Towards unbiased 3D reconstruction in single-particle cryo-electron microscopy

Dominika Elmlund

Palo Alto 2010
to my family
ABSTRACT

Cryo-electron microscopy of freestanding molecules (single-particles) plays a pivotal role in the difficult and pressing challenge of determining the structures of large macromolecular complexes. Molecular volumes are generated by aligning large sets of randomly oriented two-dimensional (2D) projection images in three dimensions (3D) before reconstruction is performed using tomographic techniques. The increasing popularity of the single-particle method is highly correlated with technical advances in instrumentation and computation. This thesis introduces new computational methods for 3D structure determination from electron microscopic projection images of single molecules. The algorithms have been developed to fill a gap in the single particle methodology – the lack of methods for ab initio 3D reconstruction of asymmetrical or low-symmetry molecules co-existing in different functional states. The proposed approach does not rely on a priori information about the structure or the character of the sample heterogeneity, which minimizes template dependence and makes the methods applicable to a wide range of single molecules. The presented algorithms constitute the basis of a new open source software package - SIMPLE (Single-particle IMage Processing Linux Engine). SIMPLE is an efficient and easy-to-use image processing system for semi-automated ab initio 3D reconstruction from challenging single-particle data sets (asymmetrical particles, significant degree of heterogeneity).
LIST OF PUBLICATIONS


III. **Elmlund D.**, Elmlund H. “SIMPLE – an image processing system for ab initio 3D reconstruction in single-particle electron microscopy”, *manuscript*

SUPPORTING PUBLICATIONS


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two-Dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-Dimensional</td>
</tr>
<tr>
<td>AAA</td>
<td>ATPases Associated with various cellular Activities</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine Triphosphate</td>
</tr>
<tr>
<td>CLI</td>
<td>Command Line Interface</td>
</tr>
<tr>
<td>CTF</td>
<td>Contrast Transfer Function</td>
</tr>
<tr>
<td>DE</td>
<td>Differential Evolution</td>
</tr>
<tr>
<td>E. coli</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>EM</td>
<td>Electron Microscopy</td>
</tr>
<tr>
<td>EF-G</td>
<td>Elongation Factor G</td>
</tr>
<tr>
<td>FEG</td>
<td>Field Emission Gun</td>
</tr>
<tr>
<td>FSC</td>
<td>Fourier Shell Correlation</td>
</tr>
<tr>
<td>GRASP</td>
<td>Greedy Randomized Adaptive local Search Procedure</td>
</tr>
<tr>
<td>HAC</td>
<td>Hierarchical Ascendant Classification</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum-Likelihood</td>
</tr>
<tr>
<td>MSA</td>
<td>Multivariate Statistical Analysis</td>
</tr>
<tr>
<td>NMA</td>
<td>Network normal Mode Analysis</td>
</tr>
<tr>
<td>NTRC</td>
<td>NADPH-dependent Thioredoxin Reductase C</td>
</tr>
<tr>
<td>OTR</td>
<td>Orthogonal Tilt Reconstruction</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA Polymerase II</td>
</tr>
<tr>
<td>PPIX</td>
<td>Protoporphyrin IX</td>
</tr>
<tr>
<td>PR</td>
<td>Phase Residual</td>
</tr>
<tr>
<td>PSF</td>
<td>Point Spread Function</td>
</tr>
<tr>
<td>RAD</td>
<td>Reference-free Alignment in a Discrete angular space</td>
</tr>
<tr>
<td>RCT</td>
<td>Random Conical Tilt</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing</td>
</tr>
<tr>
<td>SIMPLE</td>
<td>Single-particle IMage Processing Linux Engine</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SSNR</td>
<td>Spectral SNR</td>
</tr>
<tr>
<td>TEM</td>
<td>Transmission Electron Microscope</td>
</tr>
<tr>
<td>t-RNA</td>
<td>Transfer RNA</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

This thesis introduces new computational methods for three-dimensional (3D) structure determination from electron microscopic projection images of single molecules. The presented algorithms constitute the basis of a new open source software package SIMPLE (Single-particle IMage Processing Linux Engine). The material is presented in four chapters and six papers (I-VI). Chapter 1 gives a brief introduction to single-particle electron microscopy. Chapter 2 and papers I-II describe the algorithms. Chapter 3 and paper III introduce the SIMPLE package, and papers IV-VI provide examples where the SIMPLE techniques have been applied to the study of structure and function of a number of biologically relevant macromolecules. Chapter 4 summarizes the results in papers I-VI.

SINGLE-PARTICLE ELECTRON MICROSCOPY

Visualization of large macromolecular assemblies by three-dimensional (3D) structure determination is essential for understanding biological processes (O'Donoghue et al., 2010). Cryo-electron microscopy (cryo-EM) in combination with methods for 3D reconstruction from large sets of randomly oriented 2D projection images of single molecules can be used for studying structure (Saibil, 2000; van Heel et al., 2000; Yu et al., 2008), composition (Elmlund et al., 2009), and dynamics (Fischer et al., 2010; Heymann et al., 2003) of macromolecular complexes. A large number of freestanding molecules, preserved in a thin layer of amorphous ice, are imaged in an electron microscope (van Heel et al., 2000). The resulting, often very noisy, 2D projections are usually aligned in 3D (paper II) (Farrow and Ottensmeyer, 1993; Grigorieff, 2007; Penczek et al., 1994) before reconstruction of a molecular volume using tomographic techniques (Harauz and van Heel, 1986; Radermacher, 1992). The method requires only a small amount of sample (typically 10µl of protein solution with nanomolar concentration), and enables imaging under physiological or biologically well characterized conditions (Dubochet et al., 1982). Continuous advances in instrumentation and image processing algorithms have allowed a number of structures to be determined to subnanometer resolution, for example the GroEL chaperonin required for protein folding (Ludtke et al., 2008; Ludtke et al., 2004), the Gro-EL-Gro-
ES complex embracing a newly folded protein in its chamber (Clare et al., 2009), the transferrin-transferrin receptor complex responsible for iron delivery to cells (Cheng et al., 2004), the general transcription factor IID involved in the regulation of transcription (Elmlund et al., 2009), and the AAA+ magnesium chelatase assembly catalyzing the first committed step of the chlorophyll biosynthesis (paper IV). Particularly impressive results have been reported for the ribosome, where the nascent RNA chain has been visualized in distinct conformations within the ribosomal exit unit (Bhushan et al., 2010; Seidelt et al., 2009).

