Peering Beyond the Noise in Experimental Biophysical Data

Dari Kimanius

Doctoral Thesis in Biochemistry towards Bioinformatics at Stockholm University, Sweden 2019
Peering Beyond the Noise in Experimental Biophysical Data

Dari Kimanius

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry towards Bioinformatics at Stockholm University to be publicly defended on Friday 24 May 2019 at 14.00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 16 B.

Abstract

Experimental protein structure determination methods make up a fundamental part of our understanding of biological systems. Manual interpretation of the output from these methods has been made obsolete by the sheer size and complexity of the acquired data. Instead, computational methods are becoming essential for this task and with the advent of high-throughput methods the efficiency and robustness of these methods are a major concern. This work focuses on the computational challenge of efficiently extracting statistically supported information from noisy or significantly reduced experimental data.

Small-angle X-ray scattering (SAXS) is a method capable of probing structural information with many experimental benefits compared to alternative methods. However, the acquired data is a noisy reduction of a large set of structural features into a low-dimensional signal-mixture, which significantly limits its interpretability. Due to this SAXS has this far been limited to conclusions about large-scale structural features, like radius of gyration or the oligomeric state of the sample. In this thesis I present an approach where SAXS data is used to guide molecular dynamics simulations to explore experimentally relevant conformational states. The experimental data is fed into the simulations through a metadynamics protocol, which explores the experimental data through conformational sampling subject to thermodynamic restraints. I show how this approach makes it possible to use SAXS to produce atomic-resolution models and make further-reaching conclusions about the underlying biological system, in particular by showcasing de novo folding of a small protein.

Another experimental method that generates noisy and reduced data is cryogenic electron microscopy (cryo-EM). Due to recent development in the field, the computational burden has become a considerable bottleneck, which greatly limits the throughput of the method. I present computational techniques to alleviate this burden through the use of specialized algorithms capable of efficient execution on graphics processing units (GPUs). This work improves the computational efficiency of the entire pipeline by several orders of magnitude and significantly advances the overall efficiency and applicability of the method. I show how this enables the development of improved algorithms with increased capabilities for extracting relevant biological information form the data. Several such improvements are presented that significantly increase the resolution of the refinement results and provide additional information about the dynamics of the system. Additionally, I present an application of these methods to data collected on a biogenesis intermediate of the mitochondrial ribosome. The new structures provide insights into the timing of the rRNA folding and protein incorporation as well as the role of two previously unknown assembly factors during the final stages of ribosome maturation.

Keywords: Cryo-EM, mitochondrial ribosome, SAXS, molecular dynamics, metadynamics.

Stockholm 2019
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-167809

ISBN 978-91-7797-712-4

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm
PEERING BEYOND THE NOISE IN EXPERIMENTAL BIOPHYSICAL DATA

Dari Kimanius
Peering Beyond the Noise in Experimental Biophysical Data

Dari Kimanius
## Contents

List of Abbreviations iii  
List of Papers v  

1 Introduction 1  
1.1 Molecular Biology ........................................ 1  
1.2 Measurement, Error & Noise ............................. 2  
1.3 Experimental vs. Computational Data ...................... 2  
1.4 Protein Structure & Thermodynamics ...................... 3  
1.5 What is this thesis about? ................................. 4  

2 Experimental Techniques 7  
2.1 Structural Methods ......................................... 7  
2.2 SAXS – Small-Angle X-ray Scattering .................... 8  
2.3 The Electron Microscope ................................. 9  
2.4 HRTEM – High-Resolution Transmission Electron Microscopy 11  
2.4.1 Image Formation ...................................... 12  
2.5 Cryo-EM – Cryogenic Electron Microscopy ............... 13  

3 Biomolecular Simulations 15  
3.1 Molecular Dynamics ........................................ 15  
3.2 Enhanced Sampling ......................................... 18  
3.2.1 Replica exchange ....................................... 18  
3.2.2 Steered MD ............................................ 18  
3.2.3 Metadynamics .......................................... 19  

4 Computational Single-Particle Cryo-EM 21  
4.1 Beam-Induced Motion & Radiation Damage ................ 22  
4.2 CTF-Estimation, Particle Picking & Extraction ............ 22  
4.3 2-D/3-D Classification & Reconstruction .................. 23  
4.3.1 Maximum Likelihood .................................. 24  
4.3.2 Classification ......................................... 25  
4.3.3 Refinement ............................................. 26
## Contents

4.3.4 FSC – Fourier Shell Correlation ........................................ 26  
4.4 Hardware acceleration ....................................................... 27  
  4.4.1 Parallelism .............................................................. 27  

5 Biological Applications .................................................. 29  
  5.1 The Ribosome ..................................................................... 30  
    5.1.1 Protein Synthesis ......................................................... 30  
    5.1.2 The Mitochondrial Ribosome ......................................... 31  
  5.2 How to Capture the Conformational States of the Ribosome ... 31  

6 Summary of Papers ......................................................... 35  
  I Accelerated cryo-EM structure determination with parallelisation  
    using GPUs in RELION-2 .................................................. 35  
  II SAXS-guided Metadynamics ............................................... 35  
  III Characterisation of molecular motions in cryo-EM single-particle  
    data by multi-body refinement in RELION ............................ 36  
  IV Structure of the human mitochondrial ribosome in native states of  
    assembly ........................................................................... 36  
  V New tools for automated high-resolution cryo-EM structure determi-  
    nation in RELION-3 ......................................................... 37  
  VI Cryo-EM reconstruction of the chlororibosome to 3.2 Å resolution  
    within 24 h ....................................................................... 37  

7 Sammanfattning ............................................................. 39  

8 Acknowledgments .......................................................... 41  

A Appendix ............................................................................ 43  
  A.1 Diffraction Intensity of Dilute Solutions of Identical Particles .. 43  
  A.2 The Expectation-Maximization Algorithm .......................... 44  

Appendices  

References ................................................................. 47
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D, 2-D, 3-D</td>
<td>One-, two- or three-dimensional, referring in this thesis to a data sequence, spatial dimensions in an image or volume.</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio.</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics, a computational methods where classical approximations are used to simulate molecular structures and interactions.</td>
</tr>
<tr>
<td>FES</td>
<td>Free energy surface, referring in this thesis to the thermodynamic potential, also known as Helmholtz free energy.</td>
</tr>
<tr>
<td>SAXS</td>
<td>Small-angle X-ray scattering</td>
</tr>
<tr>
<td>SANS</td>
<td>Small-angle neutron scattering</td>
</tr>
<tr>
<td>MetaD</td>
<td>Metadynamics</td>
</tr>
<tr>
<td>CV</td>
<td>Collective variable or reaction coordinate</td>
</tr>
<tr>
<td>EM</td>
<td>Electron microscopy</td>
</tr>
<tr>
<td>TEM</td>
<td>Transmission electron microscopy</td>
</tr>
<tr>
<td>SEM</td>
<td>Scanning electron microscopy</td>
</tr>
<tr>
<td>Cryo-EM</td>
<td>Cryogenic electron microscopy</td>
</tr>
<tr>
<td>FSC</td>
<td>Fourier shell correlation</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>LSU</td>
<td>Ribosomal large subunit</td>
</tr>
<tr>
<td>SSU</td>
<td>Ribosomal small subunit</td>
</tr>
</tbody>
</table>
List of Papers

Paper I Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2
Kimanius D., Forsberg B. O., Scheres S. H. and Lindahl E.

Paper II SAXS-guided Metadynamics
Kimanius D., Pettersson I., Schluckebier G. and Lindahl E.

Paper III Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION
Nakane T., Kimanius D., Lindahl E. and Scheres S. H.

Paper IV Structure of the human mitochondrial ribosome in native states of assembly
Brown A., Rathore S., Kimanius D., Aibara S., Bai X. C., Rorbach J., Amunts A. and Ramakrishnan V.

Paper V New tools for automated high-resolution cryo-EM structure determination in RELION-3
Zivanov J., Nakane T., Forsberg B. O., Kimanius D., Hagen W. J., Lindahl E. and Scheres S. H.

Paper VI Cryo-EM reconstruction of the chlororibosome to 3.2 Å resolution within 24 h
Forsberg B. O., Aibara S., Kimanius D., Paul B., Lindahl E. and Amunts A.
Introduction

1.1 Molecular Biology

Living cells are vastly complex self-regulatory systems that utilize advanced chemical processes to grow and reproduce in nearly every part of the surface of the Earth. The evolutionary process that has formed each cell incorporates all factors of its past history, including availability of resources, threats from its competitors, viruses, toxins and symbiotic relations to other cells in its surrounding or host organism. Every cell will have to remain responsive to changes in its environment through cascades of molecular signaling pathways throughout its life span. It will have to maintain an osmotic balance across its outer membrane, repair damaged internal structures, handle accumulation of waste, and perhaps most importantly manage and maintain its genome.

The cell has a variety of chemical compounds at its disposal, including amino acids, nucleic acids, lipids and sugars. It will consume these resources by absorbing them from the environment through import mechanisms or producing them through complicated metabolic pathways, similar to assembly lines in a factory.

To cater for all the ongoing tasks, the cell manufactures molecular complexes, predominantly proteins, tailored for each task. Proteins are versatile molecules made up of long chains of linked amino acids coded for by the genome. During (or shortly after) the manufacturing process, each protein folds into a specific 3-D conformation, or structure, that is dictated by its 1-D sequence of amino acids and environmental factors, e.g. temperature and salt concentration. The structure in turn dictates the biochemical properties of the protein and the role it plays in the web of interacting biomolecules filling the insides of the cell.

Biological systems, as simple as a single cell, can never be understood without an understanding of their molecular building blocks. Hence scientific methods for examining molecular systems are essential in our endeavor to investigate biological phenomena. These systems are far too small to be studied directly. Fortunately,
there is a large diversity of methods at our disposal that enable some insights. These methods generate large amounts of data that have to be interpreted before conclusions can be drawn.

This thesis is focused on the computational part of such methods, in particular the bridge between experimental data and theoretical calculations. It presents algorithms for efficient calculations and rigorous methods to extract information from noisy experimental data and avoid erroneous conclusions. At the heart of this work lies thorough considerations about management of uncertainty in experimental measurements to ensure substantially supported conclusions about biophysical systems.

