PROTEIN STRUCTURE
DYNAMICS AND INTERPLAY

BY SINGLE-PARTICLE ELECTRON MICROSCOPY

HANS ELMLUND

LUND 2008

KUNGLIGA TEKNISKA HÖGSKOLAN
THESIS FOR DOCTORAL DEGREE
to my father
ABSTRACT

Single-particle cryo-electron microscopy (cryo-EM) is a method capable of obtaining information about the structural organization and dynamics of large macromolecular assemblies. In the late nineties, the method was suggested to have the potential of generating “atomic resolution” reconstructions of particles above a certain mass. However, visualization of secondary structure elements in cryo-EM reconstructions has so far been achieved mainly for highly symmetrical macromolecular assemblies or by using previously existing X-ray structures to solve the initial alignment problem. A factor that severely limits the resolution for low-symmetry (point group symmetry C\text{\textsubscript{n}}) particles is the problem of \textit{ab initio} three-dimensional alignment of cryo-EM projection images of proteins in vitreous ice.

A more general problem in the field of molecular biology is the study of heterogeneous structural properties of particles in preparations of purified macromolecular complexes. If not resolved, structural heterogeneity limits the achievable resolution of a cryo-EM reconstruction and makes correct biological interpretation difficult. If resolved, the heterogeneity instead offers a tremendous biological insight into the dynamic behaviour of a structure, and statistical information about partitioning over subpopulations with distinct structural features within the ensemble of particles may be gained.

This thesis adds to the existing body of methods in the field of single-particle cryo-EM by addressing the problem of \textit{ab initio} rotational alignment and the problem of resolving structural heterogeneity without using \textit{a priori} information about the structural variability within large populations of cryo-EM projections of unstained proteins. The thesis aims at making the single-particle cryo-EM method a generally applicable tool for generating subnanometer resolution reconstructions and perform heterogeneity analysis of biological macromolecules.
LIST OF PAPERS


PAPER NOT INCLUDED IN THIS THESIS

Cheng K., Koeck P. J. B., **Elmlund H.** Idakieva K., Parvanova K., Schwarz H., Ternstrom T., and Hebert H. (2006) ”Rapana thomasiana hemocyanin (RtH): Comparison of the two isoforms, RtH1 and RtH2, at 19 angstrom 16 angstrom resolution”, *Micron 37*, 566-576
PREFACE

I was firmly determined to study theoretical chemistry as a graduate student. The theoretical framework of statistical thermodynamics and the simulated annealing-based computational methods used to study molecular interactions in fluids and polymers fascinated me. By a coincidence, and despite my limited biochemical knowledge, the last course I attended as an undergraduate student was a course in structural biochemistry. The course was divided into three parts: X-ray crystallography, NMR-spectroscopy and electron microscopy. Unfortunately, the professor teaching NMR got ill and I never got the chance to learn about spin dynamics, but the very enthusiastic electron microscopist kindly filled the schedule with Fourier optics, theory for image formation in the transmission electron microscope and single-particle methodology. We were only two students with rather technical backgrounds attending the course and we were given a solid mathematical background to biological transmission electron microscopy. I discovered a new world – the world of large macromolecular assemblies. After seeing my first electron micrograph of single Mediator molecules and realizing the enormous potential of a method for structure determination that does not give ensemble averaged information and does not require a crystalline specimen, I knew that I would never become a theoretical chemist. This thesis marks the beginning of my hopefully long lasting relationship with single-particle electron microscopy.

Hans Elmlund, Lund 2008
# CONTENTS

Chapter 1: Background ........................................................................................................... 7
  Protein dynamics .................................................................................................................. 7
  Methods for gaining dynamic structural information ....................................................... 9
  Single-particle cryo-EM ..................................................................................................... 10
  Single-molecule techniques ............................................................................................... 11

Chapter 2: Methods ................................................................................................................. 13
  The single-particle image ................................................................................................. 13
  Coherence ......................................................................................................................... 15
  Deconvolution of the CTF by Wiener filtering ................................................................. 15
  Optimization problems in single-particle electron microscopy ........................................ 16
  The central section theorem and the concept of common-lines ....................................... 17
  Normalized correlation coefficients ............................................................................... 18
  Averaging ......................................................................................................................... 19
    The concept of 2D alignment as a property of the entire image set ............................... 20
    Principal component analysis ...................................................................................... 20
    Unsupervised classification in factor space ................................................................ 20
  The orientation search problem ...................................................................................... 22
    Simulated annealing ....................................................................................................... 24
    A generalized simulated annealing algorithm ............................................................... 24
    Reference-free alignment in a discrete angular space (RAD) ....................................... 26
  Disentangling conformational or morphological states ................................................... 27
    Orientation directed classification and sub classification ........................................... 29
    Assigning subclass averages to 3D reconstructions .................................................... 30
    Separating states directly from the class averages ...................................................... 33
    Common line correlation driven supervised classification ........................................ 34
  Reconstruction of a 3D volume ....................................................................................... 35
  Validation ......................................................................................................................... 35
  From a novel cryo data set to an intermediately resolved reconstruction ....................... 36

Chapter 3: Magnesium chelatase ........................................................................................ 39
  The chelation process and the Mg-chelatase subunits ..................................................... 39
  A quasi-atomic model of the ADP induced ID complex .................................................. 42
  Concluding remarks ......................................................................................................... 47

Chapter 4: The eukaryotic transcription machinery ........................................................... 49
  The general transcription machinery ............................................................................. 49
    The Mediator complex ................................................................................................. 51
    A model for stimulation of transcription by Mediator .................................................. 52
  TFIID ............................................................................................................................... 56
    Structural architecture of TFIID .................................................................................. 56
    Histone targeting and modification ......................................................................... 57
    Promoter recognition ................................................................................................. 58
    Activated transcription ............................................................................................... 59
    Structure of TFIID (paper III) .................................................................................... 60
    DNA binding to TFIID requires removal of TBP ......................................................... 61
    TFIID as the “mediator of the scaffold complex”? ...................................................... 62
  Concluding remarks ......................................................................................................... 65

Chapter 5: Summary and perspectives ............................................................................... 67
  Paper I ............................................................................................................................. 67
  Paper II ........................................................................................................................... 67
  Paper III .......................................................................................................................... 68
  Paper IV .......................................................................................................................... 69
  Future perspectives ........................................................................................................ 69

Acknowledgements ............................................................................................................ 70
References ........................................................................................................................... 73
# List of Abbreviations

<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>BchD, D</td>
<td>Mg-chelatase D-subunit</td>
</tr>
<tr>
<td>BchH, H</td>
<td>Mg-chelatase H-subunit</td>
</tr>
<tr>
<td>BchI, I</td>
<td>Mg-chelatase I-subunit</td>
</tr>
<tr>
<td>CDK8</td>
<td>Cyclin Dependent Kinase 8</td>
</tr>
<tr>
<td>CTD</td>
<td>Carboxy-Terminal Domain (of Rpb1 of RNA polymerase II)</td>
</tr>
<tr>
<td>CTF</td>
<td>Contrast Transfer Function</td>
</tr>
<tr>
<td>EM</td>
<td>Electron Microscopy</td>
</tr>
<tr>
<td>FSC</td>
<td>Fourier Shell Correlation</td>
</tr>
<tr>
<td>HFD</td>
<td>Histone Fold Domain</td>
</tr>
<tr>
<td>NMA</td>
<td>Network normal Mode Analysis</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>PIC</td>
<td>Pre-Initiation Complex</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>PPIX</td>
<td>Protoporphyrin IX</td>
</tr>
<tr>
<td>RAD</td>
<td>Reference-free Alignment in a Discrete angular space</td>
</tr>
<tr>
<td>Rpb</td>
<td>RNA polymerase b (II)</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing</td>
</tr>
<tr>
<td>SAGA</td>
<td>Spt-Ada-Gcn5-Acetyl transferase</td>
</tr>
<tr>
<td>SMS</td>
<td>Single Molecule Spectroscopy</td>
</tr>
<tr>
<td>SRB</td>
<td>Suppressor of RNA polymerase B (II)</td>
</tr>
<tr>
<td>TAF</td>
<td>TATA box binding protein Associated Factor</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA box binding protein</td>
</tr>
<tr>
<td>TEM</td>
<td>Transmission Electron Microscope</td>
</tr>
<tr>
<td>TFII</td>
<td>Transcription Factor class II</td>
</tr>
</tbody>
</table>
CHAPTER 1: BACKGROUND