Modern electron microscopes can deliver a resolution better than 1 Å (O'Keefe, 2008; O'Keefe et al., 2001); however, the short exposure time dictated by the sensitivity of the biological material to the electron radiation is the major obstacle towards routinely generating atomic resolution reconstructions (Henderson, 1995). Reconstructions that permit backbone tracing of the entire secondary structure have so far been obtained only for highly symmetrical assemblies (Böttcher et al., 1997; Cong et al., 2010; Jiang et al., 2008; Ludtke et al., 2008; Yu et al., 2008). For virus capsides and helical arrays, the 3D reconstruction process is aided by the symmetry-induced reduction of the orientation search space and the signal-to-noise (SNR) ratio enhancement, which greatly simplifies the alignment problem and reduces the amount of data required. Visualization of α-helices in a cryo-EM reconstruction was first achieved in the classical work on the core of the hepatitis B virus (Böttcher et al., 1997). In more recent studies, the complete atomic models of the epsilon 15 (Jiang et al., 2008) and the aquareovirus (Yu et al., 2008) icosahedral capsides have been obtained from cryo-EM density maps, marking another milestone for the method.

The number of 3D reconstructions determined to subnanometer resolution is steadily increasing, but high-resolution structure determination of particles with low or no internal symmetry still represents a considerable challenge. One problem is the generation of a reliable first 3D reconstruction (Elmlund et al., 2008; Ogura and Sato, 2006; Penczek et al., 1996; van Heel, 1987a). Another limiting factor is the intrinsic low SNR of the EM images, making high-resolution image orientation refinement a challenging problem (paper II) (Stewart and Grigorieff, 2004). It has been anticipated that millions of particles would be needed to obtain atomic resolution reconstructions of asymmetrical molecules (Zhou, 2008). However, the major obstacle towards fulfilling the promise of high-resolution visualization in the general case is
Towards unbiased 3D reconstruction in single-particle EM

conformational or compositional variability of the molecules in the specimen. Macromolecular assemblies are dynamic machines, which undergo conformational rearrangements to exert their function (Alberts, 1998). Identifying and characterizing heterogeneity in the single-particle population can be a rich source of biological information, but it still constitutes a major methodological challenge in the field, subject to vigorous research (paper I) (Fischer et al., 2010; Hall et al., 2007; Leschziner and Nogales, 2007; Penczek et al., 2006b; Scheres et al., 2007a; Shatsky et al., 2009; Zhang et al., 2008a). Another important issue is the validation of single-particle 3D density maps. Many 3D reconstruction methods routinely used in the field rely on alignment to pre-existing structures or mixed reconstructions generated from heterogeneous data. However, the possibility that the initial templates introduce model bias is not widely acknowledged.

This thesis addresses the problem of reconstructing molecular volumes from large single-particle image populations of asymmetrical or low-symmetry molecules co-existing in different functional states (papers I-III). The proposed approach does not rely on a priori information about the structure or the sample heterogeneity, which minimizes template dependence, and makes the methods applicable to a wide range of single molecules. Examples are provided of 3D reconstructions of a number of biologically relevant macromolecular assemblies (supporting papers IV-VI). In addition, the characterization of the effects of model bias in single-particle reconstructions is presented in Chapter 3.

Specimen preparation

Biological specimens consist of up to 80% water, which requires sample preparation that prevents structural collapse upon dehydration in the ultrahigh vacuum of the electron microscope column. Two techniques for specimen preservation are widely used: negative staining and cryo vitrification.

In the negative staining technique, first introduced by Brenner and Horne (Brenner and Horne, 1959), a layer of a heavy metal salt (e.g., uranyl acetate) is applied onto a carbon coated cooper grid containing a thin film of specimen solution. The excess liquid is blotted away and the suspension is allowed to dry in room temperature. The preparation gives strong contrast, but the structural information that can be gained is limited by the grain size of the stain to 15-20 Å. Drawbacks of the method include
significant distortions of the specimen due to staining and air drying, and skewed orientation distributions due to the often protein specific interactions between the molecules and the carbon support (Frank, 2006).

High-resolution macromolecular imaging requires embedding of the single-particles in vitreous ice. A perforated carbon coated cooper grid, covered with a thin film of specimen solution, is rapidly plunged into liquid ethane cooled by liquid nitrogen. The high heat capacity of the liquid ethane and the rapid plunging prevents formation of crystalline ice, resulting in a matrix of vitreous ice surrounding the individual particles, keeping them hydrated (Dubochet et al., 1981; Dubochet and McDowall, 1981; Knapek and Dubochet, 1980; McDowall et al., 1983).

Data collection and digitization

A typical electron microscope for cryo work is a 200kV instrument equipped with a field emission gun (FEG) electron source, and a liquid nitrogen cooled specimen holder. As compared to the heated tungsten filament traditionally used as an electron source, the FEG substantially improves the information transfer through the microscope, especially at high resolution (van Heel et al., 2000). While imaging dry samples is straightforward, collecting high quality cryo-EM data requires much more consideration. Vitrified specimens of unstained biological samples are very sensitive to the electron radiation, and in order not to damage the sample, a low electron dose must be used for exposure (6-10 electrons/Å²) (Unwin and Henderson, 1975). Low-dose imaging results in poor contrast due to the light atom composition of biological molecules, which makes them hard to distinguish from the vitreous ice background. To increase the image contrast, data are collected out of focus. The recorded image is distorted by noise from various sources (Baxter et al., 2009) and by the imperfections of the microscope’s lens system, described by the contrast transfer function (CTF) (Erickson and Klug, 1971; Frank and Penczek, 1995; Wade and Frank, 1977; van Heel et al., 2000). A mathematical description of the EM projection image will be given in Chapter 2. The data are recorded on photographic film or using a charge-coupled device (CCD) camera. The use of CCD substantially speeds up data collection and gives improved SNR ratio at low to medium resolution (Sander et al., 2005).
CTF correction

The CTF is a microscope-specific sinusoidal function describing the filtering characteristics of the imaging system in reciprocal space. The CTF depends on several parameters (Baker and Henderson, 2001):

\[
CTF(v) = -\left(1 - F_{amp}^2\right)^{\frac{1}{2}} \sin(\chi(v)) + F_{amp} \cos(\chi(v))
\]  

where \(\chi(v) = \pi \lambda v^2 (\Delta f - 0.5Cs \lambda^2 v^2)\), \(v\) is the spatial frequency (in Å\(^{-1}\)), \(F_{amp}\) is the fraction of amplitude contrast, \(\lambda\) is the electron wavelength (in Å), \(\Delta f\) is the defocus (in Å) and \(Cs\) is the spherical aberration coefficient of the objective lens of the microscope (in mm). The CTF oscillates between positive and negative contrast throughout reciprocal space, making different features of the object appear enhanced or suppressed in the image. The CTF oscillates at a higher rate at high spatial frequencies. No contrast is transferred at the zero crossings of the CTF, leading to complete loss of information. It is commonly suggested that this effect should be compensated for by acquiring images at different defocus settings. However, as the loss of information is minute compared to the information transferred, 3D reconstruction is possible from data acquired at a single defocus setting. An example is provided in paper II where simulated ribosome data was reconstructed using a defocus value of 2 microns.