1.2 Measurement, Error & Noise

The physical reality seems to be composed exclusively of non-ideal objects and events. There appears to be no perfect circles or lines and thus models describing our surrounding world will always remain approximative, no matter how accurate they get. This implies that every observation or measurement of a physical object or event is flawed and will be associated with an error. The scientific method relies heavily on inference, where typically multiple observations are used as support for one or multiple conclusions. The error in the observations will propagate to the conclusion and impact its certainty. Although the relative scale of the error might not be of significance to the particular conclusions drawn at one time, it is not given that it will hold for a different purpose. Error management is of central importance in every scientific study and thus every major conclusion should preferably be presented with its associated uncertainty.

Error is generally divided into systematic and random error, or bias and noise. In most cases the bias can be corrected for, e.g. by calibration. Noise on the other hand is far more difficult to handle and specialized algorithms are usually needed to correct for it. The information content in data is referred to as signal. This is the patterns in the data that contain the relevant information. In this view, any distortion causing the signal to diverge from its true value is considered noise. The ratio between the amplitude of signal and noise in the data, namely the signal-to-noise ratio (SNR), is a quantity frequently used in data processing. As the SNR drops, it becomes more difficult to detect the signal and thus the risk for misinterpretations, i.e. over-fitting, increases. Over-fitting is said to occur when the information extracted from the data lacks true statistical support within it. Most algorithms designed for processing noisy data, i.e. data with low SNR, are equipped with safeguards against over-fitting.

1.3 Experimental vs. Computational Data

Any form of meaningful scientific conclusion, even in the experimental sciences, usually involves some type of computation. The concentration of a sample for instance cannot be observed directly, why we rely on indirect measurements, such
as light absorbance, and use a computational model to infer a numerical value for the concentration. However, when speaking of computational data this is commonly not what is referred to. Computational data is usually the result of a long pipeline of algorithms that significantly alter or augment the information content of the input data. More established pipelines are increasingly becoming an integral part of experimental equipment. In fields like genomics or proteomics, where tremendous amount of data is produced with high-throughput methods, such computational pipelines are essential to make sense of the data. Here the purpose of the computational part is to augment the data into a representation with a higher information density in a rigorous way.

This thesis is primarily focused on cryogenic electron microscopy (Cryo-EM), which is a method that has many similarities with high-throughput methods, foremost due to the volume and low SNR of the data produced. The computational pipeline is extensive and consists of a series of algorithms that successively compress the information content of the data into a highly condensed representation.

Computations can also be used to either extrapolate from experimental data or interpolate between experimental data points into regions where data is missing. Rather than focus on data interpretation, these methods rely on a more complete physical model of molecular systems to provide farther-reaching conclusions, beyond what is directly available in the experimental data. Established theories in the fields of quantum mechanics and thermodynamics are used to approximate the behavior of molecular systems and make predictions.

I will also present methods where simulations of molecular systems are used together with incomplete experimental data to fill in missing information.

1.4 Protein Structure & Thermodynamics

The chemical properties of a biological molecule are strongly associated with the structural conformations displayed by it [1]. The probability of observing a system in a certain conformation is related to the free energy of that conformation. The so-called free energy surface (FES) describes the variations in free energy as the system moves between different conformational states. To understand the concept of free energy it is beneficial to distinguish between macro- and microstates. Where the line is drawn between these two depend on the observer. In protein systems the conformation of individual atoms can be considered a microstate, whereas the position of a helix can be considered a macrostate. Commonly, a distinction is made between the folded and the unfolded state of a protein, which is another example of macrostates.

Quantum mechanics can be used to accurately calculate the enthalpy of a given state of a protein, however the free energy of a macrostate is not determined only by its enthalpy. The entropic contribution to the free energy is proportional to the logarithm of the multiplicity of the state, i.e. the number of states encapsulated by the macrostate.
Given a temperature, the free energy \( F_s \) of the macrostate \( s \) is

\[
F_s = E_s - TS_s = E_s - T k_B \log \Omega_s,
\]

(1.1)

where \( \Omega_i \) is the number of microstates, \( E_s \) is the enthalpy and \( S_s \) is the entropy for macrostate \( s \). \( k_B \) is the Boltzmann constant.

The probability \( \rho_s \) of observing a conformation with the free energy \( F_s \) is in turn proportional to the Boltzmann factor,

\[
\rho_s \propto \exp \left( -\frac{F_s}{k_B T} \right) .
\]

(1.2)

The relative population, \( \rho_{i,j} \), of two states \( i \) and \( j \) can then be expressed as

\[
\rho_{i,j} = \frac{\rho_j}{\rho_i} = \exp \left( -\frac{F_j - F_i}{k_B T} \right) = \exp \left( -\frac{\Delta F}{k_B T} \right) ,
\]

(1.3)

where \( \Delta F \) is the change in free energy. A physical system will thus strive to minimize its free energy, since this will serve to place it in the state with the highest probability [2]. However, the population of two states with different free energies will become more similar as the temperature of the system increases, because there will be more thermal energy available to the system, which reduces the driving force towards lower energy states. A reduction in temperature on the other hand will cause the system to get stuck in local free energy minima.

The folding process of a protein can be understood from this standpoint. It is a stochastic process in which the protein randomly explores the large number of conformations available to it, given the thermodynamically available energy. This is simply the available thermal energy in its surrounding, which causes the protein to fluctuate between states. The number of available conformations grows exponentially as more and more amino acids are added to the chain. The conformational space quickly becomes too large to ever be explored by the system in its limited lifetime [3–5]. Hence proteins are not generally selected by evolution to explore their entire conformational space to find the native state [6, 7]. Instead the free energy landscape of most larger proteins are shaped by evolution to guide the folding dynamic towards the native state. More specifically the folding process consists of a number of semi-stable intermediate states that collectively make up a folding pathway across the FES, from the initial unfolded state to the folded state. At each folding step along the pathway the positive entropic contribution is compensated immediately or nearly immediately by an enthalpic contribution [2].

1.5 What is this thesis about?

Given the complex folding process of proteins and the vast conformational space available to even the smallest protein it is apparent that simulating these natural processes is difficult. To make such computations feasible many approximate approaches have been proposed and implemented. However, despite these efforts purely computational methods are still not efficient enough to investigate molecular
1. Introduction

phenomena of biological relevance [8]. To fully understand the underlying mechanisms that give rise to the chemical diversity and vast complexity of biological systems we instead have to employ specialized method that can incorporate experimental data into advanced computational procedures. Biophysical simulations for instance are based on a so-called forward model that can be used for augmenting experimental data and extracting all the available information.

The following chapters will outline methods for extraction of relevant biological information from experimental data and incorporation of them into computational methods. The uncertainty introduced by the noise in the experimental data will be investigated and methods for noise reduction and management presented.
Experimental Techniques

Contents

1.1 Molecular Biology ................................. 1
1.2 Measurement, Error & Noise .................. 2
1.3 Experimental vs. Computational Data ........ 2
1.4 Protein Structure & Thermodynamics ......... 3
1.5 What is this thesis about? ...................... 4

Large scale biological phenomena, like cellular reaction to external stimuli or multi-cellular coordination in a neural circuit, can be broken down into nanoscale molecular mechanisms. To fully understand biological systems we have to understand their fundamental building blocks. Structural biology brings together methods from molecular biology, biochemistry and biophysics for investigating the anatomy of biological macromolecules, and by extension their functional and dynamical properties. This chapter will outline some relevant experimental techniques at the heart of this field.

2.1 Structural Methods

Cells are in a sense mass-production facilities of biomolecules. Each cell can produce large quantities of replicas of individual molecules. Due to the prevalence of this effect most experimental methods in biology rely on a multitude of the target molecules. Instead of looking at individual molecules, data is collected for multiple identical molecules and compiled into one or several images of the molecule.

One example of this is how the quantum mechanical phenomenon of nuclear magnetic resonance (NMR) can be used to discern the presence of certain atom types and their position relative to their chemical environment. This data is then
used to calculate the complete molecular structure of the complex. Particularly in solution NMR, the sample can remain in an aqueous solution similar to its native environment. Unfortunately, as the number of atoms grows, it becomes exponential more difficult to separate the different signals, why this method is generally limited to complexes smaller than 25 kDa [9].

X-ray crystallography is another famous method, which requires that the complex is crystallized prior to measurement. The crystal is placed in a highly coherent X-ray beam that diffracts as it interacts with the crystal. The diffraction pattern produced by the crystal can then be used to calculate the 3-D structure of the complex through Bragg’s law [10]. The limiting step in this method (similar to solution NMR) is the requirement of the large quantities of the sample, needed for crystallization. If the complex does crystallize, there are no size limitations like the ones for NMR. However, as complexes become larger they usually become more difficult to crystallize and depending on the flexibility and stability of the complex it is not given that they will ever crystallize. Furthermore, it is difficult to estimate the effect of the crystallization process on the structure of the complex. This is all the more substantial for flexible structures more restricted in a crystal lattice. For the same reason crystallography experiments typically do not yield more than a single structural state of the complex.

One of the main hurdles in crystallography is the fact that the structural information is encoded into the diffraction pattern both in the amplitude, which is easily measured, but also in the phase shift of the X-rays. Unfortunately this information can not be easily retrieved experimentally and is usually estimated computationally [11]. The most common approach is to start from an approximate atomic model that is iteratively refined to better match the experimental data [12].

There are many pitfalls in all structural determination methods caused by noise. Although noise can never be completely eliminated, many measures can be taken to reduce its effects. Heterogeneity is in general considered a major contributing factor to noise. A sample is said to be heterogeneous if it is polydisperse or contains structural variations. Thus, for most samples tremendous work is invested in the purification process prior to the experiment, i.e. during sample preparation. In crystallography a sample is not likely to even crystallize unless in a highly purified condition. Polydisperisty in solution NMR renders a mixture of signals from the variations in particle type (i.e. spectral crowding), hence making the data more difficult to interpret [13].

2.2 SAXS – Small-Angle X-ray Scattering

If the crystallization step in X-ray crystallography is skipped, and the sample is placed in the X-ray beam in its solvated state, a similar diffraction pattern is produced. However, since the sample is not rectified in a crystal lattice the diffraction signal will be rotationally averaged. This removes the drawbacks that come with the process of crystallization and permits it to be applied to structurally fragile complexes [14]. This method is called small-angle X-ray solution scattering (SAXS) and generates a 1-D spectrum from the sample. The drastic information reduction
from 3-D to a 1-D means the signal becomes tremendously difficult to interpret. There are infinite numbers of molecules and/or conformations that can account for any given SAXS-profile of a protein. SAXS-data can never be used for *de novo* structural determination, but rather for measuring overall structural features, e.g. radius of gyration \[15\]. It is noteworthy is that instead of X-rays one can use neutrons in this type of experiment, which is called *small-angle neutron scattering* (SANS). All theory presented in this work related to SAXS can be expanded to SANS.