This thesis is primarily concerned with studies of computational methods for *ab initio* 3D reconstruction from homogeneous and heterogeneous populations of single-particles, imaged using a transmission electron microscope. It also deals with the application of newly developed and existing methods in structure-function studies of large macromolecular assemblies. The material is presented in five chapters and four papers (I-IV). In chapter 1 the dynamic behaviour of proteins and methods for gaining dynamic structural information are discussed briefly. The background to the newly implemented methods for *ab initio* 3D reconstruction, together with the underlying mathematical framework, is presented in chapter 2. A short introduction to Mg-chelatase – the enzyme responsible for the catalysis of the insertion of Mg$^{2+}$ into protoporphyrin IX in the first committed step of the chlorophyll biosynthetic pathway, is presented together with results from the application of the novel methods in studies of the structural organization of the enzyme in chapter 3. Structure-function studies of the Mediator complex and the general transcription factor IID, which are both part of the eukaryotic general transcription machinery, are presented in chapter 4 together with a background to eukaryotic transcription regulation, founded on results primarily available prior to the work carried out in this thesis. The last chapter, chapter 5, summarizes the results in papers I-IV, and provides suggestions for further investigations.

**PROTEIN DYNAMICS**

There is an intimate connection between macromolecular dynamics and the structure-function relationship of biological systems. The conformational rearrangement of a macromolecular assembly is a highly organized activity that may be driven by binding of smaller molecules, interactions between complexes, nucleoside triphosphate hydrolysis or other sources of energy (Alberts, 1998). The ability of macromolecular subunits or domains to move and interact with their surrounding determines the outcome of most chemical events in the cell. Proteins bind other molecules and a protein molecule physically interacts with other macromolecular structures in a very specific manner, which determines its biological function. Binding could be tight or weak and short-lived, but it often shows a high degree of specificity, as demonstrated
by the ability of proteins to discriminate between the many thousands of different binding surfaces they encounter and the few ones that they selectively bind. The exact three-dimensional arrangement and composition of amino acid side chains gives the protein surface its unique chemical properties. A conformational rearrangement may change the surface structure and alter its chemical properties. The conformational behaviour of a protein molecule therefore ultimately determines its chemistry.

Spatial and temporal coordination of chemical events in the cell requires means for regulating protein activity. There are several levels of regulation. The activity of a protein may be regulated by its occurrence in the cell, which is controlled primarily by gene expression. The mechanisms underlying regulation of gene expression at the level of DNA transcription are discussed in chapter 4. Another mean for regulation is the compartmentalization of chemical events to bounded regions of the cell, often enclosed by specific membrane structures. All DNA of the cell is for example condensed into chromatin and enclosed in the nuclear compartment, which makes the process of transcription spatially limited to the region inside the nuclear envelope. This allows for fine-tuning of the enzymatic composition in the nucleus and control of the transport of signalling molecules over the membrane. However, the far most rapid and general regulatory mechanism is to adjust reaction rates through a direct and reversible change in the enzyme responsible for the process targeted for regulation. This may be achieved by allosteric control, in which the communication between regulatory sites and the active site of an enzyme is responsible for the regulation. Conformational rearrangements of protein complexes are central to allosteric control and ligand binding may induce movements or folding events that affect the chemistry of an enzyme or ribozyme. The structural variability of the ATP induced Mg-chelatase ID-complex (paper II) offers an example of a cooperative allosteric transition, in which the hydrolysis of ATP in the ATPase active ring trigger a conformational change that is transmitted to the neighbouring ATPase inactive ring, leading to the exposure of new surface areas with presumably different chemical properties. This conformational rearrangement is required for binding of the adaptor protein and completion of the active enzyme. The composition of protein subunits into larger complexes may itself offer a way to regulate protein structure and activity, which will be exemplified in chapter 4, when discussing the role of the general transcription factor IID in transcription regulation.
METHODS FOR GAINING DYNAMIC STRUCTURAL INFORMATION

To study the dynamic behaviour of proteins by experimental methods for structure determination, time-resolved methods or methods capable of studying heterogeneous populations of molecules are required. Heterogeneity limits the applicability of ensemble averaged techniques, such as X-ray or electron crystallography or small angle X-ray scattering. A high degree of polydispersity or conformational flexibility in a preparation precludes crystal formation and a heterogeneous collection of scattering objects gives an ensemble averaged signal, impossible to interpret without additional structural information. Despite this, dynamic information may be gained by solving structures under different biochemical conditions, and the two resolved conformations of RNA polymerase II, solved with electron crystallography from two different crystal forms, is one such example (Asturias et al., 1997). Another example is the 3D reconstructions of the different binding states of Mediator, presented in paper IV, which give dynamic information about Mediator upon RNA polymerase II association. Chemical processes that can be induced in a fast and precise manner in a crystal environment may be studied by Laue diffraction and synchrotron radiation, which offers the possibility of dynamic structural information in the 150 ps-1 s range (Drenth, 1994), but the requirement of crystals with low mosaicity that are capable of withstanding conformational rearrangements in the asymmetric unit and several short exposures to extremely intense X-ray radiation is very hard to meet. NMR-spectroscopy offers the interesting possibility of gaining dynamic structural information by measuring mobility directly (Cavanagh et al., 1996; Levitt, 2001), but the limitations imposed by the relaxation properties of nuclei in large molecules and the problem of spectral overlap makes structure determination of proteins with more than 300 amino acid residues very hard, even if structural information about parts of much larger systems may be obtained. One example is the NMR study of the 20S proteasome (Sprangers and Kay, 2007), where the molecular weight limitations are overcome by using an isotope labelling scheme where methyl groups of certain aminoacid residues are protonated in an otherwise highly deuterated background. In concert with experiments that preserve the lifetimes of the resulting NMR signals, insights into pico-to nanosecond timescale side-chain dynamics are gained. The combination of conventional high-resolution ensemble averaged techniques, such as X-ray crystallography and NMR-spectroscopy, with single-particle electron microscopy and cellular tomography offers a seamless integration of resolution ranges from the detailed
level of atomic organization of macromolecular subunits to the level of intermediately resolved macromolecular complexes.

**SINGLE-PARTICLE CRYO-EM**

Single-particle cryo-EM avoids ensemble averaging and therefore has the ability of capturing reaction intermediates and resolving heterogeneous populations of molecules, see for example (Saibil, 2000). Few other methods are capable of studying the structural dynamics of entire MDa protein complexes, and there are mounting lines of evidence that large macromolecular assemblies are dynamic machines, see for example (Brink et al., 2004; Chacon et al., 2003; Frank and Agrawal, 2000; Grob et al., 2006; Heymann et al., 2003; Zhou et al., 2001). These abilities make single-particle cryo-EM particularly powerful in the study of the mechanisms underlying the biological function exerted by the individual or collective behaviour of macromolecular complexes, and their highly correlated and dynamic behaviour. The method is capable of resolving the structures of macromolecular complexes in the size range 200-10,000 kDa at a level where secondary structure elements may be identified, as demonstrated for example by (Cheng et al., 2004; Ludtke et al., 2004; Schuler and al, 2006) (paper I-III). The identification of helix regions may allow the protein or RNA backbone to be traced.

Reference-free reconstruction methods that provide verifiable 3D reconstructions directly from a set of unstained and un-tilted projections of unknown orientations are obviously advantageous. Such reconstruction methods may be constructed by using alignment protocols that rely on the *central section theorem* and the concept of common lines (Crowther et al., 1970b; Goncharov and Gelfand, 1988; Lindahl, 2001; Penczek et al., 1996; van Heel, 1987) (paper I). A problem with reference-free common lines-based methods is that a large number of projections must be simultaneously aligned, which results in the exponential growth of the sampling space with the growth of the number of projections. To circumvent this problem, the entire data set may be subjected to translational and in-plane rotational alignment, followed by 2D classification, during which images of particles in the same (or similar) orientations are grouped together into classes. For each class, a 2D average may then be calculated and the entire data set can be represented by a much smaller number of projections with enhanced signal to noise ratio. The averages can be aligned in a discrete and evenly distributed angular space, as described in paper I. A reference-free reconstruction is generated by methods for 3D reconstruction from aligned projections. A problem with
traditional common lines based methods for *ab initio* reconstruction is their sensitivity towards structural heterogeneity. This thesis addresses this problem, and novel common lines-based methods for handling structural heterogeneity in large populations of unstained single-particle images with unknown orientations are described in detail in the next chapter.