Correction for the contrast inversions of the CTF can be done by applying a Wiener filter (Frank and Penczek, 1995). The electron microscopic Wiener filter can be described as a “careful division” by the CTF (Frank, 2006), attempting to minimize the risk of enhancing noise at frequencies that have a poor SNR. Deconvolution of the CTF is compromised by the presence of noise and the recovery is never ideal.

Initial model generation

The theory for 3D reconstruction from EM projections was introduced by DeRosier and Klug (DeRosier and Klug, 1968), and first used to reconstruct the tomato bushy stunt virus (Crowther et al., 1970a). The method is based on the projection slice theorem, which states that the Fourier transform of a 2D projection is a plane intersecting the origin of the 3D object’s Fourier transform (Bracewell, 1956). Any two non-parallel 2D projections of the same 3D object will therefore share a common line in Fourier space. The relative orientations between three non-parallel noise-free
projections can be exactly determined (van Heel, 1987a), whereas the reference-free common lines-based 3D alignment of any number of projections $n$ larger than three is NP-complete (Mielikainen et al., 2004), meaning that the computation required to find the global optimum grows faster than any power of $n$. The problem’s solution can, however, be approximated using relatively simple local search optimization techniques (Elmlund et al., 2008; Ogura and Sato, 2006; Penczek et al., 1996). Because of the low SNR of the individual single-particle images, common lines-based techniques generally operate on class averages, which requires prior in-plane alignment (Penczek et al., 1992) and unsupervised classification of the data set (Lebart et al., 1984; van Heel and Frank, 1981; van Heel and Harauz, 1986a) (see Chapter 2).

If the particles reside in a single or a few preferred orientations on the carbon support, which is usually the case only for negatively stained specimens, initial volumes can be generated by the Random Conical Tilt (RCT) method (Radermacher et al., 1987). In RCT, each image field is exposed twice at 60 and 0 degrees tilt, respectively. The untilted images are used for 2D alignment (Penczek et al., 1992), and the in-plane parameters are applied to the tilted images. 3D reconstruction is possible because the in-plane angles are random, giving the tilted projections random projection directions distributed in a cone-like geometry in Fourier space. Due to the difficulties of collecting high-tilted data in cryo and the need for a double exposure, the RCT method is often applied to negatively stain samples. The artifacts due to the negative stain preparation and the complete lack of information in a cone in Fourier space compromise the quality of the reconstruction. The artificial elongation of the reconstruction due to the missing cone can be overcome by using the Orthogonal Conical Reconstruction (OCR) method (Leschziner and Nogales, 2006).

Conformational state sorting

Heterogeneity in a single-particle population may arise from different sources. Macromolecules are often unstable under purification or sample preparation conditions, leading to random loss of components. Compositional heterogeneity can also arise from partial factor occupancy. In addition, there is often conformational heterogeneity, resulting from the intrinsic flexibility of large macromolecular complexes. Heterogeneity of the kind introduced by partial factor occupancy is intensely studied in the ribosome field, and separation of states is often achieved by template matching,
using one “empty” ribosome template, lacking the associated factor, and one template generated from the heterogeneous data (Gao et al., 2004; Valle et al., 2002). The general applicability of this method is limited since it requires knowledge about the sample heterogeneity. Variance analysis has been shown to be a powerful tool for characterization and localization of structural variability (Zhang et al., 2008a). Methods that do not depend on localization of specific structural regions include Maximum Likelihood (ML)-based approaches and methods based on common lines. *Ab initio* structure determination of co-existing conformational states via ML-classification has so far been achieved only for particles with icosahedral (60-fold) symmetry (Lee et al., 2007; Yin et al., 2003). However, by using available structural data to provide a first 3D alignment, ML-classification has been used to resolve different states of the asymmetric ribosome (Scheres et al., 2007a). *Ab initio* approaches for sorting heterogeneity in 2D, before 3D alignment, have also been proposed (Elad et al., 2008; Herman and Kalinowski, 2007). The Elad et al. method resolves compositional heterogeneity using statistical analysis of images in 2D, before reconstruction of homogenous groups by angular reconstitution (van Heel, 1987a). The method is limited to cases were structural variations are localized, and can be detected by searching for peaks in the eigenimages. In the approach proposed by Herman and Kalinowski, homogenous groups are identified by a tabu search-based unsupervised classification. The method assumes that homogenous subsets can be aligned in 3D using existing techniques. The performance of this method has never been shown on experimental data. A method for simultaneous common lines-based *ab initio* 3D alignment and conformational state sorting is proposed in paper I and summarized in Chapter 2 of this thesis. This method resolves conformational or compositional heterogeneity of low-symmetry molecules, without requiring knowledge about the structure or the sample heterogeneity.

**Refinement**

Iterative orientation refinement procedures operating on the individual images are required for improving the resolution of the 3D reconstruction(s). The orientations of the particle images are determined by alignment onto a reference volume, and a 3D reconstruction with improved resolution is calculated. The process is iterated until the resolution of the 3D reconstruction is no longer improving. Assignment of orientation
parameters can be done by cross-correlating the individual projections with re-
projections of the reference volume and the particle image is assigned the orientation of
the closest matching reference (Farrow and Ottensmeyer, 1992; Goncharov and
Gelfand, 1988; Harauz and Ottensmeyer, 1984; Penczek et al., 1994). Alternatively, a
common lines-based method can be formulated in Fourier space and used to align a
target transform against a set of reference central sections (paper II) (Lindahl, 2001). In
another Fourier-based method (Frealign) for orientation refinement, the individual
projections are compared directly with the 3D volume (Grigorieff, 2007).
The refinement process is prone to significant errors due to over-fitting of the often
very noisy image data to the 3D reconstruction (Stewart and Grigorieff, 2004).
Alignment strategies based on dynamic per-particle information range adjustment are
more robust towards noise than classical approaches, where a uniform resolution range
is used for all particles, and updated as the refinement progresses. In Frealign, a weight
is assigned to each resolution shell according to the spectral SNR (SSNR). Wavelet
decomposition has been used in multi-reference projection matching to assign weights
to different spectral bands (Saad and Chiu, 2000). Sorzano et al. presented a projection
matching algorithm based on wavelet filtering, where sub-bands of the frequency space
were included in a coarse-to-fine manner, and used to select the reference projections
included in the calculations, so that the number of reference projections used decreased
with increased resolution (Sorzano et al., 2004a). Scheres et al. estimated the separate
SSNR for each defocus group in a manner introduced by (Unser et al., 2005) to obtain a
multi-resolution likelihood-based classification algorithm that changed its filtering
resolution in an automated way, depending on the quality of the current map (Scheres
et al., 2007b). This thesis introduces a common lines-based image orientation
refinement method with spectrally self-adaptive low-pass frequency limit. The spectral
range is estimated for each target transform individually, to account for the possibly
distinct imaging conditions of each individual image, and the potential anisotropic
resolution of the starting reconstruction.