The measured intensity from a SAXS experiment can be matched with an atomic model through the *Debye Formula*. The diffraction intensity of a monodisperse solution of \(N\) molecules is

\[
I_N(Q) = N \sum_i \sum_j f_i f_j \frac{\sin Qr_{ij}}{Qr_{ij}},
\]

where \(Q\) is the diffraction angle and \(r_{ij}\) is the pairwise distance between atoms \(i\) and \(j\), with form factors \(f_i\) and \(f_j\), respectively. See appendix A.1 for a derivation of this expression.

In paper II a computational methods is presented that aims to interpret the information encoded in SAXS data by enforcing thermodynamically valid system conformations. The Debye formula is a central piece in this method, since it can be used to match an atomic model of a protein to a SAXS intensity profile. Furthermore it is easily differentiable, which is essential to the method (see chapter 3). Generally this type of approach can be applied to any type of experimental data that is too noisy or too reduced to be used for direct inference about molecular structure or function. Instead a forward model is employed through which the information present in the data is disentangled and this is a recurring topic in this thesis.

### 2.3 The Electron Microscope

Light microscopy can be used to study details on the micrometer scale. The lower bound of what can be resolved with light is set by the wavelength of visible light \[16\]. This limit is given by Abbe’s diffraction limit \(d = \lambda/(2 \times NA)\), where \(\lambda\) is the wavelength and \(NA\) is the numerical aperture \[17\]. Modern optical systems can achieve a numerical aperture of about 1.4-1.6, which puts the diffraction limit at roughly 100 nm for visible light. As these resolutions are approached, it becomes increasingly difficult to distinguish details in the acquired images and the optical system is said to be diffraction-limited. Although this limit is pushed in super-resolution microscopy \[18–20\], to be able to push it to Ångström resolution, where atomic details start appearing, probing particles with shorter wavelengths are required. Light with shorter wavelengths, e.g. X-rays, interacts weakly with matter, why there are no optical lenses that can bend them efficiently and accurately to generate a magnified image. In crystallography the diffraction within the sample is used to generate intensity measurements in Fourier space, which has some drawbacks.

Electron microscopy (EM) is a method where electrons are used as the probing particles, rather than photons. Many similarities exist between this methods and light microscopy, despite their use of fundamentally different particles. Electrons,
Figure 2.1: (A) Experimental setup for X-ray crystallography. (B) Schematic example of acquired diffraction spots and the central unscattered beam. (C) Experimental setup for small-angle X-ray solution scattering. Defining the scattering angle $\theta$. (D) Schematic example of acquired rotationally averaged SAXS diffraction profile. (Illustration by the author)

In contrast to photons, are electrically charged and will be subject to the Lorentz force, $\mathbf{F}$, as they move through a magnetic field $\mathbf{B}$. This force is given as

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}. \quad (2.2)$$

Here $q$ denotes the elementary charge and $\mathbf{v}$ the velocity of the electron.

Through this phenomenon, magnetic lenses can be used to manipulate the path of an electron beam in a similar fashion as for light and thus make it possible to use them to achieve magnified image formation.

The wavelength of matter particles is given by the de Broglie wavelength through the Planck constant, $h$, as $\lambda = h/p$, where $p$ is the momentum of the electron, which with the relativistic effect accounted for can be approximated as

$$pc \approx \sqrt{2 \times E_K \times m_0 c^2}. \quad (2.3)$$

Here $E_K$ is the kinetic energy of the electron, $m_0$ its resting mass, and $c$ the speed of light.

For an electron accelerated from rest to 1 keV, the corresponding de Broglie wave-
2. Experimental Techniques

Figure 2.2: (A) HRTEM imaging. Illustration of the scattered wave function in different planes along the optical axis in the microscope column. (B) Acquired image in real space, showing ribosomal particles. (C) The Fourier representations of image in (B) together with a slice showing the fitted contrast transfer function. (Illustration by the author)

Length is

\[
\lambda = \frac{hc}{pc} \approx \frac{12.4 \text{ keV} \times \text{Å}}{\sqrt{2} \times (1000 \text{ eV}) \times (0.511 \text{ MeV})} = 0.4 \text{ Å}. \tag{2.4}
\]

Therefore, even at low energies the electron wavelength is comparable to the atomic radius, which makes them excellent for probing atomic details. In high-resolution transmission electron microscopy (HRTEM) wavelengths smaller than 4 pm are achieved by accelerating the electrons with voltages above 100 kV. In applications within material science, acceleration up to a megavolt can be used [21].

2.4 HRTEM – High-Resolution Transmission Electron Microscopy

There are several different applications for electron microscopes. For instance in scanning electron microscopy (SEM), electrons are reflected off the sample before they hit the detector, which yields sharp images of microscopic surface topologies. SEM is generally limited to resolutions in the nanometer range and above, which is well suited for investigating the surface of objects like insects and whole cells.

In Transmission electron microscopy (TEM) the detector is positioned on the opposite side of the sample from the electron source. Hence, the detected electron beam has passed through the sample before hitting the detector and carries information from the entire cross-section of the sample. Viewed from the detector, the 3-D sample content is projected into the 2-D plane of the detector. The thicker the sample is, the more data is loaded onto the beam and it becomes more difficult to extract the relevant information. Therefore, great effort is invested in making the sample as thin as possible.
2.4.1 Image Formation

The type of imaging in TEM is called phase-contrast imaging, where the difference between the propagation speeds of the probing particle in different materials (i.e., the reflective index) are used to generate contrast in the resulting images. Therefore a beam with high temporal coheres is desired. As the electron beam moves through the sample there is a non-zero probability that it will pass through and not interact with the sample. The phase-difference between these particles and the ones that do interact with the sample generate constructive and destructive interference similar to what is observed in a crystal lattice.

Formally the phase $\rho(r)$, dependent on the 2-D coordinate in the image plane, $r$, is generated through the electrostatic crystal potential of the sample, $V_C(r, z)$ as

$$\rho(r) = \frac{\pi}{\lambda U_A} \int_{z_0}^z V_C(r, z) dz .$$

Here $U_A$ is the acceleration voltage and the integral is across the thickness of the sample. Assuming a small $z_0$ and a small scattering angle (i.e., the weak phase object approximation) we can express the scattered wave function as $\psi_{sc}(r) = \exp(-i\rho(r)) \approx 1 - i\rho(r)$. According to Bragg’s law it will generate a Fourier transform as it propagates through the microscope column and reaches the so-called back focal plane [22]. Hence the wave function at this state can be written as

$$\psi_{bfp}(q) = F[\psi_{sc}(r)](q) = \psi_{sc}(q) ,$$

where $F$ denotes the Fourier transform and $q$ is the 2-D reciprocal coordinate. After the scattered beam passes the back focal plane it enters into the region where the microscope lenses focus it onto the detector, placed in the image plane. One might expect that this is a plain inverse Fourier transform of the wave function at the back focal plane, $F^{-1}[\psi_{bfp}(q)]$. However, it is at this point that the impact of the transfer characteristics of the lenses become relevant [22]. This effect can be accounted for by the transfer function, $t(q)$, which instead yields a wave function at the image plane given as

$$\psi(r) = F^{-1}[\psi_{sc}(q)t(q)] .$$

With the weak phase object approximation the detected real space intensity becomes

$$I(r) = \psi(r)\bar{\psi}(r) = 1 + 2\{\rho(r) \ast C(r)\} ,$$

where $\ast$ denotes the convolution operator, $C(r)$ is called the contrast transfer function (CTF) and equals the imaginary part of $t(r)$ [22].

The CTF for an image, in its simplest form, depends on the defocus, $c_1$, and the third order spherical aberration, $c_3$, which are configurations of the microscope optics at the moment the image was acquired. Ignoring astigmatism it can be defined as

$$C(q) = \sin \left( -\frac{\pi}{\lambda} \left( q^2 c_1 + \frac{1}{2} q^2 c_3 \right) \right) .$$

This function modulates the image contrast as well as the signal, why it is essential for estimating the SNR at different values of $q$ (see section 4). Usually it is estimated from the image itself.
2.5 Cryo-EM – Cryogenic Electron Microscopy

Biological macromolecules are dynamical systems commonly suspended in water or lipids and thus subject to Brownian motion [23, 24]. To make it possible to study them in an electron microscope to higher resolution, Dubouchet and coworkers developed the method of cryofixation through plunge freezing in the 1980s [25, 26]. This method is used extensively in modern cryogenic electron microscopy (Cryo-EM). In the most common variation of this method the biological specimen is placed on a flat copper grid mesh that is ultra-rapidly cooled to liquid nitrogen temperatures (-196 °C). This prevents the formation of crystalline ice, which would otherwise destroy or obscure the target molecule [27]. The type of ice that is instead formed is called vitreous or amorphous ice, since it is in a glass like state. At this state the sample is also protected from the vacuum present in the microscope column that otherwise would vaporize it.

This technique is used in three closely related methods: electron crystallography, electron cryotomography and single-particle cryo-EM. This thesis will be limited to the later of these three, where a large number of 2-D images of the biological molecule is combine to reconstruct an accurate 3-D structure.

Perhaps the most important factor in the development of single-particle cryo-EM has been the advent of direct electron detectors. They use a CMOS-based active pixel sensor and were proposed as an alternative to CCD cameras that indirectly detect electrons via a device called a scintillator, which renders a significant improvement in the SNR [28–31]. In addition, state of the art direct electron detectors can collect data at sufficient speed to enable a single collection to be divided over multiple frames, similar to the operation mode of a movie camera [32]. This enables correction for the so-called beam-induced motion (see section 4.1), which significantly reduces motion blur that otherwise would be a considerable resolution limiting-factor.

Additionally large quantities of data can be collected in a relatively short amount of time, which allows for more sophisticated computational methods to mine the data. This type of approach enables a great deal of flexibility and has been an important part of the success of the method.
If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

— John von Neumann

3
Biomolecular Simulations

Contents

2.1 Structural Methods ................................................. 7
2.2 SAXS – Small-Angle X-ray Scattering .......................... 8
2.3 The Electron Microscope ........................................... 9
2.4 HRTEM – High-Resolution Transmission Electron Microscopy ........................................... 11
  2.4.1 Image Formation ............................................... 12
2.5 Cryo-EM – Cryogenic Electron Microscopy .................. 13

Computational chemistry encapsulates a large variety of methods with the purpose to investigate physical and chemical properties of atomistic systems, which spans from crystalline metal alloys in material science or small organic molecules to more complex heterogeneous macromolecules in aqueous solutions. These methods strive to computationally reproduce models of atomic ensembles that collectively display realistic behavior, with the purpose to predict and explore observable mesoscopic or macroscopic system properties. In material science this can be conductivity, strength or elasticity, and in biology the binding affinity of a drug to its target receptor. Collectively these methods rely on the fact that atomistic phenomena are the fundamental factors in all the behaviors of chemical systems. The choice of method depends on the computational capability of the method to generate sufficient data and the reliability of that data for answering the relevant questions.