**SINGLE-MOLECULE TECHNIQUES**

Single-molecule techniques have started to change the way we think about biochemical processes. The study of chemical reactions at the single molecule level allows for direct measurements of distributions of molecular properties, rather than their ensemble averages. A rapidly growing field is the field of single-molecule fluorescence spectroscopy (Weiss, 2000). Experimental conditions that synchronizes all protein molecules in an active preparation in time along a certain reaction pathway is not in general possible to establish. The ability of single molecule spectroscopy (SMS) to measure conformational dynamics of biomolecules and record asynchronous time trajectories of physical properties that would be hidden for classical ensemble averaged spectroscopy is therefore appealing. Single-particle cryo-EM and SMS are complementary in the sense that the visualization of structurally distinct subpopulations can be coupled to investigations of chemical properties of subpopulations by SMS. One example is the study of the molecular mechanisms underlying muscle contraction. Here, the actin filament moves past the myosin filaments as a result of the hydrolysis of ATP. The molecular entanglement between actin and myosin has been studied by cryo-electron microscopy (Holmes et al., 2003) and the mechanism for nucleotide release has been revealed. Time-resolved SMS has been used to track conformational changes in myosin (Forkey et al., 2003), which may be interpreted in context of the available structural data. SMS has also been used to study the asynchronous behaviour of the prokayotic RNA polymerase (Herbert et al., 2006; Neuman et al., 2003). In (Burley and Roeder, 1998) a challenge facing molecular biologists was predicted to be the task of going beyond the static pictures provided by the classical ensemble averaged methods for high-resolution structure determination and characterize the kinetic and thermodynamic properties of the large transcriptionally active nucleoprotein complexes present in the eukaryotic cell. Perhaps the combination of SMS and single-particle electron microscopy provides the required methodology.
CHAPTER 2: METHODS

A description of the transmission electron microscope and the physics of electron scattering and image formation is beyond the scope of this thesis, and it has been the subject of excellent books by J. C. H. Spence (Spence, 1988) and L. Reimer (Reimer, 1993). This chapter is not aimed at describing the complete single-particle methodology, which has been the subject of a very comprehensive and stringently formulated work by J. Frank (Frank, 2006). The focus of this chapter is rather to describe the novel method for \textit{ab initio} reconstruction from homogeneous cryo-EM single-particle populations of randomly oriented and unstained macromolecules, utilized in paper I-III, and the method for separation of conformational or morphological macromolecular states from heterogeneous single-particle populations, used to study the ATP-fueled motions of the Mg-chelatase ID-complex in paper II, and used to resolve the heterogeneous TFIID population in paper III.

THE SINGLE-PARTICLE IMAGE

Two sources of contrast exist in the TEM image. Amplitude contrast refers to imperfect electron transparency of the object being imaged, leading to absorption or inelastic scattering of electrons, so that areas with greater absorption appear darker. A pure phase object is electron transparent and merely scatters the incident illumination elastically, which gives rise to phase contrast. The amplitude/phase contrast ratio is dependent on the atomic species and heavier atoms give a larger portion of amplitude contrast. Vitrified specimens of unstained macromolecular complexes are to a very good approximation pure phase objects and scattering is relatively weak, so that the theory for image formation by a weak phase object applies (Reimer, 1993; Spence, 1988). Let $o(x,y,z)$ denote the 3D electrostatic potential distribution of the thin object being imaged in the TEM. Considering only the elastic scattering interactions of the electrons with the specimen, the 2D image, $im(x,y)$, represents a projection through the $o(x,y,z)$ potential, with the direction of the projection determined by the direction of the electron beam, $z$. The structure factor is encoded in the spatial distribution of the electron wave. The projected potential is convoluted with (1) the point spread function, $h(x,y)$ of the microscope for a bright field image formed with a central objective aperture, and (2) the envelope function, $e(x,y)$. The Fourier transform of the envelope
function describes the Fourier amplitude decay in reciprocal space. To the image a random vector, \( n(x,y) \) is added, which represents the additive noise term:

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

(2.1)

The Fourier representation of this image is:

\[
\mathcal{Z}\{\im(x, y)\} = O(u,v)\text{CTF}(u,v)E(u,v) + N(u,v)
\]

(2.2)

\( O(u,v) \) is the structure factor function, which represents a central section of the 3D object’s Fourier transform (Crowther et al., 1970b). The point spread function of the imperfect optical imaging system of the TEM reproduces a point object as an airy disc, leading to a correlation between nearby lying pixels in the digitized image. This feature of all coherent imaging systems results in that not all input spatial frequencies give the same contrast in the output image. This behaviour is described by a transfer function. The contrast transfer function (CTF) of an electron microscope is the Fourier transformation of its point spread function. The CTF is a sinusoidal function, resembling the first order Bessel function, which oscillates at a higher rate at higher spatial frequencies. An analytical expression of the CTF may for example be found in (Frank, 2006; Reimer, 1993; Spence, 1988), and the overall dependence of the non-astigmatic CTF on resolution, wavelength, defocus and spherical aberration is (Baker and Henderson, 2001):

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

(2.3)

where \( \chi(v) = \pi \lambda^2 (\Delta f - 0.5C_s \lambda^2 v^2) \), \( v \) is the spatial frequency (in Å\(^{-1}\)), \( F_{\text{amp}} \) is the fraction of amplitude contrast, \( \lambda \) is the electron wavelength (in Å), \( \Delta f \) is the defocus (in Å) and \( C_s \) is the spherical aberration of the objective lens of the microscope (in Å). Most important to note is the dependency of the CTF on defocus, leading to contrast inversion of different regions of the reciprocal space with different defocus settings. In the zero crossings of the CTF no information other than noise is present. It is therefore important to acquire the images contributing to a reconstruction at different defocus settings, in order to avoid systematic loss of information. The alternating positive and negative zones of the CTF are rotationally symmetric provided that the axial astigmatism is fully compensated. The combined effect of defocusing and lens aberrations is of crucial importance for the TEM imaging system. By collecting data at
defocus settings lower than that of Scherzer focus, the contrast of the low-dose image is improved, but a deconvolution of the CTF has to be performed in order to retrieve high-resolution information (Unwin and Henderson, 1975). Other means for improving the contrast in images of frozen hydrated specimens exist. By lowering the acceleration voltage of the electron source the interactions between the electrons and the specimen get more pronounced (Spence, 1988), which improves the contrast but increases the risk of damaging the specimen by the electron radiation.

**COHERENCE**

In the weak phase object approximation (Frank, 2006; Reimer, 1993; Spence, 1988), the high-resolution details in an electron micrograph arise from the coherent interference between the scattered and the unscattered electron wave. The envelope component, $E(u,v)$ ultimately limits the resolution. Its damping character is mainly due to partial temporal and spatial coherence of the electron beam. High temporal coherence is achieved by a stable voltage supply to the electron source, resulting in a narrow energy spread of the electrons. The tip from which the electrons emerge in the electron source cannot be made infinitely thin. Therefore the electrons do not emerge from the same spot and cannot be described by the same quantum mechanical wave function. This leads to partial spatial coherence. Other factors contributing to the decay of high resolution Fourier components are ice-thickness, image drift, specimen charging, heterogeneity and alignment errors (Frank, 2006; Jensen, 2001).