Validation and resolution
Currently, the most convincing method for validating a 3D reconstruction is to
compare the density with atomic structures of the entire complex or its submodules. If
the resolution does not allow for an unambiguous fitting of existing structures or no a
priori structural information is available, other less direct criteria can be used, such as the uniform distribution of the angular orientations assuring that all views needed to completely define the structure are present, high similarity between class averages and re-projections of the density map, and clear improvement of the resolution throughout the refinement. However, due to the high levels of noise, these criteria can also be fulfilled for erroneous 3D reconstructions if model bias is present (see Chapter 3).

The assessment of resolution of the single-particle reconstruction is another widely debated issue in the field (Frank, 2006; Henderson, 1995; Orlova and Saibil, 2004; Penczek et al., 1992; van Heel et al., 2000; van Heel and Harauz, 1986b; Yang et al., 2003; Zhang et al., 2008b; Zhou, 2008). The Fourier shell correlation (FSC) (Saxton and Baumeister, 1982), the phase residual (PR) calculated in resolution bands (Van Heel, 1987b), and the SSNR (Unser et al., 2005) have been proposed as resolution measures in different settings. As all these methods display different degrees of sensitivity to the noise present in the reconstructions, it is critical to confirm the estimated resolution by examining the resolved features, such as α-helices, bulky side chains, and strands in β-sheets (Zhou, 2008). Confirmation of the resolution is straightforward for a high-resolution reconstruction, the difficulties lie in determining the resolution of intermediately resolved maps.
CHAPTER 2: METHODS

THE PROJECTION IMAGE AND ITS DEGREES OF FREEDOM

In a TEM, the interaction between the incident electron beam and the 3D potential distribution of the specimen, \( o(x, y, z) \), is described by the phase object approximation (Cowley and Moodie, 1957). The 2D projection of \( o(x, y, z) \) is created from image intensity measurements. Electron scattering can be classified as elastic or inelastic. The former does not involve energy transfer, and carries the high-resolution information. Inelastic scattering gives rise to an undesired background in the image. Vitrified thin biological specimens are to a very good approximation electron transparent, thus the phase change due to elastic scattering is small, and image formation by the weak phase object approximation applies (Misell, 1976).

The electron microscopic image \( p(x, y) \) represents a projection of \( o(x, y, z) \) over the \( z \) dimension onto a plane orthogonal to the electron beam. The structure factor is encoded in the spatial distribution of the electron wave. The projected potential is convoluted with the point spread function (PSF), \( h(x, y) \), which describes the degree of spreading of the point object by the imaging system. The PSF is the real space representation of the CTF. In addition, the image is affected by the limited spatial coherence of the electron source caused by the finite source size, and the limited chromatic coherence due to the energy spread of the electrons. These effects are described by the envelope function \( e(x, y) \) (Frank, 1973), which limits the resolution by damping the higher spatial frequencies in the image’s Fourier transform \( \mathcal{F}\{p(x, y)\} \). The additive noise term is represented by a random vector \( n(x, y) \):

\[
p(x, y) = \int_{-\infty}^{+\infty} o(x, y, z) dz \otimes h(x, y) \otimes e(x, y) + n(x, y) \tag{2.1}
\]

where \( \otimes \) denotes a convolution. The Fourier representation of the EM projection is:

\[
\mathcal{F}\{p(x, y)\} = O(u, v) \text{CTF}(u, v) E(u, v) + N(u, v) \tag{2.2}
\]

where \( O(u, v) \) is the structure factor function, representing a central section of the 3D object’s Fourier transform (Crowther et al., 1970b).

Five degrees of freedom are needed to describe the 3D orientation of a 2D projection image: three Euler angles \( \psi, \theta, \varphi \); and two translational degrees of freedom \( t_x, t_y \). For
particles co-existing in multiple conformations the additional state assignment parameter, \( s \), needs to be accounted for. The windowed 2D projection image of a single molecule from a heterogeneous data set can be completely described as follows:

\[
p(x, y)_{\psi, \theta, \varphi, t_x, t_y, s} = T_{t_x, t_y} \left( \int_{-\infty}^{+\infty} R_{\psi, \theta, \varphi} \{ o_s(x, y, z) \} dz \right) \otimes h(x, y) \otimes e(x, y) + n(x, y) \quad (2.3)
\]

where \( T_{t_x, t_y} \) is a 2D translation matrix, and \( R_{\psi, \theta, \varphi} \) is a 3D rotation matrix. Assuming that the CTF deconvolution has been performed, 3D reconstruction from heterogeneous data sets requires \( 6N \) parameters to be determined, where \( N \) is the number of images:

\[
\{(\psi_i, \theta_i, \varphi_i, t_x, t_y, s_i)\}_{i=1}^N
\]  

(2.4)

**Unsupervised classification – dimensionality reduction and signal enhancement**

Unsupervised classification can be used to significantly reduce the dimensionality of the single-particle data by dividing it into subsets of similar images. Each group can be represented by a class average with improved SNR. Classification and averaging are meaningful only for an in-plane aligned data set. If no reference is available for 2D alignment, the reference-free in-plane alignment method based on the concept of alignment as a feature of the entire data set (Penczek et al., 1992) can be used to provide in-plane parameters of sufficient quality for the classification to resolve heterogeneity due to variations in projection direction (paper I). The reference-free in-plane alignment starts with finding a “random approximation” of the global average by randomly selecting, aligning, and averaging two images from the stack. This first approximation of the global average is used as a reference for alignment of a third randomly selected image, and a new average is generated. This initialization procedure continues until all images are aligned and added. The “randomly” estimated average is then iteratively refined. In each step of the refinement, an image is removed, and a “moving average” is calculated from the remaining images. The removed image is aligned to the “moving average”, and the global average is updated. The procedure cycles over all images until the rotations do not change significantly.
Towards unbiased 3D reconstruction in single-particle EM

The in-plane aligned projections form a multivariate data set, where each $D \times D$ pixel image can be represented by a vector in a $D \times D$ dimensional space. Principal Component Analysis (PCA) is a linear transformation that can be used to decompose the total inter-image variance into mutually orthogonal components, i.e. eigenvectors, ordered according to decreasing magnitude (Koeck et al., 1996; Lebart et al., 1984; van Heel and Frank, 1981). The image vector can be reconstituted in this factor space as linear combinations of the PCA basis without significant loss of information (Frank, 2006). Hierarchical ascendant classification (HAC) can be used to form all nested partitions of images based on their pair-wise distances in factor space, giving rise to a classification of the data into homogeneous groups (Frank, 2006; Lebart et al., 1984). Proximity in factor space directly reflects the similarity between images, and subsets of images sharing the same structural characteristics, i.e. same conformational state and view, are grouped together.