3.1 Molecular Dynamics

Ideally, the time-dependent relativistic Schrödinger equation can accurately describe the time evolution of a quantum mechanical system. However, this level of accuracy quickly becomes too complicated for systems more complex than the
equilibrium state of a few atoms. Popular methods based on the Born-Oppenheimer
approximation, like the Hartree–Fock method, have a time complexity of $O(N^4)$,
where $N$ is the number of atoms [33]. Due to their accuracy these methods have
had tremendous success in organic chemistry, where they are used to simulate
simple systems subject to a negligible entropic effect.

To expand beyond organic chemistry and explore for instance macromolecular
complexes consisting of thousands or millions of atoms, further approximations have
to be made. The large entropic effect on the thermodynamics and kinetic of these
systems require simulations that sample phase space for a valid estimation of their
free energy. At this point, models based on a quantum mechanical description fail
to scale and need to be augmented by approximations. Molecular dynamics (MD)
simulations are based on empirically parameterized classical approximations, where
atoms are considered the smallest units in the calculations. Numerical integrators
are used to integrate Newton’s equation of motion

$$m_i \frac{\delta^2 \mathbf{r}_i}{\delta t^2} = \mathbf{F}_i, \quad i = 1..N$$  \hspace{1cm} (3.1)

for a system of atoms with masses $m_i$ and positions $\mathbf{r}_i$. At each step of
the integration, the force $\mathbf{F}_i$ acting on each atom $i$ in a system of $N$ atoms, can be
calculated through the potential energy, $U$, as

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}_1, ..., \mathbf{r}_N)$$  \hspace{1cm} (3.2)

where $\nabla_{\mathbf{r}_i}$ is the differential operator with respect to the position of atom $i$.

Although this approach might not yield an accurate temporal evolution of
the system (since the exact initial state is unknown), it provides a relevant
sampling of the conformational space [34–36]. Because experimental measurements
(i.e. observations of macroscopic properties) are time-averages of the available
confrontational states, they are equivalent to the ensemble-averages generated
through simulations, given sufficient sampling (i.e. ergodicity). This is further
assuming that the potential, $U$, accurately describes the interactions of the system.

There are many well established MD engines; among the most known are
Gromacs, NAMD, CHARMM and Amber [34, 36–38]. While these software packages
provide an efficient implementation of the MD integrator, the definition of the
potential function $U$ is distributed as force fields. Large effort has been invested
into these parameters to maximize the similarities between simulation results and
known biochemical data [39–42].

Although the accuracy of force fields is of critical importance, many other
physical factors have to be accounted for in MD to produce realistic ensembles.
Real-world experiments in biology predominantly take place at constant pressure
and temperature. To produce isothermal-isobaric ensembles with the same statistical
properties as real physical systems, algorithms called thermostats and barostats
are incorporated into simulations. The purpose of a thermostat is to maintain, on
average, a target temperature. This is done by weakly coupling the system with a
heat bath and effectively adding or removing kinetic energy by rescaling the atomic
velocities at each integration step. Many suggested algorithms, like the Berendsen
Figure 3.1: Illustration of a simulation box of a lactose permeate transmembrane protein (purple), consisting of 12 α-helices inserted in a lipid bilayer (transparent grey) and surrounded by explicit water molecules (blue). This system facilitates the transport of lactose across the periplasmic membrane in bacteria. Water molecules are seen entering into the protein channel (from below) as the protein is in its open state facing the extracellular side of the membrane. (Illustration by author [55])

thermostat, have been show to cause unphysical artifacts, e.g. by violating the principle of equipartition of energy [43–45]. Current state of the art production simulations usually use the Bussi–Donadio–Parrinello thermostat that is know to efficiently produce accurate physical ensembles [46]. Similarly, the pressure of the system is tuned with a barostat by volume rescaling [47, 48].

The solvent environment of a molecule can be simulated either with explicit solvent molecules or through what is called implicit solvent, where the solvent is instead represented by a continuous medium. It has been show to be difficult to create implicit solvent models that account for non-uniform solvent distribution in interface regions and cavities [49]. Although implicit solvent models are known to enhance the sampling rate of conformational states, their ability to produce realistic ensembles over longer time-scales remains to be shown [50].

To limit the number of solvent molecules the simulation box is commonly made as small as possible. Periodic boundary conditions are used to alleviate the unphysical effects of a stiff wall or vacuum surrounding the simulation box. Instead, the simulation box can be imagined as the unit cell in an infinite periodic lattice. This enables an efficient handling of the long-range interactions that achieves a computational complexity of $O(N \log N)$, which consecutively yields a significantly boosted sampling speed for larger complexes [51–53]. This improved scalability, makes the higher amount of sampling required for a valid estimation of the free energy of macromolecular systems accessible for simulation. However, as simulations become longer the requirements on the accuracy of force fields become increasingly important [54].
3.2 Enhanced Sampling

Despite remarkable computational efficiency, MD simulations still fall short of reaching timescales to provide answers to many important biological questions. Although this is likely to gradually improve, given the development in computational power, it is important to consider the underlying cause for this limitation, which can primarily be attributed to the rough shape of the FES of biological systems. This is the collective effect of many local minima that trap the system and force the simulation to sample limited regions of conformational space \[56, 57\]. The most important functional aspects of proteins involve large conformational changes that require crossing many free energy barriers. Examples of such mechanisms include the opening and closing of channels and transporters or the large-amplitude movements in the catalytic activity of enzymes \[58, 59\].

The probability of surmounting a free energy barrier for a physical system is exponentially related to its size. This can restrict access to certain relevant parts of the phase space, which will be sampled less efficiently, or not at all during the simulation. To improve the sampling rate of MD, enhanced sampling methods alter this behavior and force the simulation to surmount barriers and explore larger regions. Although this will give an unphysical sampling distribution, the idea is to do this in such a way that the underlying physical distribution can be recovered knowing the applied modifications, but through much less sampling.

3.2.1 Replica exchange

Replica exchange is perhaps one of the best known such methods that was proposed by Swedensen for spin systems \[60\] and later reformulated for MD by Sugita and Okamoto \[61\]. The strength of this method is that it is based on multiple parallel simulations that can be executed largely separated, which is well suited for large computational clusters used for simulations. Each simulation is tuned to a different temperature, successively ranging from physiological temperatures and up. Occasionally, simulations with neighboring temperatures exchange their ensemble coordinates based on a Metropolis acceptance criterion \[61\]. The simulations running at physiological temperatures will explore the relevant parts of phase space while the ones at higher temperatures help expand the search region \[62\]. One of the major drawbacks of this method is that the number of ensembles required to reach an optimal acceptance ratio of \(20\%\) during the exchange step becomes increasingly larger with increased system size \[63\]. Given the availability of computational resources, this renders an upper boundary for the system size, at which point the computations become inefficient, despite the increased parallelism \[64\].

3.2.2 Steered MD

Another method is to establish a conformational pathway, along which different properties of the molecular system are of relevance. This could for instance be the pore radius of a channel or the distance between a ligand and its binding pocket. This type of reduced, more compact, description of a system property can
be defined as a function of the system coordinates,
\[ \xi = f(r_1, \ldots, r_N) . \] (3.3)

This defines a so-called collective variable (CV) or a reaction coordinate.

Umbrella sampling is an example of a steered MD method that can perform a more directed exploration of the conformational space, given a CV. Through multiple simulations with biased umbrella potentials, exploration of local intermediate states is performed and the resulting occupancy is used to calculate the FES along the pathway. This is achieved by providing a series of conformations in intervals along the CV and use them to define a harmonic (or umbrella) potential that restricts each simulation to sampling the immediate vicinity of these conformations. The bias force pulling the system away is calculated and used to derive the slope of the underlying free energy landscape. The measure of deviation can for instance be based on root mean square deviation (RMSD) [65].

Since conformations along the CV have to be provided, this method cannot be used to explore new conformational states.

### 3.2.3 Metadynamics

Metadynamics (MetaD) is yet another powerful method to enhance sampling that was suggested by Laio and Parrinello in 2002 [66]. It uses a history-dependent biasing potential to prevent the system from getting trapped in previously visited regions in the CV-space.

For a set of \( M \) CVs, \( \xi_i \) \((i = 1, 2, \ldots, M)\), which span a phase space with the FES \( F(s) \), the exploration of the free energy landscape will be driven by a thermodynamic force given as
\[ F_i = -\frac{\partial F}{\partial \xi_i} . \] (3.4)

In the case where the free energy is a potential defined within the CV-space, it can be shown [67] that the first order expansion of the dynamics is given by the discretized evolution equation
\[ \xi_{i}^{t+1} = \xi_{i}^{t} + \delta \xi \frac{F_i^t}{|F_i^t|} , \] (3.5)

where \( \xi^t \) denotes the CV at the simulation step \( t \), \( \delta \xi \) is a step size and \( F_i^t / |F_i^t| \) is the normalized thermodynamic force at step \( t \). This type of sampling would head towards the closest local minimum of the system and get stuck, which shows the previously mentioned limitations of regular MD. MetaD was proposed to address this problem by depositing repulsive Gaussian potentials at points previously visited by the system and hence creating a bias force added to the thermodynamic force, which yields the discretized total force
\[ \tilde{F}_i^t = -\frac{\partial F}{\partial \xi_i} - \omega \frac{\partial}{\partial \xi_i} \prod_{i < t} \exp \left( -\frac{[\xi_i(t) - \xi_i^t]^2}{2\sigma_i^2} \right) , \] (3.6)
where $\omega$ dictates the height of the Gaussian (i.e. the strength of the applied bias force) while $\sigma_i$ controls the width of the potential. The sum is over all the deposited Gaussian potentials.

This modified form of dynamics will yield a system that can be viewed as a random walk process that diffuses through the conformational space. As it gets stuck in any local region in the subspace defined by $\xi$ it eventually builds a large enough repulsive potential to push it away. Eventually all the local minima in the FES will be filled with repulsive potentials and the system will move efficiently over the entire CV-space. At this point the collective set of all the Gaussian potentials will be an inverse approximation of the underlying free energy landscape, up to a constant, which is calculated as

$$V(\xi) \approx -\lim_{t \to \infty} V_G(\xi, t) = -\omega \sum_{t' \leq t} \exp \left( -\sum_{i} \frac{[\xi_i(t) - \xi_i(t')]^2}{2\sigma_i^2} \right). \quad (3.7)$$

Here $V_G$ defines the sum of the Gaussian potentials deposited at an infinite number of simulation steps.