**DECONVOLUTION OF THE CTF BY WIENER FILTERING**

The goal of a Wiener filter is to filter out noise that has corrupted a signal. The design of the filter must take the transfer function of the optical system into account. The parametrization of the electron microscopic CTF may be performed manually, as for example in *ctfit*, which is part of the Eman suite of programs (Ludtke et al., 1999), or with some kind of computational curve fitting approach. A straightforward way to correct for the distortions of the CTF is to perform a division of the image’s Fourier transform by the CTF, but this naïve approach would risk amplifying noise in the regions where no other information is present. The electron microscopic Wiener filter can be described as a “careful division” of the CTF, also taking the noise and the
Fourier amplitude decay into account. Following the formalism in (Frank, 2006), the image in Fourier notation is:

\[ \Im \{im(x, y)\} = O(u, v) CTF(u, v) E(u, v) + N(u, v) \]

\[ = O(u, v) H(u, v) + N(u, v) \]  

(2.4)

We seek an estimate \( \hat{O}(u, v) \) minimizing the expectation value of the squared difference between this estimate and the structure factor function:

\[ \left\langle \left| O(u, v) - \hat{O}(u, v) \right|^2 \right\rangle = \min \]  

(2.5)

and look for a filter function, \( S(u,v) \) with the property:

\[ \hat{O}(u, v) = S(u, v) (O(u, v) H(u, v) + N(u, v)) \]  

(2.6)

This filter function is obtained under the assumption that there is no correlation between \( O(u, v) \) and \( N(u, v) \):

\[ S(u, v) = \frac{H^*(u, v)}{\left| H(u, v) \right|^2 + P_{\text{noise}}(u, v)/P_{\text{obj}}(u, v)} \]  

(2.7)

where \( P \) denotes a power spectra. \( P_{\text{noise}}(u, v)/P_{\text{obj}}(u, v) \) represents the noise to signal ratio, and it may be estimated by a constant value. To “correct” the Fourier transforms of the individual particles of a defocus group (a group of particle images sharing the same CTF), their Fourier transforms are multiplied component-wise with the Wiener filter. The additive noise to signal term in the denominator of equation (2.7) prevents excessive noise amplification in the neighbourhood of \( H(u,v) = 0 \). The denominator is dominated by the noise to signal ratio at high resolution due to the envelope component of \( H(u,v) \). A high constant value of the noise to signal ratio therefore leads to suppression of finer details in the image, similarly to applying a low pass filter. In this thesis, CTF-correction is performed under the assumption that all single molecule images from the same micrograph share the same CTF.

**Optimization problems in single-particle electron microscopy**

Many of the optimization problems encountered in single-particle electron microscopy are related to alignment. Five degrees of freedom are required to completely describe
the 3D orientation of a 2D projection image: Three Euler angles and two translational
degrees of freedom, see for example (Lindahl, 2001):
\[
\{\psi_i, \theta_i, \phi_i, x_i, y_i\}_{i=1,N}
\] (2.8)

5N parameters are related to the orientations, where N is the number of images in the
data set. There are a number of other parameters to consider, some of which have
traditionally been treated as micrograph invariant. These parameters are: magnification,
defocus, astigmatism, Fourier amplitude decay and beam tilt. In this thesis, these
parameters have been handled in the traditional approach. One point that is important to
emphasize is the requirement of a homogeneous set of particles for inclusion in a 3D
reconstruction. Image selection and elimination of structurally deviating particles has
traditionally been performed manually or by the use of multivariate statistics (van Heel
and Frank, 1981). However, the recent technical development in the field (Gao et al.,
2004; Penczek et al., 2006; Scheres et al., 2007) has opened the door to fully automatic
optimization procedures for resolving conformational or morphological states or
sorting out heterogeneous particle views from large populations of unstained single-
particle images. One such method, founded on simulated annealing optimization of the
joint common line correlation coefficient is presented in the supplementary information
to paper III, but for completeness of this method chapter, it is described below.

**THE CENTRAL SECTION THEOREM AND THE CONCEPT OF COMMON-LINES**

The basic theory of 3D reconstruction from projection images in electron microscopy
was developed by DeRosier and Klug (Derosier and Klug, 1968) and the theory was
later used to reconstruct the tomato bushy stunt virus (Crowther et al., 1970a). The
theory is based on the projection slice theorem (Bracewell, 1956), which states that the
Fourier transform of the projection of a 3D density distribution corresponds to a central
section through the 3D volume’s Fourier transform:

\[
\Im \{im(x, y)\} = \int \int \int o(x, y, z) \exp \{-2\pi i (xf_x + yf_y)\} dz dx dy
\]

\[
= \int \int o(x, y, z) \exp \{-2\pi i (xf_x + yf_y)\} dx dy dz
\] (2.9)
The z-direction is arbitrary and the Fourier transform of a 2D projection along a certain direction is thus identical to the Fourier transform of a plane through the origin, normal to the projection direction. Two projections will therefore share a common line, and their relative orientations will be fixed up to a rotation around the axis defined by this line. Three non-parallel projections will generate three common lines that unambiguously determine their relative orientations, except for the ambiguity of enantiomorphism, which cannot be resolved from independent projections alone.

**Normalized correlation coefficients**

The 2D normalized correlation coefficient may be used as a similarity measure between the image pairs, \( g(x,y) \) and \( f(x,y) \):

\[
\frac{\int \Re \{ F(\tilde{s}) G^*(\tilde{s}) \} d\tilde{s}}{\sqrt{\int |F(\tilde{s})|^2 d\tilde{s}} \sqrt{\int |G(\tilde{s})|^2 d\tilde{s}}}
\]

with \( \tilde{s} = (u,v) \) being the reciprocal lattice vector. For sampled transforms, the integration signs may be replaced by summations. If the Fourier transforms are sampled on a \( D \times D \) orthogonal lattice, with the lowest and highest frequencies of the Fourier components along the axes being \( 1/d \) and \( D/2d \) Å respectively, where \( d \) Å is the wavelength of the first-order Fourier component and \( d/D \) being the sampling distance in the images from which the Fourier transforms are calculated, a 2D normalized correlation coefficient is defined:

\[
C_{2d} = \frac{\sum_{k \in \Omega} \Re \{ F(\tilde{k}) G^*(\tilde{k}) \}}{\sqrt{\sum_{k \in \Omega} |F(\tilde{k})|^2 \sum_{k \in \Omega} |G(\tilde{k})|^2}}
\]

with \( \Omega : \{ n \in \mathbb{Z} ; |d/n| \in [r_1, r_2] \} \), where \( r_1 \) and \( r_2 \) are the high- and low-resolution limits (in Å). \( C_{2d} \) will thus be a real number \( \in [-1,1] \). If we are interested in judging the quality of the relative orientations between images, rather than their similarity, a joint common line correlation coefficient may instead be calculated. After a set of common line pairs
related to a set of Euler angles has been generated, the normalized joint common line
correlation coefficient may be calculated (Lindahl, 2001):

\[ C_{\text{line}}[\{l_{1i}, l_{2i}, \ldots, l_{Mi}\}] = \]

\[ \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{kl} \text{Re}\{F_{l_{1i}}(kl)F_{l_{2i}}^*(kl)\}}{\sum_{kl} |F_{l_{1i}}(kl)|^2 \sum_{kl} |F_{l_{2i}}(kl)|^2} \]

(2.12)

where \( M \) is the number of common line pairs, \( \{l_{1i}, l_{2i}\} = \{(x_1, y_1), (x_2, y_2)\} \),
\( |l_{1i}| = |l_{2i}| = 1 \) is the \( i \)th pair of common lines in the coordinate system of the Fourier
transformed images, \( \{l_{1i}, l_{2i}\} \) denotes the pair of 2D Fourier transforms to which they
apply, and \( F_{l_{1i}}(l) \) is the value in the point \( l \) of the Fourier transform \( f_{l_{1i}} \). \( C_{\text{line}} \) will thus
be a real number \( \in [-1, 1] \). The sampling points do not in general coincide with the
sampling points of the Fourier transforms of the projections, and for a spatially limited
object sampled on a Cartesian grid, the continuous Fourier transform is obtained from
the discrete one by the sinc function as a convolution kernel, as described in (Lindahl,
2001). Furthermore, formula (2.12) assumes that all projections have the same origin
and to account for an eventual origin shift, a phase shift of \((2\pi / D)(s_x, \varepsilon_x + s_y, \varepsilon_y)\) radians for the Fourier component \( F(\varepsilon_x, \varepsilon_y) \) has to be introduced for an origin shift of \((s_x, \varepsilon_y)\).