If a high quality in-plane alignment has been established, large data sets of electron microscopic projections can be efficiently grouped into homogenous classes in 2D by combining PCA with HAC. This is true when the variations in the data are due to differences in both projection direction and conformational state (paper I).

**JOINT AB INITIO 3D ALIGNMENT AND CONFORMATIONAL STATE SORTING**

Solving the problem of simultaneous ab initio 3D alignment and conformational state sorting addresses an important issue in the single-particle field as it enables reconstruction of molecular volumes free from model bias. Alignment to pre-existing structures or mixed reconstructions carries a great risk of introducing artifacts that template-based refinement will never overcome (see Chapter 3). A method for joint ab initio 3D alignment and conformational state sorting that aims at minimizing model bias and increasing the convergence radius of existing single-particle 3D reconstruction methods is presented in paper I and summarized below (see also Figure 2.1 below).

The method is formulated in Fourier space, and operates on class averages rather than raw images. In the first step, the RAD (Reference free Alignment in a Discrete angular space) algorithm (Elmlund et al., 2008) is used to align the class averages representing different views and conformational states in a single 3D reference frame. RAD uses simulated annealing (SA) for the combinatorial optimization. Detailed descriptions of SA can be found elsewhere (Kirkpatrick et al., 1983; Locatelli, 2000; Marinari and
Parisi, 1992). The low-pass filtered correlation, calculated over common lines between the permuted projection and the remaining “fixed” projections, is used to evaluate an SA transition:

\[
C_{SA} = \frac{\sum_{i=p}^{K} \sum_{m=r}^{hp} R e\{c_{i,m,o_i}^* c_{p_i,m,o_{p_i}}\}}{\left| \sum_{i=p}^{K} \sum_{m=r}^{hp} |c_{i,m,o_i}|^2 \sum_{i=p}^{K} \sum_{m=r}^{hp} |c_{p_i,m,o_{p_i}}|^2 \right|}
\]

where \( \phi_i = \{\psi_i, \theta_i, \varphi_i\} \) and \( 0 \leq \varphi_i, \psi_i \leq 2\pi \quad 0 \leq \theta_i \leq \pi \) \hspace{1cm} (2.6)

where \( i \) denotes the index of the Fourier transformed class average, \( p_i \in \{1,2, ..., K\} \) is the index of the Fourier transformed class average with altered angular orientation, \( m \) is the Fourier index, \( \psi, \theta, \varphi \) are the three Euler angles, and

\[
c_{i,m,o_i} = A_{i,m,o_i} + i\Phi_{i,m,o_i} \\
c_{p_i,m,o_{p_i}} = A_{p_i,m,o_{p_i}} + i\Phi_{p_i,m,o_{p_i}} \quad \hspace{1cm} (2.7)
\]

describe the real \( A \) and complex \( \Phi \) parts of the common line components in the respective coordinate systems of the planes \( i \) and \( p \) at Fourier index \( m \). \( r^{hp} \) and \( r^{lp} \) are the high-pass and low-pass temperature-dependent (see below) Fourier index limits, respectively.

The most important improvement to the original RAD algorithm is the complexity reduction introduced by restricting the number of class averages that are allowed to be assigned the same projection direction to the number of conformational states assumed to be present. This significantly reduces the search space, which leads to a marked improvement of solution quality when aligning large populations of class averages. The constraint is formulated based on the assumption that no two class averages represent the same view and conformational state. To further improve the convergence stability of RAD, a parameter that controls the direction of the angular orientation search is coupled to a dynamically controlled low-pass filter that excludes high frequencies at high temperatures and includes Fourier components of higher frequency in the low-temperature transitions. The coupling between the acceptance probability and the low-pass filter improves the robustness of the reference-free 3D alignment towards noise and reduces the amount of computation.
Labeling of class averages according to conformational state is accomplished before the reconstruction of volumes. A mixed continuous-discrete local search optimizer based on greedy randomized adaptive search (GRASP) (Feo and Resende, 1995) and differential evolution optimization (DE) (paper II) (Brest et al., 2006; Storn and Price, 1997) is used to maximize the spectral common line correlation coefficient over $5K$ continuous orientation parameters ($3K$ Euler angles and $2K$ translational degrees of freedom) and $K$ discrete state assignment parameters, where $K$ is the number of class averages:

Fig 2.1: Flowchart for the method. In the rhombic box initialization of variables is performed, rectangular boxes describe computations, and diamond boxes represent decision nodes. The hexagon divided into three regions represent a loop, with the initialization defined in the upper left region, the increment defined in the lower left region, and the stopping condition defined in the rightmost region. $X$ is the set of input projections and $I$ is their assigned in-plane parameters. $A$ is the set of class averages generated by unsupervised classification. $t$ is the number of states to resolve and $T$ is the upper limit of $t$. For each $t$ a 3D alignment $E(t)$ is obtained by RAD and the discrete solution is fed to a program maximizing a $6K$-dimensional spectral common line correlation coefficient $C(O)$ (equation 8) that determines $5K$ continuous orientation degrees of freedom and $K$ discrete state assignment parameters, where $K$ is the number of class averages. *Ab initio* reconstructions(s) are calculated and used to refine only in-plane parameters by projection matching, leading to progressively improved class homogeneity.
\[ \mathbf{O}^* = \max_{\mathbf{O}} \mathcal{C}_{\text{GRASP/DE}}(\mathbf{O}) = \max_{\mathbf{O}} \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \sum_{m=2 \lambda}^{j_p} \sum_{m_{\text{max}}}^{j_p} \text{Re} \left( c_{i,m,o}^* c_{j,m,o}^* \right)}{\left( \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \sum_{m=2 \lambda}^{j_p} \sum_{m_{\text{max}}}^{j_p} |c_{i,m,o}|^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \sum_{m=2 \lambda}^{j_p} \sum_{m_{\text{max}}}^{j_p} |c_{j,m,o}|^2 \right)^{\frac{1}{2}}}} \]

where \( \mathbf{O} = \{ \mathbf{o}_i \}_{i=1}^{K} \) and \( \mathbf{o}_i = \{ \psi_i, \theta_i, \varphi_i, t_{x_i}, t_{y_i}, s_i \} \)

\[ 0 \leq \varphi_i, \psi_i \leq 2\pi \quad 0 \leq \theta_i \leq \pi \quad -s_{\text{max}} \leq t_{x_i}, t_{y_i} \leq s_{\text{max}} \]

(2.8)

where \( \psi, \theta, \varphi \) are the three Euler angles, and \( s_{\text{max}} \in \mathbb{R} \) is an externally defined constant that controls the degree of translation. \( s_i \) is the integer state assignment parameter. The conditional statement \( \text{if } s_i = s_j \) in the second summation over \( j \) ensures that only common lines between projections assigned to the same state are contributing to the correlation coefficient.