The bias potential, $V_G$, will in turn generate a bias force on each atom in the system. As shown in eq. (3.2), to calculate this force for atom $i$, the gradient with respect to the atomic coordinates $r_i$ need to be calculated as

$$F_i(r, t) = -\nabla_r V_G(r, t) = -\omega \sum_{t' \leq t} \nabla_r \exp \left( -\sum_{i} \frac{[\xi_i(r, t) - \xi_i(t')]^2}{2\sigma_i^2} \right) =$$

$$= \omega \sum_{t' \leq t} \sum_{i} \frac{\xi_i(r, t) - \xi_i(t')}{\sigma_i^2} \exp \left( -\sum_{i} \frac{[\xi_i(r, t) - \xi_i(t')]^2}{2\sigma_i^2} \right) \nabla_r \xi_i(r, t). \quad (3.8)$$

This shows that the parameter $\omega$ will have a direct effect on the size of the (unphysical) forces generated by the bias potential and it can thus be used to stabilize the simulation. But more importantly, this shows that MetaD requires an explicit expression for each CV, $\xi_i$, as well as an explicit expression for its gradient with respect to the atomic coordinates, which limits the method to CVs that are differentiable.

In paper II we show that the Debye Formula, see eq. (2.1), is differentiable and can be employed to guide MD simulations with experimental SAXS data through the MetaD protocol. This approach enhances the sampling rate of the simulations along a CV defined by the SAXS data, which provides a measure to combine the information available in the experimental data and the force fields to sample thermodynamically valid conformations close to the observed state. This can be used to disentangled the mixture of signals from different domains in the SAXS data, as well as correct for the noise present in the data.
If we are uncritical we shall always find what we want: we shall look for, and find, confirmations, and we shall look away from, and not see, whatever might be dangerous to our pet theories.

— Karl Popper

4

Computational Single-Particle Cryo-EM

Contents

3.1 Molecular Dynamics .......................... 15
3.2 Enhanced Sampling .......................... 18
  3.2.1 Replica exchange ......................... 18
  3.2.2 Steered MD ............................ 18
  3.2.3 Metadynamics ............................ 19

The advent of direct electron detectors with high output frame rates that make it possible to produce large amounts of data through automated modes of operation has been central to the success of single-particle cryo-EM in recent years [68]. Powerful classification algorithms have made it possible to use more heterogeneous datasets, which for instance can be highly polydisperse samples, to achieve equally good reconstructions. In effect, the sample is purified computationally rather than by arduous and often damaging biochemical procedures [69]. This high-throughput approach generates a need for data processing pipelines with equal or preferably higher throughput capabilities.

This chapter will provide an overview of the most central parts of this computational pipeline starting from raw microscope data all the way to the final reconstructed 3-D high resolution electron density map. Furthermore it will outline the work done to achieve orders of magnitude faster computations in the most significant bottlenecks along the pipeline. This work was implemented in the popular software package RELION, which is predominantly targeted at single-particle cryo-EM data refinement [70].
4.1 Beam-Induced Motion & Radiation Damage

As the sample is irradiated by the negatively charged electron beam, some of the electrons that pass through it are absorbed. The deposition of charge leads to significant deformations of the ice, which is known as beam-induced motion [71, 72]. This effect can be divided into two components: uniform whole-frame motion and nonuniform local motion that varies across the frame [73]. Unless corrected for, this phenomenon leads to a blurring effect that will significantly impact the performance of downstream calculations.

The energy deposited into the sample by inelastically scattered electrons also causes structural damage that increases with longer exposure times. In effect, the high-resolution data is distorted as more data is collected from the same region of the sample, which implies that the initial frames will contain more structural information than the subsequent ones [74, 75].

Both the above-mentioned distortions can be accounted for during the step called motion correction, which is typically the initial step of the data refinement process. The correlation between neighboring subframes is used to derive a least-squares estimate of relative motion [73]. To account for the nonuniform motion, the frame is divided into multiple segments and a similar analysis is performed on each patch. A deformation function that varies smoothly across the entire image area is calculated and used to remap pixels from every subframe into a single-frame averaged micrograph. Using known data about the effect of radiation on biological samples, the amount of radiation damage per frame can be estimated based on the electron dose and the frames weighted accordingly.

4.2 CTF-Estimation, Particle Picking & Extraction

The 2-D phase contrast transfer function of each micrograph describes the information distribution in the Fourier domain and is elemental in the reconstruction phase of the refinement. Fortunately, both the signal and noise components of the image data are modulated by the CTF-function (see section 2.4.1), which can be retrieved by fitting it to the Fourier amplitudes (power spectra) of the image (see fig. 2.2). The defocus parameter and the angle defining the astigmatism are estimated by fitting the CTF-function to the power-spectra of the image [76, 77]. Most implementations assume small levels of astigmatism and initially estimate a single defocus value in a 1-D rotational average of the image. This estimation is then expanded into a 2-D and the full parameter set is fine-tuned. These three parameters are stored in a database and fed downstream alongside the image data.

After motion correction and CTF-estimation the single-frame micrographs are usually subject to particle picking. Here, subregions that look like projections of the target complexes are identified and extracted from the micrographs as smaller images for further processing. This step can be extremely labor-intensive if done manually, since millions of asymmetrical units can be required for the subsequent steps in the...
pipeline [71, 78]. It is highly desirable to at least partially automate this task. Semi-supervised machine learning approaches have been very successful for this [79, 80].

The identified particles are extracted from the micrographs and stored for subsequent processing. At this stage the data size has decreased dramatically compared to the movie representation of the data. It is however not common practice to discard the raw data, since some calculations further downstream require access to it and future algorithmic advances might cause us to re-process the data.

4.3 2-D/3-D Classification & Reconstruction

One of the key strengths of cryo-EM is the capability to handle data with high levels of heterogeneity. This is a direct consequence of the heavy emphasis on the computational side of the method. Shifting the purification requirements from the experimental side to the computational side yields a larger number of unknown parameters, which increases the algorithmic complexity and the risk of over-fitting.

Each particle image is a transmission projection of a 3-D object down to 2-D, which results in major information loss. However, since a dataset contains a large number of projections of identical particles in different orientations, the 3-D structural information can still be reconstructed. A particle image in a single-particle cryo-EM dataset is generally associated with five spatial parameters that are unknown. To uniquely describe arbitrary orientations of an object in three spatial dimensions, a minimum of three rotational parameters are required. Two additional translational parameters are required to describe the center of the particle in the projection plane. These parameters will have to be determined for each particle to correctly reconstruct the underlying 3-D object.

It can be shown that two projection images from arbitrary orientations of the same 3-D object will share a line in Fourier space, through the common line principle [81]. A family of methods employ this to determine the five unknown parameters by minimizing the discrepancy of pair-wise common lines of all the images in the dataset. The angular reconstruction method was proposed in 1987 as an efficient approach for doing this [82]. However, this family of methods is subject to three major shortcomings:

1. The information content used in the pairwise comparison is reduced to just a line from each image. This results in lower robustness against noise, which becomes a considerable issue given the extremely low SNR in cryo-EM data.

2. It is very difficult to implement the pairwise comparison between the images in a computationally efficient manner. The entire dataset must be accessed simultaneously, which creates an upper bound on dataset size.

3. There is no true common line between two non-identical particle projections, which will limit the method to strictly heterogeneous dataset.

Despite numerous efforts [83–86], to date this approach has not been successful for 3-D reconstruction of single particle cryo-EM data.
4.3.1 Maximum Likelihood

A more recent approach uses a robust maximum-likelihood estimator to optimize the parameters. Initially suggested for 2-D images in 1998, this method calculates an average representation of the data and iteratively seeks to improve it. Based on a probabilistic model, the objective function is defined as the probability \( f = P(\Theta | \Psi) \) for the model \( \Theta \), defined as the collective set of estimations of the latent variables (i.e. unknown parameters), given the entire dataset of \( N \) images, \( \Psi = \{X_1, \ldots, X_N\} \). Maximizing the objective function maximizes the probability of the model given the data. This approach removes the pair-wise comparisons between images in favor of comparisons between the images and the model. In each iteration, the parameters are tuned in such a way that the model improves and its discrepancy to the dataset decreases. This algorithm will eventually converge to a model given the dataset in a least-square sense [87]. The final model is then used to generate a density map through the technique of back-projection.

Most importantly, this method can incorporate the entire image in its comparisons rather than just a line, which significantly improves its robustness to noise. Additionally the comparison is done to a common model rather than to every other image in the dataset, which from a data access point of view is much more efficient.

A very successful approach for maximizing the likelihood function, \( P(\theta | X) \), employs Bayes' theorem,
\[
P(\Theta | \Psi) \propto P(\Psi | \Theta) P(\Theta),
\]
(4.1)
to convert the objective function to instead be expressed as the probability of observing the data, \( \Psi \), given the model, \( \Theta \), i.e. the posterior probability. Here the model \( \hat{\Theta} \) that maximizes the posterior is called the maximum a posteriori estimate. The probability of observing image \( i \) in a given arbitrary configuration of the unknown rotational and translational parameters, \( \phi \), is defined as
\[
P(X_i | \phi, \Theta) = \prod_j \frac{1}{2\pi \sigma_{ij}^2} \exp\left(-\frac{|X_{ij} - C_{ij} P^\phi M(\Theta)|^2}{2\sigma_{ij}^2}\right).
\]
(4.2)
Here \( j \) iterates over the Fourier components, \( \sigma_{ij}^2 \) is the estimated noise variance, \( C_{ij} \) is the CTF and \( P^\phi \) are the components of the projection operator acting on the map \( M(\Theta) \), which is defined by the current model \( \Theta \). The marginal probability can be obtained from a sum over all \( \phi \). Further assuming that the measurements are independent, we can write the marginal posterior probability for the entire dataset, \( \Psi \), as
\[
P(\Psi | \Theta) = \prod_i \int_\phi P(X_i | \phi, \Theta) d\phi
\]
(4.3)
Assuming that all Fourier components are independent and normal distributed (i.e. smoothness) one can define the prior probability as
\[
P(\Theta) = \prod_j \frac{1}{2\pi \tau_j^2} \exp\left(-\frac{|M_j(\Theta)|^2}{2\tau_j^2}\right),
\]
(4.4)
where $\tau^2_j$ is the unknown variance of $M_j$.