**AVERAGING**

In order not to burn the very sensitive vitrified biological specimen to ashes in the
electron microscope, the “principle of shared suffering” must be applied. By letting
each molecular image only be illuminated with a part of the dose required to create the
final image, the radiation damage is minimized and high resolution information is
preserved. In order to make class averages for use in reference-free alignment,
averaging over several projections in the same (or similar) view is performed. The
degree of over-sampling in a certain projection direction is a trade-off between the
extent to which one whish to improve the signal to noise ratio and the resolution
desired for the averaged projection. For averaging to be meaningful, the two
translational degrees of freedom and the in-plane rotational angle must be optimized
such that the 2D correlation between each image pair of the entire image set is
maximized.
The concept of 2D alignment as a property of the entire image set

In this thesis, an iterative method for reference-free 2D alignment that avoids selecting individual images as references is used (Penczek et al., 1992). By generalizing the alignment between two images to a set of $N$ images, the following definition is proposed: a set of $N$ images is aligned if all images are pair-wise aligned (Frank, 2006). In the Penczek method for reference-free 2D alignment, each image numbered $i$ is aligned to a partial average of all other images iteratively.

Principal component analysis

Each $D \times D$ pixels image of the single-particle data set can be viewed as a vector in a $D \times D$ dimensional space. All images together form an “image cloud” in this space. Data mining seeks to describe such multi-component data vectors in a simplified and comprehensible way. Principal Component Analysis (PCA) is a statistical tool for performing data mining or dimensionality reduction, and it has been extensively used in the single particle field. The mathematical formalism of PCA may for example be found in (Frank, 2006; Koeck et al., 1996; Lebart et al., 1984). PCA can be described as finding the directions of the maximum extension of the “image cloud” and reducing the dimensionality by describing each image as a linear combination of a set of orthogonal basis vectors, found by diagonalization of the variance matrix of the image set. This problem is an eigenvalue problem, and the orthogonal basis vectors are therefore referred to as eigenvectors. Reconstitution in factor space refers to the reconstitution of each image vector as a linear combination of the basis vectors. It has been shown that this can be achieved with 60 factors or less without loss of significant variations in the data set, see (Frank, 2006) and references therein. PCA reconstitution provides a clear geometric representation of the information, noise reduction and computational efficiency in algorithms that measure distances between image vectors, in order to judge their similarity.

Unsupervised classification in factor space

The information condensed by PCA to a lower dimensionality hyperspace may be used for clustering projections into classes of particle views projected in the same (or similar) orientation and originating from molecules in the same conformational or
morphological state. Two clustering techniques have been used in this thesis: k-means clustering, described for example in (Penczek et al., 1996), and hierarchical ascendant classification, described in for example in (Lebart et al., 1984). These algorithms are implemented in Spider (Frank et al., 1996) and described in detail in (Frank, 2006). After classification of a data set aligned in 2D into quasi-homogenous groups (classes), the dimensionality of the reference-free 3D alignment problem is reduced

\[ \{\psi_i, \theta_i\}_{i=1}^{#C} \]  

(2.13)

to 2\#C parameters for a conformationally/morphologically homogenous data set, where \#C is the number of classes, represented by their average. The first orientation search can then be performed in a discrete and evenly distributed angular space (paper I). This process can be viewed as points sliding on the surface of a sphere, with one point corresponding to the orientation of a 2D image. A natural objective cost function for an optimization involving this kind of search is the negative normalized joint common line correlation coefficient (Lindahl, 2001), which measures the quality of a given “point configuration”. For a heterogeneous data set

\[ \{\psi_i, \theta_i, s_i\}_{i=1}^{#C} \]  

(2.14)

with \(1 \leq s_i \leq #S\), where \#S is the number of conformational/morphological states present in the population, the dimensionality of the 3D alignment problem may only be reduced to 3\#C parameters, presupposed that the classification has resolved the heterogeneity. If the variations due to heterogeneity are not dependent upon the initial orientation assignment, a reference-free alignment scheme (Goncharov and Gelfand, 1988; Ogura and Sato, 2006; Penczek et al., 1996; van Heel, 1987) (paper I) may be used to find approximate orientations for each individual image in the data set, and the heterogeneity may be separated by classification of orientation directed classes (described below). For a heterogeneity that does not affect the initial orientation assignment the points must, apart from sliding on a spherical surface, jump between different layers of an “angular superspace” (illustrated in Fig. 2.1 and described formally below). For a heterogeneity that disrupts the initial orientation assignment, an extension of the angular space must be applied already in the first round of reference-free alignment, and the “point moving” optimization algorithm must be performed on the surfaces of as many spheres as conformations present in the population.
THE ORIENTATION SEARCH PROBLEM

The computational complexity of orientation search problems in cryo-EM has been the subject of a very interesting report from the department of computer science of the University of Helsinki, Finland (Mielikäinen et al., 2004). In this report it is stated that “…while several variants of the problem are NP-hard (Nondeterministic Polynomial-time hard), inapproximable and fixed-parameter intractable, some restrictions are polynomial-time approximable within a constant factor or even solvable in logarithmic space”. In practice, this means that for several variants of the problem exhaustive optimization procedures are not computationally feasible. The orientation search problem in single-particle electron microscopy is dominated by Expectation Maximization-type procedures of repeatedly finding the best reconstruction for a set of fixed orientations and the best orientations for a fixed model, see for example (Frank et al., 1996; Grigorieff, 2007; Lindahl, 2001; Ludtke et al., 1999; van Heel et al., 1996). Procedures of this kind require an initial reference reconstruction, which has traditionally been generated for example by using the Random Conical Tilt method (RCT) (Radermacher et al., 1987) or the method of angular reconstitution (van Heel, 1987). Because of the manual assignment of tilt pairs and the technical challenges involved in collecting high quality tilted cryo data, the RCT method and its related Orthogonal Conical Reconstruction (OCR) method (Leschziner and Nogales, 2006), are best performed by using negative stain specimen preparation (Brenner and Horne, 1959), which provides the contrast necessary for manual tilt-pair assignments and offers the ability of minimizing charging during tilted data collection via the double sandwich layer specimen preparation technique, see for example (Valentine et al., 1968). The negative stain preparation severely limits the resolution of a reconstruction and it is highly questionable if a low-symmetry negative stain 3D reconstruction in
Protein structure dynamics and interplay

general is capable of taking a cryo data set to its upper resolution limit. Attempts to refine cryo data sets of the Mg-chelatase enzyme (chapter III) and the L-Mediator complex (chapter IV), using negative stain reconstructions as initial references, to a resolution better than 20 Å failed (H. Elmlund and J. Lundqvist, unpublished observations).

The problem of reference-free alignment of projection images is NP-hard if the number of projections is equal to or larger than three (Mielikäinen et al., 2004). Several methods aimed at solving this very computationally intense problem have been developed. The method of angular reconstitution is based on common line correlation driven exhaustive orientation search for three class averages simultaneously. The alignment of the three, approximately noise free projections, is then used to align the complete set of class averages. Until the mid nineties, the method of angular reconstitution was the only true *ab initio* method for non-symmetric particles. In 1996 P. Penczek published a common lines based method (Penczek et al., 1996) for orienting several class averages simultaneously. The simultaneous minimization method presented by Penczek serves to maximize the joint common line correlation coefficient of the image set. The main advantage of an approach for simultaneous alignment of a large number of single particle class averages is that the risk of ending up in a false minimum due to an unlucky choice of projections is minimized. The Penczek method for reference-free 3D alignment has been used for a number of published 3D reconstructions, see for example (Azubel et al., 2004; Craighead et al., 2002), and it has been a source of inspiration in developing the method for reference-free alignment in a discrete angular space (RAD), presented here in paper I and summarized below. Recently, a simulated annealing based method for *ab initio* 3D reconstruction that utilizes a 2D correlation based cost function and the weighted backprojection reconstruction method for calculating the volume was published (Ogura and Sato, 2006). The 3D reconstructions generated by this approach look promising at low resolution, but the very computationally intense cost calculation requires the interpolation of a map at each iterative step. The common line formulation of the reference-free 3D alignment problem offers a more sensitive cost function and a greater flexibility to explore different alignment strategies. Furthermore, good class-averages have potentially more high resolution information than their resulting 3D reconstruction and staying longer with the class averages and applying common lines based reference-free alignment schemes iteratively, using orientations from
combinatorial optimization by simulated annealing in a discrete angular space as initial 
value configuration, has a profound effect on the resolution and interpretability of the 
initial map (paper I). The field of cryo-electron microscopy of icosahedral viruses has 
also recognized the efficient performance of simulated annealing optimization. A 
multi-path simulated annealing optimization algorithm has been used to resolve 
secondary structure elements in icosahedral virus reconstructions (Liu et al., 2007).