The method relies heavily on the quality of class averages used for analysis. The reference-free 2D alignment algorithm (Penczek et al., 1992) provides in-plane parameters of sufficient quality for unsupervised classification to resolve heterogeneity due to differences in projection direction. Any templates generated from these first class averages will, however, be biased by conformational heterogeneity. To overcome this limitation, an iterative approach is proposed, where projection matching (Penczek et al., 1994) onto the ab initio generated reconstructions is used to improve the quality of the in-plane parameters, which in turn will drive the classification to resolve conformational heterogeneity. To avoid model bias, classification, orientation determination, and state assignment are all done independently of any template.

**High-resolution image orientation refinement**

A number of factors limit the ability of the refinement to converge to a high-resolution reconstruction. One problem is the often poor quality of the reconstruction used for initializing the process. Another obstacle is the existence of false maxima in the scoring function used for orientation search. Two sources are responsible for this effect: the existence of multiple peaks in the correlation function landscape due to the multiple correlation maxima existing along the tilt-axis orthogonal to the correct projection direction, and the noise (paper II) (Sigworth, 1998). A method for common lines-based orientation search in Fourier space, capable of dealing with high levels of
Towards unbiased 3D reconstruction in single-particle EM noise and poor starting reconstruction quality, is presented in paper II and briefly described below (see also Figure 2.2 below).

The algorithm is based on continuous global optimization of a cross common line correlation coefficient calculated in a particle dependent resolution range that is dynamically and automatically adjusted during the course of refinement. The spectral self-adaptation simultaneously accounts for the improving resolution of the reference volume and the information content in the individual images. Each round of refinement starts with exhaustive search in a discrete angular space, where each particle is aligned against a set of evenly distributed reference sections, using the translation parameters from the previous round. Only non-parallel planes contribute to the correlation. The 60 best discrete solutions are used to initialize a continuous space optimizer based on differential evolution (DE) that searches all parameters. The fittest solution (DE1 superscript) is used to define a local region of the parameter space in which the final round of DE is performed. Convergence is mapped by calculating the FSC plots between successive reconstructions.

Fig 2.2: Flowchart for the method. In rhombic boxes initializations of variables are performed, rectangular boxes describe computations, and diamond boxes represent decision nodes. Constant parameters include the high-pass frequency limit, the number of particles, the number of reference central sections, the Nyquist frequency, etc. The first round of refinement is performed using a conservatively chosen low-pass frequency limit (r3) common for all particles. In following rounds, the alignment data from the preceding round is read from disk, and a low-pass frequency limit is calculated accordingly.

A first alignment is performed exhaustively in a discrete angular space while keeping the translation parameters fixed. The 60 best discrete solutions are used to initialize a continuous space optimizer based on differential evolution (DE) that searches all parameters. The fittest solution (DE1 superscript) is used to define a local region of the parameter space in which the final round of DE is performed. Convergence is mapped by calculating the FSC plots between successive reconstructions.
best discrete angular solutions are used to initialize a continuous global optimizer, based on the DE algorithm (Brest et al., 2006; Storn and Price, 1997), which solves the following five dimensional optimization problem:

$$
o_i^* = \max_{o_i} \mathcal{C}_{DE}(o_i) = \max_{o_i} \frac{\sum_{i=1}^{R} \sum_{m=r_{hp}}^{r_{lp}} \text{Re}\left[ c_{i,m,o} e^{i \varphi_i p_i m o p_i} \right]}{\sqrt{\sum_{i=1}^{R} \sum_{m=r_{hp}}^{r_{lp}} \left| c_{i,m,o} \right|^2 \sum_{i=1}^{R} \sum_{m=r_{hp}}^{r_{lp}} \left| e^{i \varphi_i p_i m o p_i} \right|^2}}$$

where \( o_i = \{ \psi_i, \theta_i, \varphi_i \} \) and \( o_i^* = \{ \psi_{i^*}, \theta_{i^*}, \varphi_{i^*}, x_{i^*}, y_{i^*} \} \)

$$0 \leq \varphi_i, \psi_i, \psi_{i^*}, \varphi_{i^*} \leq 2\pi \quad 0 \leq \theta_i, \theta_{i^*} \leq \pi \quad -sh_{\max} \leq t_{x_{i^*}}, t_{y_{i^*}} \leq sh_{\max} \quad (2.9)$$

where \( i \) denotes the index of the reference central sections, \( i \in \{ 1, 2, ..., R \} \) and \( p_i \) is the index of the particle Fourier transform being aligned.

A conservative low-pass frequency limit common for all particles is chosen for the first round of refinement. In following rounds, a particle-dependent low-pass frequency limit is estimated according to the orientation found in the preceding round by using a Fourier shell common line correlation, defined as:

$$FCC(m) = \frac{\sum_{i=1}^{R} \text{Re}\left[ c_{i,m,o} e^{i \varphi_i p_i m o p_i} \right]}{\sqrt{\sum_{i=1}^{R} \left| c_{i,m,o} \right|^2 \sum_{i=1}^{R} \left| e^{i \varphi_i p_i m o p_i} \right|^2}} \quad (2.10)$$

Here, the summation over resolution \( (m) \) has been removed, resulting in a common line correlation coefficient with spectral dependence. \( FCC \) is used to automatically estimate the Fourier index \( r_{lp} \) of the low-pass limit. The dynamic filter is constructed for each individual image transform by selecting a resolution range with strictly positive shell correlation:

$$FCC(m) > 0 \Rightarrow r_{hp} \leq m \leq r_{lp} \quad (2.11)$$

A particle quality measure \( Q \), that maps the agreement between the target projection and the reconstruction, is calculated after the generation of a volume:

$$Q = \sum_{m=r_{hp}}^{r_{lp}} FCC_{\psi, \theta, \varphi, t_x, t_y}(m) \quad (2.12)$$

The method requires CTF phase correction, since the resolution of the filter would otherwise halt at the first sign shift.
CHAPTER 3: SIMPLE

The increasing popularity of single-particle cryo-EM is highly correlated with the technical and methodological advances in the field. Automated procedures for data acquisition allow for large single-particle sets to be obtained relatively quickly (Carragher et al., 2000; Lei and Frank, 2005; Potter et al., 1999; Suloway et al., 2005). The bottleneck is the image processing, which may proceed for weeks before it is even possible to determine if the data set is of quality sufficient for obtaining a high-resolution structure. Several mature software packages for single-particle reconstruction are available, e.g., SPIDER (Frank et al., 1996), EMAN (Ludtke et al., 1999), IMAGIC (van Heel et al., 1996), XMIPP (Sorzano et al., 2004b), but they are designed primarily for very experienced users, and a typical processing scheme involves extensive decision making and application of many different procedures, before the reconstruction is generated.