Due to the monotonic property of the log-function, maximizing $\log x$ is equivalent to maximizing $x$. It turns out that it is computationally more efficient to maximize the log-likelihood and hence we define the objective function as

$$f(\Psi, \Theta) = \log P(\Psi|\Theta) + \log P(\Theta)$$  

Maximizing this function will maximize the probability of the model, $\Theta$, given the entire dataset, $\Psi$, in a maximum a posteriori sense. This can be done in multiple ways. One technique for doing this is the expectation-maximization algorithm, where the objective function is maximized via an update formula. In the case of eq. (4.5) this formula is

$$M(\Theta) = \frac{\sum_i \int_0 \Gamma \sum_j B_j^\theta \frac{C_j X_j}{\sigma_j^2} + \frac{1}{\tau^2_j}}{\sum_i \int_0 \Gamma \sum_j B_j^\theta \frac{C_j X_j}{\sigma_j^2}}$$

where $B_j^\theta$ is the back-projection operator, i.e. the inverse of the projector operator and $\Gamma$ is a function of both $\Psi$ and $\Theta$ (see appendix A.2). This algorithm is implemented in RELION for 2-D classification, 3-D classification and refinement [88].

4.3.2 Classification

A maximum likelihood approach is flexible in the sense that additional parameters can be added to the objective function, which enables one to perform the optimization across multiple models rather than a single one. This is done by introducing an additional optimization parameter per particle in addition to the five spatial ones, namely a weight distribution across the multiple models. This way, the optimization can simultaneously perform a classification of the dataset, which is key to handling heterogeneous datasets. The classification protocol can be viewed

Figure 4.1: Illustration of the major steps in the refinement pipeline of single-particle cryo-EM, from particle picking and data extraction to the iterative refinement process. The refinement iterates through the data and maximizes the probability of the model given the image data through the expectation-maximization approach. FT is short for Fourier transform and purple marks objects in Fourier space. (Illustration by the author, adapted from paper I)
as an implicit k-means optimization where the objective is to move the centroids, i.e. models, to the center of each cluster of particles belonging to the different types of biological complexes [86, 89].

A 2-D classification is usually performed before a 3-D classification due to considerations of computational time and resource availability. Here only a single rotational parameter is considered, which reduces the total number of parameters per particle to four, three spatial and one for class weights. This reduced complexity and the fact that the algorithm is dealing with images rather than 3-D objects significantly improves calculation speed (no projections are for instance required). 2-D classification is only used as a coarse filtration step that discards all the junk images from the true particle images, because the classification will be performed on the orientation of the particles rather than particle type (i.e. 3-D structure). This is to enrich the ratio of biologically relevant particles in the dataset in the computationally more expensive 3-D classification.

4.3.3 Refinement

After 2-D and 3-D classification the dataset is reduced to one or multiple subsets corresponding to different particle types. Within each subset the data is considered homogeneous enough for a final reconstruction, where the aim is to push the resolution of the generated density map. It is at this point that the risk of over-fitting becomes substantial, mainly due to the rapid drop in SNR for high-resolution features in the map. To alleviate this risk the so-called Gold standard is used, where the dataset is split into two halves and reconstruction is performed on them separately. This way over-fitted features in the map can be measured through a drop in correlation between the two halves. This is possible due to the fact that signal should similar (i.e. correlated) between the halves, whereas noise is not. in other words, if a molecular feature truly exists it should look roughly the same within the two halves, otherwise it cannot be distinguished from noise and should hence not be reinforced in refinement.

4.3.4 FSC – Fourier Shell Correlation

The above mentioned correlation measurement is based on the Fourier shell correlation (FSC) that for a shell with radius $q$ from origin is given as

$$FSC(q) = \frac{\sum_{q_i \in q} F_1(q_i) \cdot \bar{F}_2(q_i)}{\sqrt{\sum_{q_i \in q} |F_1(q_i)|^2 \cdot \sum_{q_i \in q} |F_2(q_i)|^2}}, \quad (4.7)$$

where $F_k(q_i)$ denote the Fourier component, $q_i$, of the half-map $k$. $\bar{F}$ denotes the complex-conjugate of $F$. The sums are carried out over the components, $q_i$, belonging to the shell, $q$. It can be shown that the FSC is the ratio of signal in the total power of the data, $FSC(q) = S(q)/(S(q) + N(q))$, where $S(q)$ and $N(q)$ are the power of signal and noise, respectively [90]. This can in turn be used to estimate the global resolution of a map. In modern crystallography the resolution is estimated based on the correlation between the calculated electron density map and
Figure 4.2: Illustration of the hierarchy of parallelism, with decreasing levels of task complexity from left to right. Each colored tile represents an example task that is further expanded on to the left. Starting from individual images, classes, orientations (and translations) and eventually on to individual image pixels (or Fourier coefficients), the work is split into very small tasks that can be executed in parallel. (Illustration by author, adapted from paper 1)

the atomic model generated from the same map. Since an atomic model is usually a requirement for refining crystallography data (see section 2.1), there is always one available for this correlation measurement, which is not the case for cryo-EM. To make resolution estimates comparable between the two methods the 0.143 FSC cutoff criterion was derived, which is based on a statistical reasoning that takes into account the fact that the correlation measurement is between two noisy maps rather than between the map and a noise-free atomic model. [91, 92]. The global resolution is estimated by the resolution of the Fourier shell before the FSC falls below 0.143.

4.4 Hardware acceleration

The maximum likelihood algorithm in RELION is by far the most computationally expensive part of the entire pipeline. It is instrumental throughout 2-D classification, 3-D classification and refinement. These calculations used to require weeks on computational clusters, which imposed considerable limitations on the throughput of cryo-EM as a method. Additionally, the user only had a limited number of attempts at classification and refinement. Given the flexible nature of the computational part of cryo-EM, the user should ideally instead investigate the dataset in a more exploratory fashion. It can for instance be mined for rare conformational states or further purified by altering the classification parameters, which requires multiple passes of classification and refinement. This puts the efficiency of the implementation of the algorithm at the heart of the practical capabilities of cryo-EM.

4.4.1 Parallelism

A closer look at eqs. (4.2) and (4.3) suggests that for the evaluation of the marginalized posterior probability, the projection operator $P^\phi$ has to be evaluated
many times for a large number of parameter configurations per particle. This is one of the most important limiting steps in the calculations. Another limiting step is the back-projection operator, $B^\phi$, which is also required to be done many times for each particle to generate the map $M(\Theta)$. The multitude of projections and back-projections are computational tasks that have to be carried out at some point in the calculations. However, as long as they are done at a certain point in the calculations it does not matter in what order they are done.

It turns out that this property is shared by many computational tasks at different levels in the implementation of a maximum likelihood algorithm. These are tasks that can be carried out in isolation from each other. This property is known as parallelism and is inherent to an algorithm. Each parallel task can be carried out in isolation from the other ones, which enables efficient distributed calculation on multiple computational units. Each factor in the product presented in eqs. (4.2), (4.3) and (4.4) can be a separate computational task.

However, it is not trivial to implement an algorithm in a way that exposes this inherent parallelism. The size and complexity of a tasks dictates what computational resources are best suited or most efficient for carrying it out. For instance at the highest level of complexity one can consider the computational task for each particle, $X_i$, in eq. (4.2) and at the lowest level the individual Fourier coefficients of the projection operator, $P^\phi_j M(\Theta)$. The central processing unit (CPU) of a computer is a general-purpose computational unit that can handle a large variety of computational task and is very well suited for large complex task. There are however other computational units, like the graphics processing unit (GPU), that are instead purpose-built for handling a large number of smaller tasks. GPUs excel at very small tasks that are repeated many times, like the processing of individual pixels in computer graphics. The projection of each Fourier component is very similar to tasks commonly performed in image processing, why GPUs are well suited for it. On the contrary CPUs are more suited for managing the larger tasks that are performed few times, like handling data access to and from disk. Acceleration of a computational task refers to the use of dedicated hardware for faster and/or more efficient execution. Often this includes a complete reformulation of the algorithm to better handle the data flow between the different computational units and the use of heuristic technique to improve the overall efficiency with negligible loss in precision.

In paper I and V, work related to the acceleration of the maximum likelihood method is presented.
Evolution is a process that depends on amplifying things that almost never happen.

— Daniel Dennett

5

Biological Applications

Contents

4.1 Beam-Induced Motion & Radiation Damage . . . . . . 22
4.2 CTF-Estimation, Particle Picking & Extraction . . . . 22
4.3 2-D/3-D Classification & Reconstruction . . . . . . . 23
  4.3.1 Maximum Likelihood . . . . . . . . . . . . . . . . . . . . 24
  4.3.2 Classification . . . . . . . . . . . . . . . . . . . . . . . . 25
  4.3.3 Refinement . . . . . . . . . . . . . . . . . . . . . . . . . 26
  4.3.4 FSC – Fourier Shell Correlation . . . . . . . . . . . . . . 26
4.4 Hardware acceleration . . . . . . . . . . . . . . . . . . 27
  4.4.1 Parallelism . . . . . . . . . . . . . . . . . . . . . . . . . 27

One of the main strengths of Cryo-EM is enabling structure determination of mega-dalton complexes, such as ribosomes, splisosomes and viruses [75]. Although some of these complexes have been studied in their fully assembled form with X-ray crystallography [93–95], it is far less difficult to do so with cryo-EM. Thus with the advent of high-resolution cryo-EM the number of published very large multi-protein structures has rapidly increased along with our understanding of them.

Ribosomes are a family of complexes that are exceptionally well suited for cryo-EM [96]. They produce relatively high contrast images, due to their RNA content, which generates increased SNR and significantly aids particle alignment. Hence they are commonly used for validating and developing cryo-EM as a method. Although this thesis is heavily focused on method development, some attention will be directed towards the application of cryo-EM on ribosomal complexes and related biological discoveries.
5.1 The Ribosome

The production process of a protein starts as the blueprint for it, stored in the genome as DNA, is transcribed into RNA. This type of RNA is known as messenger RNA (mRNA) and serves to convey information from the genome to the ribosomal complexes that translate it into a peptide chain. The ribosome is a large multi-protein complex that contains a large amount of ribosomal RNA (rRNA). It is made of two subunits, historically named the large subunit (LSU) and the small subunit (SSU), which comprise a few RNA chains and an irregular number of proteins. The total RNA content in different species and organelles varies roughly from 40% to 60% [97].