Simulated annealing

Simulated annealing (SA) represents a collection of stochastic algorithms that are 
generalizations of a Monte Carlo method for examining the equations of state and 
frozen states of n-body systems (Metropolis et al., 1953). SA algorithms have been 
used to solve many combinatorial optimization problems, like the travelling salesman 
problem. The stochastic SA algorithm, first proposed by (Kirkpatrick et al., 1983), is 
derived in analogy with a physical system. The melting of a substance at a very high 
temperature, followed by a slow cooling, may return the substance to crystalline state at 
a global free energy minimum. By simulating this process in numerical optimization 
algorithms, ergodic functions that are hard to treat with traditional methods can be 
minimized. To get a stable convergence of SA, the time spent on each temperature 
level should be sufficient for the system to reach a steady state (Rajasekaran, 2000). It 
has been shown that SA converges in the limit to a globally optimal solution with a 
probability of 1 (Mitra et al., 1986), but a time bound for convergence is not given. A 
true global minimization of a multidimensional ergodic or noisy cost function would 
require such a slow annealing rate that the computation time would correspond to the 
time required to solve the problem exhaustively, and one therefore has to be satisfied 
with the best possible solution achieved within a feasible computation time.

A generalized simulated annealing algorithm

Optimization by SA requires the definitions of the notions state, transition, 
temperature and cost. Let:

$$ X = \left( \begin{array}{c} x_{11} \cdots x_{1N} \\ \vdots \\ x_{M1} \cdots x_{MN} \end{array} \right) 
\quad x_{pq} \in \mathbb{Z}^+, \ 1 \leq x_{pq} \leq L, \ 1 \leq p \leq N, \ 1 \leq q \leq M $$

(2.15)
describe the state or solution of the combinatorial problem. The simulated annealing algorithm used in this thesis is written in object-oriented Fortran 95 and it is generalized in the sense that it accepts any arbitrary cost function and any number of parameters. A transition between two states is defined in different ways, depending on the nature of the combinatorial problem. If a row of $X$ may contain equal integers, a transition is described as the perturbation:

$$\left(x_{S1} \ldots x_{SN}\right) \rightarrow \left(x'_{S1} \ldots x'_{SN}\right)$$

(2.16)

where $x'_{S1} \ldots x'_{SN}$ denotes a series of integer random numbers not equal to $x_{S1} \ldots x_{SN}$ with $1 \leq x_{Sq} \leq L$. $S$ represents an incremental iteration variable $1 \leq S \leq M$. If a row of $X$ is not allowed to contain equal integers, the following requirement must be fulfilled:

$$x'_{Sm} \neq x'_{Sn} \text{ with } m \neq n \text{ and } 1 \leq m, n \leq N$$

(2.17)

The mapping $cost: X \rightarrow \mathbb{R}$ returns the objective cost value of a solution. The cost value should provide an estimate of the quality of a given state and the temperature is simply an unsigned control parameter in the same unit as the cost. The SA algorithm always accepts a true downhill transition; else the temperature controls the acceptance probability of a transition $state \rightarrow state^*$ according to:

$$P_{state \rightarrow state^*} = \exp\{-\frac{(cost^*-cost)}{kT}\}$$

(2.18)

where $cost^*-cost$ is the cost difference, $T$ is the temperature and $k$ is the Boltzmann factor, which only serves to scale the cost difference to fit the temperature interval used in the annealing. At very high temperatures the SA algorithm accepts essentially all transitions. By letting the SA algorithm stabilize at each temperature level and reach a steady state, followed by an annealing according to:

$$T_{s+1} = tT_s$$

(2.19)

where $t$ is a problem-specific temperature update constant $0.5 \leq t \leq 0.99$, a global minimization of a multidimensional ergodic cost function can be achieved (Rajasekaran, 2000).
Reference-free alignment in a discrete angular space (RAD)

A detailed description of the RAD-algorithm is found in paper I, but for completeness of this method chapter it is summarized here. Let:

\[ X = \{ x_i \}_{i=1}^{\#X} \]  
(2.20)

be the set of \#X class averages subjected to a RAD-simulation and let:

\[ E = \{ e_j \}_{j=1}^{\#E} \]  
(2.21)

be the set of \#E evenly distributed projection directions. \( L \), is a list of ordered pairs:

\[ L = \{ (x_i, e_j) \}_{k=1}^{N} \text{ with } N \leq \#X, \#E, \ 1 \leq i_k \leq \#X, \ 1 \leq j_k \leq \#E \]  
(2.22)

and there are no identical pairs for \( m \neq n \) and \( 1 \leq m, n \leq N \)

which defines a state. A transition between two states is defined as the perturbation:

\[ \{ (x_{ik}, e_{jk}) \}_{k} \rightarrow \{ (x_{ik}, e_{jk'}) \}_{k} \]  
(2.23)

with \( j' \) being an integer random number \( 1 \leq j' \leq \#E \) for one \( k \). Thus, a transition describes how one class average changes its associated orientation. A schematic overview of the RAD-algorithm is found in paper I. The negative normalized joint common line correlation coefficient, \( C_{\text{line}} \) (see above) is used as an objective cost function, calculated in a user controlled resolution interval using Strul (Lindahl, 2001) modules. To accomplish a transition, a permutation motor consisting of a state generator and an acceptor function is required. Initialization of the RAD state generator involves automatic selection of the \( N \) class averages chosen to be as dissimilar and highly populated as possible (paper I). A subset of \( N \) Euler angle triplets are randomly selected from \( E \) and used as initial orientations. From the resulting initial state an initial cost is calculated. The RAD-solution is perturbed by transitions between discrete angular configurations, in a similar fashion to the early annealing attempts on the ribosome (Penczek et al., 1996). The acceptor function always accepts a true downhill transition; else the transition probability for:

\[ \{ (x_{ik}, e_{jk}) \}_{k} \rightarrow \{ (x_{ik}, e_{jk'}) \}_{k} \]  
(2.24)

is given by:
where $\text{cost}^*-\text{cost}$ is the cost difference, $k=10^{-4}$ is the Bolzmann factor and $T$ is the temperature. The annealing rate is controlled by the temperature update function:

$$T_{s+1} = tT_s$$

and the choice of initial temperature. A high initial temperature combined with a careful annealing increases the chance of global minimization (Locatelli, 2000) and by varying the control parameters in several simulations optimal values are found, resulting here in a temperature update constant, $t$ of 0.9 and an initial temperature, $T$ of 1,000,000. The maximum number of rearrangements at each temperature level in RAD is limited to ten times the number of projection averages to align. The algorithm is not capable of handling the ambiguity of enantiomorphism and an additional experiment must be performed to determine the absolute hand (Rosenthal and Henderson, 2003).

**Disentangling conformational or morphological states**

An obstacle towards a complete understanding of the structural organization and dynamics of large macromolecular assemblies by single-particle cryo-EM is polydispersity of the preparations due to conformational or morphological variability. If not resolved, structural heterogeneity may limit the achievable resolution of a 3D reconstruction and mislead its biological interpretation. If resolved, the heterogeneity instead offers a tremendous biological insight to the dynamic behaviour of a molecule. There is an intimate connection between macromolecular dynamics and the structure-function relationship of biological systems. The fundamental process of transcription offers one example of highly dynamic machinery, where large conformational changes accompany the initiation process. Examples of dynamic transcription complexes are pol II, with its flexible clamp domain (Kostek et al., 2006), the jack-knife like conformational rearrangement of Mediator upon pol II association (paper IV), and the conformational breathing of TFIID (paper III). Yet another example of a highly dynamic cellular process is the translation of mRNA into proteins, which involves numerous RNA and protein molecules that bind to and dissociate from the ribosome. The 30S ribosomal subunit has for example been shown to rotate with respect to the 50S subunit in a rachet-like movement upon interaction and action of the elongation factor EF-G (Frank and Agrawal, 2000). Conformational heterogeneity thus occurs.
naturally in the macromolecular world and chemical processes that induce conformational rearrangements in populations of large macromolecular complexes are often rate limited by the mass transport, leading to a mixture of conformational states upon rapid freezing. Single-particle cryo-EM offers the possibility of visualizing these different states (Brink et al., 2004; Fu et al., 2007; Gao et al., 2004; Heymann et al., 2003; Heymann et al., 2004; Penczek et al., 2006; Scheres et al., 2007), although methods for studies of this kind are not yet completely explored. Most methods for 3D reconstruction from images of individual weak phase objects, such as images of unstained biological macromolecules, require a large number of images representing the same underlying 3D volume projected in different directions. If this condition does not hold due to structural heterogeneity, the angular space used in the alignment has to be extended to account for the additional variability in the images.

