estimated density histogram is used as a constraint in real space and the radial structure factor is used in reciprocal space reconstruction. Hence the name VISDEM implying **V**isualization **I**mprovement by **S**tructure factor and **D**ensity histogram correction for cryo-**EM** data.

In the paper, we have shown sharpened density maps using VISDEM for three different resolutions taken from the EMBD. They are the transient receptor potential channel V1 (TRPV1) (EMDB-5778), chaperonin GroEL/ES density map (EMDB-2325) and the mycobacterium tuberculosis fatty acid synthase multienzyme complex (EMDB-2538) with a resolution of 3.3 Å, 8.9 Å and 18 Å respectively. The results show that the VISDEM sharpened maps are better than the B-factor sharpened maps and comparable to the sharpened maps by using the correct radial structure factor and density histogram applied from the crystal structure. The improvement of visualization by VISDEM was quantified by both the FSC curve and the density correlation coefficients between the VISDEM maps and the atomic model maps computed from the crystal structures.

Finally, a new method has been developed to estimate the positional error of atomic coordinates during real space refinement, which is discussed in detail in Chapter 4. Atomic models built and refined with intermediate-resolution (3.5–5 Å) cryo-EM density maps are prone to error since such density maps do not provide clear atomic details. A clear estimate of the error of the atomic coordinates helps to judge the level of reliability of the determined structures. There are well established

methods in X-ray crystallography to measure the precision of refined structures. In contrast, in cryo-EM no method is available to determine these errors. Our method determines the error of the positional precision during real space refinement of cryo-EM data.

The approach presented here is based on the cross-validated cross-correlation coefficient, C_{free}, which measures the fit of a model to the data. The refinement program DireX computes the C_{free} value during the refinement. Ideally, the model that yields the highest C_{free} value should be the best model, closest to the true structure. However, the C_{free} value has a statistical error which means all models within a certain range of C_{free} values have the same estimated quality and cannot be distinguished. Here, the positional error is defined as the root mean square fluctuation (RMSF) of all models within this range of C_{free} values. Ensembles of models generated with different approaches have been tested. For the generation of the ensembles MDFF and DireX with different restraints were used, yielding different extent of conformational sampling. The measured positional precision during real-space refinement of atomic structure against 5 Å and 7 Å density maps were shown in the result section of Chapter 4.