A unit of measurement frequently used to describe the ribosome and its different structural elements is the Svedberg unit (S). This is the sedimentation coefficient observed during centrifugation, which vaguely correlates with a particle's size and weight [98]. Employing this unit, the prokaryotic 70S ribosome consists of a 50S and a 30S subunit, while the eukaryotic 80S cytosolic ribosome consists of a 60S and 40S subunit [97]. Similarly, ribosomes contained in some eukaryotic organelles are made up of 70S particles, which points to their prokaryotic evolutionary origin [99].

The hub-like role of the ribosomes within the biomolecular network of the cell introduces a high level of conservation in their structure and function, which makes them an important target for antibiotic treatment [100].

5.1.1 Protein Synthesis

Protein synthesis is initiated through the recruitment of a ribosome-initiator tRNA (transfer-RNA) complex to the initiation codon of an mRNA. The prokaryotic initiation complex forms through the direct interaction between rRNA and mRNA [101–103], whereas the eukaryotic counterpart relies primarily on protein-RNA and protein-protein interactions, which enables a more complex regulatory mechanism [104, 105].

The ribosome has three canonical binding sites for tRNA within a tunnel-like structure formed between the two subunits: A (aminoacyl-tRNA), P (peptidyl-tRNA), and E (exit) [106, 107]. The primary purpose of tRNA is to carry individual amino acids to the ribosomal binding site and match its template RNA to the base triplets in the mRNA (see fig. 5.1). Shortly after initiation, the first tRNA is located at the P site ready for the elongation cycle to commence. The first step in the cycle involves the binding of a tRNA to the A site followed by the formation of a peptide bond. Once the peptide bond is formed, the tRNA in the P site is deacylated and the nascent poly-peptide is now covalently bound solely to the tRNA in the A site. Through a series of conformational changes the deacylated leading tRNA is transfered to the E site where it is eventually expelled, whereas the trailing tRNA is translocated to the P site. The matching codon on the mRNA is also translocated along with the tRNA at this step. At this point the A site is vacant and the cycle is ready to repeat.

Throughout this process the ribosome traverses a series of large conformational changes, amongst which are the ratchet-like rolling correlated with tRNA and
mRNA translocation [108–113].

5.1.2 The Mitochondrial Ribosome

The mitochondria in eukaryotic cells have their own independent genome that shows substantial similarities to that found in prokaryotic cells. The mitochondrial proteome is partially synthesized by the mitoribosome (mitochondrial ribosome) and partially by the cytoplasmic ribosome. The human mitochondria contain a small circular DNA of about 16 kbp that encodes 37 genes. The ribosome plays an important role in regulating the distribution of the different proteins active mainly in the respiratory chain [114]. All of the 82 proteins in the mitoribosome are produced by the cytosolic ribosome, transported to and imported into the mitochondria. Despite this, the entire rRNA contents is produced within the mitochondria [115, 116].

The similarities to the prokaryotic ribosome pose a major hurdle in antibiotic treatment. Thus, the distinctions are of great importance in development of new antibiotics. Despite this, our understanding of the mitoribosome is far behind that of the cytosolic one.

In paper IV, work related to the assembly process of the mitoribosome prior to the formation of the initiation complex is presented. The human mitoribosome contains about half the length of rRNA, compared to the prokaryotic counterpart and it contains an additional 36 proteins. This increases the complexity of its assembly process and we show that this has yielded a different more complex assembly mechanism.

5.2 How to Capture the Conformational Sates of the Ribosome

Cryo-EM images capture all the states of the sample present after the freezing process. The canonical and the hybrid states of the ribosome with populated tRNA at the different sites is an example of a set of states that can be captured [106].

Some form of classification has to be applied to extract the data subsets belonging
to each state, as described in chapter 4. However, it turns out that a trivial global classification approach is not always sufficient to distinguish the desired states. The reason for this is that the overall variability in the signal tends to over-shadow the one coming from the structural features important for differentiating the states. It is therefore desirable to classify the variability of the signal coming from a specific region selected by the user. This can be achieved by suppressing all the signal not belonging to the region of interest by subtracting it from the images [117].

Let us define the region of interest as $\Omega$. Then the best maximum a posteriori estimate of the signal not belonging to this region is $M_{\Omega}(\hat{\Theta})$ (see section 4.3.1), which we subtract from the data before running the classification. This will leave the signal from $\Omega$ somewhat intact and the classification should yield a better clustering of this signal. The clustering in turn yields a number of subsets of the dataset that can be used to further refine $\Omega$. Assigning the tRNA-binding sites to $\Omega$ will for instance focus the classification to the multiple states differentiated by the occupancy of the different sites. Another example is the degree of folding of the rRNA interfacing the two subunits observed in the premature LSU before the formation of the initiation complex, as presented in paper IV.

This is a powerful method for differentiating discrete states and improving the resolution of biologically important regions. However, in case of the ribosome, the LSU and SSU exhibit a conformational change that is non-discrete. Instead, the rolling-like motion turns out to be a continuous degree of freedom. Let us define the regions belonging to LSU and SSU as $\Omega_{LSU}$ and $\Omega_{SSU}$. Due to the higher amount of signal belonging to $\Omega_{LSU}$, the refinement algorithm tends to focus on this region. It is thus desirable to subtract this region to improve the refinement of $\Omega_{SSU}$. A refinement performed immediately after subtraction (no classification is required) will maximize $\Theta$ only with respect to $\Omega_{SSU}$, which we write as $\hat{\Theta}_{\Omega_{SSU}}$. With this refinement the estimate $M(\hat{\Theta}_{\Omega_{SSU}})$ will provide a more accurate estimation for the signal belonging to $\Omega_{SSU}$. Subtracting this new improved estimate of $\Omega_{SSU}$ will yield a better subsequent refinement for $\Omega_{LSU}$.

This process can be repeated for an iterative improvement of the refinement result. Paper III presents a protocol for performing this iterative process for two or more domains.
Figure 5.2: Illustration of part of the processing pipeline of the mitochondrial ribosome large subunit presented in paper IV. Here different regions of a biologically relevant substructure (shown in green) are further improved using focused classification with two different masks A and B (left). The first classification serves only to identify particles containing the substructure, whereas the second classification aims to improve the consistency of the structural features of the substructure. After each classification the particles belonging to one or multiple selected classes are further refined. This approach yields an improved refinement of the specific substructure, but can result in deterioration of the signal from the other regions. (Illustration by author, adapted from paper IV)
Summary of Papers

I Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2

The computational cost for processing of single-particle cryo-EM data is rapidly increasing with the demand for the application of the method and the improved throughput of detector technology. The large computational cost is however becoming a considerable bottleneck, which is posing a significant hurdle for groups that only have access to limited computational resources. Additionally it hinders the development of more advanced algorithms that improve automation and refinement results but require more computations. This paper presents an accelerated implementation of algorithms used in the popular cryo-EM refinement software package RELION. The particle picking, classification and refinement steps were identified as the most computationally intensive steps and were accelerated using graphics processing units (GPUs). This work has yielded orders-of-magnitude faster computations for these steps and has significantly impacted the total time required for the entire refinement pipeline. Most importantly, this has made it possible to perform refinement on a single workstation, rather than a computational cluster.

II SAXS-guided Metadynamics

Small-angle X-ray scattering (SAXS) can be used for investigating structural features in biomolecules. However, the data produced is a rotational average of 3-D structures, which implies that the structural information is encoded into a mixture that is difficult to disentangle. An infinite number of conformations can produce a single SAXS profile, why it is desirable to apply some additional restraint to filter the improbable ones. In this paper a collective variable (CV) is presented that can
III. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION

The classification algorithms frequently used in single-particle cryo-EM are efficient at identifying discrete conformational states of macromolecules. They are however less ideal for structurally flexible complexes that show a continuous degree of freedom. This paper describes a new tool integrated into the refinement pipeline of RELION that addresses this limitation. It enables the user to define structural domains, or bodies, within the complex that are flexible in relation to each other. This information is then used to perform focused refinement of the individual bodies and account for their flexibility. Additionally a post-processing tool is provided with the purpose to discern the primary modes of motion through a dimensionality reduction approach, known as principal component analysis (PCA). The method is applied to two test cases, a cytoplasmic ribosome from Plasmodium falciparum, and the spliceosomal B-complex from yeast, which demonstrates how this method can be useful for gaining insight into dynamical properties of flexible biological macromolecules.

IV Structure of the human mitochondrial ribosome in native states of assembly

Large complexes like ribosomes are far less trivial to assemble than single protein molecules that usually fold on their own. For this reason there are cellular mechanisms in place to facilitate the assembly process of some of these complexes. Due to the reduced rRNA content of mitoribosomes relative to their evolutionary origin, the bacterial ribosome, their assembly process is complicated. This paper presents two assembly intermediates of the LSU, determined by cryo-EM to 3Å resolution, which belong to the final stages of the assembly process just before the LSU is merged with the SSU.

The late-stage protein incorporation and the folding process of the rRNA that interfaces the SSU is investigated. Furthermore the structure of the assembly factors that prevent the premature merging of the two subunits is presented. This includes the ribosomal silencing factor (RsfS), present in bacteria, and two additional factors, L0R8F8 and mt-ACP, previously not implicated in ribosomal assembly.
V  New tools for automated high-resolution cryo-EM structure determination in RELION-3

This paper presents a new version of RELION and all the additional new tools that comes with it:

1. A new implementation that can use CPU resources with capabilities similar to GPUs, namely efficient calculations of very small tasks. This is know as CPU vectorization and provides improved computational speed and efficiency when no GPUs are available or when hardware limitations prevents the use of GPUs, for instance due to insufficient GPU memory.

2. Particle picking can be performed without the requirement for the user to provide a reference. Instead Laplacian-of-Gaussian filtering is used to identify regions that show typical particle characteristics.

3. A new tool is presented for improving the CTF-estimation and correcting for the effects of a tilted beam on a per-particle basis. This is useful when the height of the particles varies across the micrograph due to uneven ice, in which case a global CTF estimation is inadequate.

4. Improved motion correction that performs a per-particle estimation of the motion rather than based on patches of the micrograph.

These new tools are applied to previously published data and shown to collectively yield an improvement of 0.2-0.7 Å in resolution compared to previous refinement results, which shows the impact of improved data processing.

VI  Cryo-EM reconstruction of the chlororibosome to 3.2 Å resolution within 24 h

The efficiency of the improved processing pipeline makes it feasible to start the refinement process already during the data collection. This type of workflow is called on-the-fly processing and is investigated in this paper. Here, an optimal organization of the workflow is presented that can be used to shorten the time between initiation of data collection and the first refined 3-D structure. This type of workflow is most valuable in cases when cryo-EM is used as a screening tool, in for instance drug binding experiments.