It has been shown that dynamic models, based on elastic network normal mode analysis (NMA), can be used to generate references for multireference refinement when the source of heterogeneity is conformational variability. NMA has been able to capture functionally relevant macromolecular motions of large assemblies, even without additional structural data (Brink et al., 2004; Chacon et al., 2003; Tama et al., 2002), but with atomic structures at hand, NMA and cryo-EM may describe functionally important motions at quasi-atomic resolution, as demonstrated by the dynamic reorganization of the functionally active ribosome (Tama et al., 2003). Under certain circumstances a complex case of structural heterogeneity may be resolved by simple methods that for example rely on density measurements in high variance regions for assigning particles separated by protocols for classification in factor space to different structures (Grob et al., 2006; Kostek et al., 2006; Penczek et al., 2006). The special case of a binary and locally discrete structural heterogeneity due to partial ligand occupancy has for example been extensively studied in the ribosome field (Gao et al., 2004; Penczek et al., 2006; Valle et al., 2002).

In a few cases, structural heterogeneity has been disentangled from a single heterogeneous population of randomly oriented macromolecules without a priori knowledge about the structural variability (Fu et al., 2007; Penczek et al., 2006; Scheres et al., 2007). Separation of conformational states without the use of any guiding structural template is a inherently difficult challenge. In paper II-III of this
thesis it is shown that such separations are indeed possible for low-symmetry particles, and the underlying theory is described below.

**Orientation directed classification and sub classification**

If, to a first approximation, the heterogeneity does not affect the initial orientation assignment, conformational or morphological states may be separated as follows. Let:

\[
X = \{x_i\}_{i=1,\#X}
\]  

be the heterogeneous set of \#X particles and let:

\[
O = \{o_i\}_{i=1,\#X}
\]  

be their assigned orientations. For a translationally aligned set of particles, an arbitrary orientation is described by a triplet of Euler angles and in-plane rotational alignment is achieved by rotating the set according to the azimuthal Euler angle. Let:

\[
S = \{s_i\}_{i=1,\#X}
\]

be the set of integer numbers that assigns each particle to a conformational state, with \(1 \leq s_i \leq \#S\), where \#S is the number of states present in the population. Let:

\[
E = \{e_j\}_{j=1,\#E}
\]

be a set of \#E evenly distributed Euler angles. An orientation directed classification is performed by sorting clusters of particle images on the basis of closeness (\(\leq 5^\circ\)) to the evenly distributed set of \#E projection directions by using the Euclidean norm, and \#E classes are formed:

\[
C = \left\{ \left\{ x_{j,k}, o_{j,k}, s_{j,k} \right\} \right\}_{j=1,\#E}^{1 \leq k \leq \#X, k_p \neq k_q \text{ for } p \neq q}
\]

\#Cj is the number of particles in class \(j\). \(l\) maps a particle with its defined orientation and unknown state to a class according to the orientation directed classification. Each class will be heterogeneous and contain \#S states. To resolve these states, each class is subjected to automated k-means classification (Penczek et al., 1996) to form \#S subclasses. If the heterogeneity is separable by the Euclidean norm, measured without
restraint to certain regions, this approach separates the \#S states and the problem left to solve is to assign the \#S projected conformers to the \#S 3D reconstructions.

**Assigning subclass averages to 3D reconstructions**

The task of assigning the \#S projected conformers to the \#S 3D reconstructions is equivalent of finding the \#S vertices connecting the proper set of \#E subclass averages, which forms the nodes of an \#S-directed graph (illustrated in Fig. 2.2 below).

![Illustration of the graph determination problem of assigning subclass averages to 3D reconstructions for a binary heterogeneity. The set of arrows correspond to a subset of evenly distributed orientations, with each arrow pointing on the two resolved conformational or morphological states in each orientation. Each subclass average forms a node in the bi-directed graph and the two arcs of the graph (red and blue) connect the set of subclass averages that correspond to one of the two 3D reconstructions.](image)

An assumption made from here on is that the orientation for each particle is equal to the orientation used for generating the class, giving:

\[
C = \left\{ \left\{ x_{j,i}, e_{j,i}, s_{j,i} \right\}_{j=1,\#E} \right\}_{i=1,\#S} \quad (2.32)
\]

The dimensionality of the problem is reduced by forming the subclass averages, resulting in:

\[
C = \left\{ \left\{ (x_{m}, e_{j,s}, s_{m}) \right\}_{m=1,\#S} \right\}_{j=1,\#E} \quad (2.33)
\]

where \#S <\#Cj and \(\bar{x}_{m}\) defines a subclass average. All possible configurations of:

\[
\left\{ \left\{ (e_{j}, s_{m}) \right\}_{m=1,\#S} \right\}_{j=1,\#E} \quad (2.34)
\]
form the “angular superspace”. The number of configurations in this space is \( \#S^{(nE-1)} \).

For the case of 116 evenly distributed orientations and two conformational states the “angular superspace” has \( 4.2 \times 10^{34} \) configurations, and a calculation aimed at finding the optimal solution by exhaustively calculating all non-redundant pair-wise common line correlation coefficients would require a computing time in the order of several years, even on the fastest CPU available, which stresses the need for an efficient combinatorial optimizer. In this thesis, common line correlation driven simulated annealing (SA) in the “angular superspace” is used to solve the graph determination problem of assigning each resolved conformer to a structure. SA optimization requires the definition of state, temperature and cost. \( C \) describes the state (formula (2.32)) and a transition between two states is defined as the perturbation:

\[
\{(\bar{x}_{m_j}, e_j, s_{m_j})\}_{m=1,8S} \rightarrow \{(\bar{x}_{m_j}, e_j, s'_{m_j})\}_{m=1,8S}
\]

with \( \{s'_{m_j}\}_{m=1,8S} \) being a series of integer random numbers \( 1 \leq s'_{m_j} \leq \#S \), \( s'_{p_j} \neq s'_{q_j} \), \( p \neq q \) for one \( j \). Thus, a transition describes how the subclass averages in one orientation change their associated conformational state. SA can be applied to \( C \) if an objective cost function \( \text{cost} : C \rightarrow \mathbb{R} \) is defined. The cost function provides an estimate of the quality of a given state and the temperature is simply an unsigned control parameter in the same unit as the cost. A schematic overview of the algorithm is presented in Fig. 2.3. The negative normalized joint common line correlation coefficient is used as an objective cost function, calculated in a defined resolution interval using Strul (Lindahl, 2001) modules. To accomplish a transition, a permutation motor consisting of a state generator and an acceptor function is required. Initialization of the permutation motor involves assignment of the series of random integer numbers:

\[
R = \{(r_{m_j})_{m=1,8S}\}_{j=1,8E} \quad 1 \leq r_{m_j} \leq \#S, \quad r_{p_j} \neq r_{q_j} \quad \text{for} \quad p \neq q
\]

as coordinates for the \( \#S \) directed graphs. From the resulting initial state:

\[
C_{\text{init}} = \{(\bar{x}_{m_j}, e_j, r_{m_j})\}_{m=1,8S}\}_{j=1,8E}
\]

an initial cost is calculated. The acceptor function always accepts a true downhill transition; else the transition probability for:

\[
\text{state} \rightarrow \text{state}^* \Leftrightarrow \{(\bar{x}_{m_j}, e_j, s_{m_j})\}_{m=1,8S} \rightarrow \{(\bar{x}_{m_j}, e_j, s'_{m_j})\}_{m=1,8S}
\]
is given by:

\[ P_{\text{state} \rightarrow \text{state}} = \exp\left\{-\frac{(\text{cost}^* - \text{cost})}{kT}\right\} \]  

(2.39)

where \( k=10^{-4} \) is the Boltzmann factor and \( T \) is the temperature. ‘steady’ and ‘frozen’ are Boolean variables (see Fig. 2.3 below).