AN OVERVIEW OF SIMPLE

SIMPLE (Single-particle IMage Processing Linux Engine) (paper III) is a new open source software package that combines well-established image processing techniques (image manipulation, 2D alignment, unsupervised classification) with the above described methods for joint \textit{ab initio} 3D alignment and conformational state sorting (paper I), and high resolution image orientation refinement (paper II). The package provides a number of independent programs, each performing a specific task, that together form an efficient and easy-to-use image processing system for semi-automated \textit{ab initio} 3D reconstruction from challenging single-particle data sets (asymmetrical particles, significant degree of heterogeneity). SIMPLE is divided into a back-end object-oriented FORTRAN 95/2003 library implementing the core algorithms, and a front-end modular Perl library communicating with the user via command-line interfaces (CLIs). SIMPLE is portable to any UNIX-like platform, but optimal performance is achieved on distributed computing systems, where the program execution can be divided between nodes.

The SIMPLE methods have so far been used to generate 3D reconstructions of the AAA+ assemblies of cobalt and magnesium chelatase, visualized in different
conformational states induced by ATP hydrolysis, which have given important insights into the mechanism of the enzymes (supporting papers IV-V). The magnesium chelatase structure was among the first low symmetry reconstructions at subnanometer resolution of a previously unknown structure. The 12-subunit yeast RNA polymerase II (pol II) was visualized for the first time in both clamp open and clamp closed conformations (paper I). Crystallographic analysis only revealed the clamp closed form. The E-site t-RNA lacking ribosome conformation, which has not been resolved by other methods applied to the same well-characterized data set of the EF-G bound ribosome, has been revealed (paper I). Yet another successful example is provided by the reconstruction of the C2 symmetric NADPH-dependent thioredoxin reductase C (NTRC), resolved to a ~10 Å resolution (supporting paper VI).

**RUNNING SIMPLE**

*Ab initio* 3D reconstruction in SIMPLE is a two-stage process (see Figure 3.1). First, preliminary models are generated by using two programs: `align2D.pl` and `align3D.pl` that are cycled until convergence (Figure 3.1A). A stack of ~25k CTF corrected high-defocus single-particle images is required as an input. The `align2D.pl` program is responsible for in-plane alignment (-inplane flag) and unsupervised classification (-hcl, -cvags flags) of the particle stack. In the first round, reference-free in-plane alignment (Penczek et al., 1992) is performed. In consecutive rounds the in-plane parameters are established via projection matching (Penczek et al., 1994) onto the *ab initio* generated templates. The `align3D.pl` (-rfree flag) program takes the masked stack of class averages as an input, and performs joint *ab initio* 3D alignment and conformational state labeling (paper I and Chapter 2). The loop should be continued until the number of resolved states, and the quality of the resulting reconstructions from two consecutive rounds are stable.

The initial volumes are refined to high resolution (Fig. 3.1B) using the spectrally self-adapting orientation refinement method (paper II and Chapter 2) that is executed by giving the –rbased flag to `align3D.pl`. The input volumes should be masked, but never low-pass filtered. The centered (by `center.pl`) and Fourier transformed (by `ftstack.pl`) stack of particles used for refinement is usually much larger than the stack used for initial model generation. Reconstructions are calculated using `reconstruct.pl` only from particles with high $Q$ values (eq. 2.12) (the threshold is established by `findQ.pl`).
Towards unbiased 3D reconstruction in single-particle EM

Fig 3.1: SIMPLE processing scheme. (A) Preliminary model generation loop: ~25k stacked particle images are inputted to `align2D.pl` for in-plane alignment (-inplane flag) and unsupervised classification (-hcl and -cavgs flags). In the first round, reference-free in-plane alignment is performed. In consecutive rounds, the in-plane parameters are established via projection matching onto the *ab initio* generated templates. The stack of masked class averages is used by `align3D.pl` (-rfree flag) for *ab initio* 3D alignment and conformational state sorting. The resulting 3D volumes are inputted to another round of `align2D.pl`. The loop continuous until the number of resolved states and the resolution of the resulting reconstructions from two consecutive rounds are stable. (B) Refinement loop: The preliminary volumes are refined using `align3D.pl` (-rbased flag) that takes masked reconstructions, centered (by `center.pl`) and Fourier transformed (by `ftstack.pl`) particle images as input. The FULLstack.spi used for refinement (loop B) is usually larger than the stack used in the preliminary reconstructions generation loop (loop A). Reconstructions are generated by `reconstruct.pl` using only particles with high $Q$ values (`findQ.pl`).

**MODEL BIAS**

SIMPLE is unique with respect to other existing packages for single-particle reconstruction in that it enforces *de novo* structure determination for low-symmetry
particles. No pre-existing knowledge about the structure or the sample heterogeneity is used at any point of the processing, which significantly reduces the model bias. Unfortunately, the problem of model dependence has not been carefully studied in the field, and there seems to be many beliefs, rather than facts, circulating around the subject. It has for example been claimed that “The bias of initial 3D templates can be mitigated, if not eliminated altogether, by employment of a 3D version of a multireference alignment(...) The method constitutes a version of unsupervised classification in which class membership and orientation parameters are estimated simultaneously.” (Spahn and Penczek, 2009). In many existing approaches the initial reference volumes are either derived by randomly partitioning the data set (Penczek et al., 2006a; Sander et al., 2010; Scheres et al., 2007a), by adding the ‘neutral’ template to the pool of existing reconstructions (Gao et al., 2004), or by generating templates from Normal Mode Analysis (NMA) of the structure determined from the entire data (Brink et al., 2004). The general assumptions in these cases are that “the 3D map will be correct in its general features” (Spahn and Penczek, 2009), and that “the emergence of a significant density mass that is connected to the density of the large molecule is irrefutable evidence that the ligand has been localized, because the additional density built up even though it had no counterpart in the 3D reference.” (Frank, 2006). Both statements are of course true, but they do not exhaust the subject, since the reverse problem of what will happen if some features of the reference are not present in the data is not addressed. In this case, the effects of model bias can be significant enough to mislead the biological interpretation.

**AB INITIO VERSUS TEMPLATE-BASED APPROACHES**

The SIMPLE methods have been tested on a publicly available experimental cryo-EM data set containing 10000 images of 70S *E. coli* ribosomes. (see Paper I). As determined by supervised classification (Gao et al., 2004), the population was reported to be composed of an equal mixture of the two ribosome states, where the EF-G bound ratcheted ribosome contained a single t-RNA in the hybrid P/E site, whereas the EF-G lacking unracheted ribosome contained t-RNAs in the classic A/A, P/P and E/E positions. A larger overlapping data set has been used to test the performance of the ML-classification method (Scheres et al., 2007a), which essentially reproduced the results from the supervised classification. After processing this data using SIMPLE, the
EF-G lacking ribosome did not show any signs of the t-RNA density in the E/E position. Moreover, the L1 protuberance had swung away from the part of the intersubunit space that is overlapping with the E-site region (Figure 3.2a). This contrasts with what has been reported for template-based ML-classification (Scheres et al., 2007a) and supervised classification (Gao et al., 2004), but agrees with the results from the *ab initio* double Multivariate Statistical Analysis (MSA) method (Elad et al., 2008). The discrepancy is explained by the model bias introduced in the template-based approaches. To visualize this effect, the ribosome data set was subjected to five rounds of supervised classification using two reference volumes, low-pass filtered to 50 Å, representing an unratcheted ribosome with A/A, P/P and E/E site t-RNAs, and a ratcheted ribosome with the P/E site t-RNA, but stripped of EF-G. The experiment shows that the density not present in the reference appears if present in the data (the EF-G), but it is also evident that the density present in the reference does not disappear only because it lacks support in the data (the E/E site t-RNA) (Figure 3.2b).