Additionally a proof of concept is presented where a high-resolution structure of the chloroplast ribosome is acquired within 24 hours starting from fresh spinach leaves. This shows the potential high-throughput capabilities of the method essential in a screening context.
Sammanfattning

Metoder för experimentell strukturbestämning utgör en fundamental del av den biologiska vetenskapen. Moderna metoder med hög genomströmning genererar stora mängder brusig svärtolkad data, vilket för manuell tolkning omöjligt. Istället har specialiserade beräkningsmetoder fått en allt större vikt för tolkning av experimentell data och i takt med den ökande genomströmningen av data ökar samtidigt kraven på effektivitet och robusthet på de fundamentala algoritmerna i dessa beräkningsmetoder. Denna avhandling fokuserar på de utmaningar som föreligger i utvinning av biologiskt relevant information med statistisk stöd i experimentell biologisk data.

Småvinkelröntgenspridning (SAXS) är en metod för granskning av biologiska molekyler i lösning, som har vissa experimentella fördelar jämfört med andra metoder för strukturbestämning. Den genererade datan är dock en väldigt brusig och reducerad representation av den tredimensionella strukturen, vilket medför att denna metod generellt är begränsad till analys av grova strukturella egenskaper, t.ex. molekylens tröghetsradie eller oligomeriseringstillstånd. I denna avhandling presenterar jag en beräkningsmetod där SAXS-data används för att styra molekylodynamik (MD) simuleringar. SAXS-datan används som referens för att styra sampling av konformationsrymden till experimentellt relevanta tillstånd, samtidigt som termodynamiska villkor är uppfyllda. Jag visar att denna typ av metod kan användas för att utforska proteintillstånd som är av större biologisk relevans.

Jag presenterar ett flertal förbättringar på de existerande algoritmerna samt helt nya algoritmer som tidigare inte var praktiskt användbara p.g.a. deras höga beräkningskostnad. Vidare presenterar jag arbetet med applikationen av dessa metoder på cryo-EM data på en konstruktions-intermediär av den mänskliga mitokondriella ribosomen. Denna analys resulterar i identifieringen av två tidigare okända faktorer som är involverade i konstruktionsprocessen samt nya rön om väckningsprocessen av den centrala rRNA:n.
Firstly, I would like to express my sincere gratitude to my advisor Erik Lindahl for creating a working environment that primarily nurtures creativity and independence. I am deeply grateful to you for instilling a radiant sense of self-confidence by never doubting my ability to accomplish something difficult. This gift will forever fuel my passion for science and shine the way in my hunt for knowledge.

My sincere thanks also go to Alexey Amunts for granting me the means to enter the realm of experiential sciences. I will always admire your passion for science and drive to accomplish great things.

I would like to thank Magnus Andersson for introducing me to scientific research in molecular biology, initially as my master thesis advisor and later as co-author on my first publication. I truly enjoyed working with you.

Björn, I would not have asked for anyone else to share this journey with than you. Your rare focus combined with a strict critical realist mindset make you the ideal co-worker in research and scientist in general. I cannot imagine anything other than a successful career ahead of you. In the meantime make sure to take every opportunity to make dad-jokes, since they define you as a caring father.

Marta, you are like a warm and welcoming ray of sunshine with a contagious laughter like no other. Your enthusiasm for science and technology has been a true inspiration for me.

Shintaro, the lab is not the same without your humor, good judgment and constantly snapping fingers. Alex (brorsan), your humor should be assigned world heritage status. As a professor you will be a sight to behold. Thanks Victor and Annemarie for the amazing company and intriguing discussions.

Saman, thank you for motivating me to more physical exercise and helping me get introduced to machine learning applications to gene expression. Iman you are an incredible inspiration and friend. Thank you Reba for supporting my laboratory work every bit of the way and Urska, with your dedication and hard work, success is never too far away. Christian B., I wish there was more time
for collaborative projects, since I genuinely enjoyed each and every encounter with you. And Stefan, thank you for fixing problems I never knew I had and being such an awesome friend and co-worker while doing it.

At this point I would like to thank all of my amazing co-workers at TCB-lab, Amunts’ lab, the cryo-EM facility and Delemotte’s army for creating such an incredible working environment. I will miss each and every one of you dearly. Additionally, I would like to extend my gratitude to my co-advisor Arne Elofsson, mentor Peter Brzezinski and coworkers at Department of Biochemistry and Biophysics (DBB) at Stockholm University for their support.

Without the support of friends and family this journey would not have been as pleasant and enjoyable. My sincere thanks goes to all of my extended family for their invaluable support.

I would like to thank Oscar, my closest friend, for his aid throughout my studies. Your creative and calculating mind make for an astonishing comedic engine I can never grow tired of. Adam, Jens and Sammy, thank you for all the deep discussions about life and technology that have been like food for the mind.

Bahar, my sweet little sister and friend, you have become a phenomenal human being with a sharp mind and a sparkle in your eyes that speak of a rare combination of wisdom and mischief. I am more than proud of what you have become and cannot wait to see what will become of you.

To my parents, I would like to express my deepest gratitude, for being there for me in every aspect of my life. Your sacrifices and struggles made me who I am and there is no way I can ever repay you. Thank you for blessing my life with love and wonder.

Isabel, there are no words to describe my gratitude to you. We have spent our entire adult life together and share countless wonderful memories. I cannot imagine life without you.
A.1 Diffraction Intensity of Dilute Solutions of Identical Particles

Assuming an elastic scattering process, the magnitude of the wavelength $\lambda$, of an incident photon is transferred to the scattered one. Thus the wave vector of the incident photon, $k_i$, and the scattered photon, $k_s$, can be defined as

$$k_i = \frac{2\pi}{\lambda} u_i,$$

$$k_s = \frac{2\pi}{\lambda} u_s,$$

where $u_i$ and $u_s$ are unit vectors pointing in the direction of the incident and scattered beam, respectively. The scattering vector can thus be defined as

$$Q = k_s - k_i .$$

With a scattering angle as defined in fig. 2.1, its amplitude becomes

$$Q = \frac{4\pi}{\lambda} \sin \theta .$$  (A.1)

Furthermore it can be shown [10] that the scattering amplitude of a single atom located at $r$ is

$$A = f \exp (iQ \cdot r) ,$$  (A.2)

where $f$ is the atom form factor. Consequently for a set of atoms this becomes

$$F(Q) = \sum_j f_j \exp (iQ \cdot r_j) .$$  (A.3)
The quantity measured during an experiment is the intensity of this scattered wave, \( I(Q) \), which for a monodisperse solution of \( N \) molecules become

\[
I_N(Q) = N \langle |F(Q)|^2 \rangle .
\] (A.4)

The average is over all the possible rotations of the molecule. Evaluating it with the expression

\[
\langle \exp(-iQ \cdot r) \rangle = \frac{\sin Qr}{Qr}
\] (A.5)

we arrive at the famous **Debye Formula**

\[
\langle |F(Q)|^2 \rangle = \sum_i \sum_j f_i f_j \frac{\sin Qr_{ij}}{Qr_{ij}},
\] (A.6)

where \( r_{ij} \) is the distance between atom \( i \) and \( j \). This yields the diffraction intensity

\[
I_N(Q) = N \sum_i \sum_j f_i f_j \frac{\sin Qr_{ij}}{Qr_{ij}},
\] (A.7)

for a monodisperse solution of \( N \) molecules.

### A.2 The Expectation-Maximization Algorithm

Eq. (4.5) defined the objective function that if maximized also maximizes the likelihood defined in eq.(4.1). There are multiple methods for doing this [79, 118], here we will explore a *fixed-point* for doing this, namely the **expectation-maximization** algorithm, since this is, as of this writing, the most popular approach. Taking the gradient of the objective function in eq.(4.5) with respect to \( M(\Theta) \) yields

\[
\nabla_M f = \nabla_M \log P(\Psi|\Theta) + \nabla_M \log P(\Theta)
= \frac{1}{P(\Psi|\Theta)} \int_\phi \nabla_M P(X_i|\phi, \Theta) d\phi + \nabla_M \log P(\Theta) .
\] (A.8)

Evaluating the gradient for the first term yields

\[
\nabla_M P(X_i|\phi, \Theta) = \nabla_M \prod_j P(X_{ij}|\phi, \Theta)
= P(X_i|\phi, \Theta) \sum_j \frac{\nabla_M P(X_{ij}|\phi, \Theta)}{P(X_{ij}|\phi, \Theta)}
= P(X_i|\phi, \Theta) \sum_j \frac{1}{\sigma_{ij}^2} (X_{ij} - C_{ij} P_j^\phi M(\Theta)) C_{ij} P_j^\phi
\] (A.9)

Similarly, the second term yields

\[
\nabla_M \log P(\Theta) = \frac{\nabla_M P(\Theta)}{P(\Theta)} = -\frac{M(\Theta)}{\tau_j^2} .
\] (A.10)
To simplify, we also define

$$\Gamma(X_i, \phi, \Theta) = \frac{P(X_i | \phi, \Theta)}{\int_{\phi'} P(X_i | \phi', \Theta) d\phi'}.$$  \hspace{1cm} (A.11)

Now, the gradient of the objective function in eq. (A.8) becomes

$$\nabla_M f = \sum_i \int_\phi \left( \sum_j \frac{1}{\sigma_{ij}^2} \left( X_{ij} - C_{ij} \phi^\phi M(\Theta) \right) C_{ij} \phi^\phi \right) \Gamma d\phi - \frac{M(\Theta)}{\tau_j^2}$$ \hspace{1cm} (A.12)

Since we are seeking the extremum of $f$, we put this expression equal to zero and solve for $M(\Theta)$, which yields

$$M(\Theta) = \frac{\sum_i \int_\phi \Gamma \sum_j B_j^\phi \frac{C_{ij} X_{ij}}{\sigma_{ij}^2}}{\sum_i \int_\phi \Gamma \sum_j B_j^\phi \frac{C_{ij} X_{ij}}{\sigma_{ij}^2} + \frac{1}{\tau_j^2}},$$ \hspace{1cm} (A.13)

where $B_j^\phi$ is the back-projection operator, i.e. the inverse of the projector operator. This expression defines the update-formula for the iterative maximization of the objective function that will yield an $M(\Theta)$ that maximizes the likelihood probability defined in eq. (4.1) in a maximum a posteriori sense.
References


References 53