![Fig. 3.2: Reconstruction of the E. coli ribosome.](image)

(a) Reconstructions from SIMPLE. Well-known quaternary structure regions are indicated. (b) Reconstructions from supervised classification. The E-site t-RNA is present in both reconstructions.
Methods that rely on alignment to pre-existing templates or mixed reconstructions do not assure self-consistency of the single-particle data, since the correlation will always “find” and reproduce features of the inputted reference, even if matching completely unrelated structures. This is due to the low SNR of the cryo-EM data. To illustrate this effect, the experimental ribosome data described above was reconstructed by supervised classification, using 10- and 12-subunit pol II density maps as templates. The two pol II density maps were scaled to match the molecular weight of the *E. coli* ribosome, and low-pass filtered to 50 Å, before generation of 2000 reference projections of each volume. Five rounds of supervised matching of the ribosome data onto the pol II references resulted in two “conformational states” resolved to ~20 Å. Projections of the volumes in five evenly distributed angular orientations agreed disturbingly well with projection averages (Figure 3.3). In contrast, application of a single iteration of SIMPLE method, when using pol II templates for in-plane alignment only, resulted in feature-less blobs without resemblance to either pol II or the ribosome. The FSC-plots show ~50 Å at FSC=0.5, and projections show little resemblance to their corresponding class averages (Fig. 3.3c shows one of the reconstructions).

In conclusion, the choice of reconstruction method is of crucial importance for the outcome, and if templates are used, validation with an *ab initio* approach should be provided.
Fig. 3.3: Reconstruction of real ribosome data by alignment to RNA polymerase II. (a) Example ribosome images (left) and templates used for supervised matching (right). (b) Results from supervised matching, left: projections (upper row) versus class averages (lower row), middle: reconstructions, and right: FSC-plots. (c) Corresponding results from a single iteration of SIMPLE approach, when using pol II templates for in-plane alignment.
CHAPTER 4: PAPERS

PAPER I


Main conclusion: A common lines-based optimization-driven method enables ab initio 3D reconstruction from challenging single-particle data sets (asymmetrical particles, significant degree of heterogeneity), and reduces model bias.

My contributions: Design of research (together with HE), method development and analysis of results (together with HE), data processing and figure preparation, writing of the manuscript (together with HE).

PAPER II


Main conclusion: Fast and robust convergence to high-resolution is demonstrated for a single-particle refinement method based on differential evolution optimization of a spectrally self-adaptive common line correlation coefficient on real and simulated data.

My contributions: Design of research (together with HE), method development and analysis of results (together with HE), data processing and figure preparation, writing of the manuscript (together with HE).

PAPER III

D. Elmlund and H. Elmlund “SIMPLE – an image processing system for ab initio 3D reconstruction in single-particle electron microscopy”, manuscript

Main conclusion: Presentation of the easy-to-use single-particle image processing system SIMPLE for challenging data sets (asymmetrical particles, significant degree of heterogeneity).
My contributions: Coding of all front-end modules, all executable programs, and a few back-end classes, data processing and figure preparation, writing of the manuscript (together with HE).

**PAPER IV**


Main conclusion: An ATP-fueled and Integrin-I mediated conformational rearrangement of the *R. capsulatus* magnesium chelatase AAA+ activation complex is required for the holoenzyme assembly.

My contributions: Data collection of the ATP-induced complex, initial data processing.

**PAPER V**


Main conclusion: Subunits CobS and CobT of the *B. melitensis* cobalt chelatase form a two-tiered ring structure with six subunits in each ring organized as a trimer of dimers.

My contributions: Data collection and processing (together with JL).

**PAPER VI**

R.P. Wulff et al. “NADPH-dependent thioredoxin reductase C and its possible involvement in the cyclase reaction of the chlorophyll biosynthetic pathway”, *manuscript*

Main conclusion: NTRC in combination with its target 2-Cys peroxiredoxin stimulates cyclase activity in barley.

My contributions: Data collection (together with JL), design of the processing scheme used, helped with validation of the reconstruction.
ACKNOWLEDGEMENTS

Hans Elmlund – for sharing with me your passion for science in general, and EM in particular, your never ending enthusiasm, having more ideas per minute than can ever be tested, and for always believing in me. It has been a lot of fun but also a lot of struggle. We did a great job.

Claes Gustafsson, Hans Hebert and Philip Koeck – for giving me a lot of freedom and trust, for always supporting me both scientifically and financially, and for patiently answering my many questions (special thanks to Philip). Thank you for letting me perform science in, what I believe is, a very untraditional way.

Joakim Lundqvist – for putting up with me and all my questions already when I was a master student at Molecular Biophysics in Lund, and for your enthusiasm and hard work that helped us develop and improve SIMPLE. I had a great time first in the “monkey cage” and then at Carlsberg!

Mats Hansson – for your hospitality, enthusiasm, and trust in us. It was absolutely wonderful to be a part of your group at Carlsberg.

André Müller – for your enthusiasm and ability to learn quickly.

Salam Al-Karadaghi – for taking me as a master student and for throwing me directly out on deep waters.

Kimberley Cheng – for being very helpful and friendly and for all your pep talk in the emails.

Martin Lindahl – for being patient with my questions.

Pasi Purhonen – for always helping me out with the microscope.

Other colleagues at NOVUM: Caroline, Sunny, Emmanuella, Olga, Paulina, Zsolt, Henrik, Neli, and Venus for all fika chitchats, scientific advice or help with endless paper work.

All members of the Carlsberg team: Simon, Shakira, Ragna, Marzena for taking us in, all crazy discussions at a lunch table, and all the fun we had together.

Tack till Alf och Karin och alla Lundbergare och Salomonsonare för att ni har hjälpt mig att hantera honom (ni vet vem). Tack för ert stöd och varma välkomnande av mig in i familjen.
Na koniec specjalne podziękowania dla mojej rodziny, zwłaszcza moich rodziców, za życzliwość, wsparcie, częste odwiedziny i jeszcze częstszes rozmowy na skypie. Bez Was nie byłoby mnie tu!
REFERENCES


automated acquisition of 1000 electron micrographs a day. Ultramicroscopy 77, 153-161.