‘frozen’ is set true if no change of the cost is observed between ten successive temperature changes and ‘steady’ is set true if the cost stabilizes during the rearrangement at a given temperature level, or the number of maximum rearrangements is reached. The principles underlying the simulated annealing optimization are otherwise the same as described above. To relax the initial assumption about a heterogeneity not affecting the initial orientation assignment, each of the \#S conformational groups of subclass averages is subjected to reference-free alignment over all five degrees of freedom using the orientations from the orientation directed classification as initial value configuration. Thus, each average is aligned against the
others iteratively, until the joint common line correlation coefficient in the given resolution interval converges. The $\#S$ 3D reconstructions are interpolated and used as references for refinement of the orientations of the individual images assigned to each group, before applying a supervised classification scheme (Gao et al., 2004), which serves to improve the separation and the resolution.

**Separating states directly from the class averages**

The method described here has not been used in any of the papers presented in this thesis, but for completeness of the theory for separating conformational or morphological states it is included. The method has generated promising results for a severely heterogeneous RNA polymerase II population containing substoichiometric amounts of the elongation factor spt4/5 (H. Elmlund, unpublished results). If the first approximation of an orientation assignment independency of the heterogeneity does not hold, the common line correlation driven simulated annealing must be performed in an “angular superspace” already in the first stage of reference-free alignment. Let:

$$X = \{\bar{x}_i\}_{i=1,\#X}$$ (2.40)

be the conformational/morphological heterogeneous set of $\#X$ 2D aligned class-averages subjected to reference-free alignment and let $ES$ be a space containing the set of $\#ES$ evenly distributed Euler angles, $\{e_j\}_{j=1,\#ES}$ and the set of $\#ES$ series of integer numbers, $\{s_{jk}\}_{k=1,\#S}$:

$$ES = \{(e_j, s_{jk})\}_{j=1,\#ES} \quad 1 \leq s_{jk} \leq \#S, \quad s_{jp} \neq s_{jq} \text{ for } p \neq q$$ (2.41)

where $\#S$ is the number of states present in the population. The list of ordered triplets:

$$L = \{\{\bar{x}_i, e_{ij}, s_{ij}\}\}_{i=1,\#X}$$ (2.42)

defines a state and a transition between two states is defined as the perturbation:

$$\{\bar{x}_i, e_{ij}, s_{ij}\}_i \rightarrow \{\bar{x}_i, e_{ij}, s_{ij}'\}_i$$ (2.43)

with $j'$ and $k'$ being integer random numbers $1 \leq j' \leq \#ES$ and $1 \leq k' \leq \#S$ for one $i$. Thus, a transition describes how one class average changes its associated Euler angle triplet and conformational state. As an objective cost function, the joint common line
correlation coefficient is used, just as described above. Initialization of the permutation motor involves random assignment of a subset of $\#X$ randomly selected Euler angle triplets withdrawn from $ES$ and assignment of a series of random integer numbers:

$$ R = \{r_i\}_{i=1,\#X} \quad 1 \leq r_i \leq \#S $$  \hspace{1cm} (2.44)

as an initial solution:

$$ L_{\text{init}} = \{(\bar{x}_i, e_i, r_i)\}_{i=1,\#X} $$  \hspace{1cm} (2.45)

from which an initial cost is calculated. The principles underlying the simulated annealing optimization are otherwise the same as described above.

**Common line correlation driven supervised classification**

The most commonly used approach for resolving heterogeneous single-particle populations is to use the principle of supervised classification, see for example (Gao et al., 2004), and the technique of projection matching (Frank et al., 1996). This approach requires 3D references representing the different states. Separation is achieved by comparing the experimental data with sets of evenly or quasi-evenly distributed projections of the different reference reconstructions and as an objective cost measurement the negative normalized 2D correlation coefficient may be used. Classification is based on the reference that gives the lowest cost. The situation may be complicated by the coexistence of many transition states. In cases where trustworthy start and end point states are known, a binary supervised classification scheme may be combined with histogram classification, where the differences formed between the two cost values assigned to each experimental image are plotted in a histogram and subpopulations are defined by histogram division (Gao et al., 2004). The dependency on *a priori* available information severely limits the general applicability of this approach, but when combined with the methods presented here, conformational states may be separated without the use of *a priori* structural information. In this thesis, the negative normalized common line correlation coefficient, calculated in Strul (Lindahl, 2001), is used as an objective cost measurement in the supervised classification scheme, which serves to improve the initial separation and resolution of the *ab initio* resolved templates.
RECONSTRUCTION OF A 3D VOLUME

A 3D reconstruction is calculated from the aligned sets of individual particle images or class averages. A number of ways to achieve this have been developed, deploying different mathematical formalisms, see for example (Grigorieff, 2007; Lanzavecchia et al., 1999; Penczek et al., 2004; Radermacher, 1992). In this thesis, reconstructions from aligned sets of class averages are calculated by using Fourier interpolation reconstruction, as implemented in Spider (Frank et al., 1996). From large sets of aligned individual particle transforms, reconstructions are calculated in VolRec - an algorithm developed by Martin Lindahl, which is described in detail in paper I.

VALIDATION

In (Frank, 2006) it is stated that “agreement between input projections and reprojections cannot be taken as an indication that the structure is correct, since they amount to little more than comparison between data placed into the 3D Fourier transform and data extracted at the same place.” The electron microscopic image is to a very good approximation a true projection, and the projection operator is the integral of the electrostatic potential distribution of the specimen over the direction of projection. The projection operator takes the entire structure into consideration and does not correspond to only partial information about the structure. If simply “placing” data in the 3D Fourier transform (random orientations) and comparing projections with reprojections, not much similarity will be observed. During the refinement of TFIIID in paper III, the heterogeneous reconstruction exhibited a rather poor similarity between projections and reprojections, whereas separation into two states substantially improved the similarity, which shows that the comparison is an important validity measure. In combination with the observation of a well-behaved Fourier shell correlation plot and a clear improvement of the resolution of the reconstruction in a biologically interpretable way (not scattered density) during the refinement cycles, a high degree of similarity between projections and reprojections offers a way to judge the quality of a reconstruction. Therefore, it is important to show refinement data, FSC-plot and comparison between projections and reprojections. The recognition of parts of a complex in independent reconstructions from different preparations of different binding states of a molecule, offers a convincing case of self-consistency, as demonstrated for example by the Mediator reconstructions presented in paper IV. The comparison
between parts of a reconstruction, and corresponding parts (subunits) solved with traditional methods like X-ray crystallography or NMR-spectroscopy, offers another self-consistency measure, as demonstrated in paper I, where the excellent fit of the atomic I-subunit model validates the novel method for \textit{ab initio} reconstruction. If the resolution does not allow for secondary structure visualization and no \textit{a priori} structural information is available, the attachment of a known, small and rigid structure with a conspicuous shape to the molecule under investigation may provide a means for validation. If the reconstruction of the labelled data set reproduces the correct structure that offers a strong self-consistency measure.

In this thesis, a fixed 0.5 Fourier shell correlation threshold value has been used to estimate the resolution of the presented 3D reconstructions. The resolution criterion in single-particle cryo-EM has been the subject of a very intense and lengthy debate that is beyond the scope of this chapter. Objective measurements of resolution in cryo-EM is hard, and perhaps the best way to estimate the resolution of a reconstruction is to simply look at it and see what kind of structural information it gives.

\textbf{FROM A NOVEL CRYO DATA SET TO AN INTERMEDIATELY RESOLVED RECONSTRUCTION}

The methods presented in this chapter can be utilized to take a cryo data set of a previously unknown macromolecular structure to an intermediately resolved reconstruction. With the use of distributed computing for performing refinement by using Strul (Lindahl, 2001) in combination with VolRec (written by M. Lindahl and presented in paper I), a resolution level that allows for the identification of secondary structure elements is achievable for high-resolution cryo data, without using \textit{a priori} structural information. This is demonstrated by the reconstructions presented in paper I-III. The methods may also be used to gain \textit{ab initio} dynamic information by studying heterogeneous populations of molecules, as demonstrated in paper II-III. A processing scheme for single-particle 3D reconstruction is presented in Fig. 2.4.
Fig. 2.4: Processing scheme for single-particle 3D reconstruction.
Magnesium chelatase (Mg-chelatase) is an ATP-dependent multiprotein enzyme that inserts Mg$^{2+}$ into protoporphyrin IX (PPIX) in the first committed step of the chlorophyll and bacteriochlorophyll biosynthetic pathways. In plants, the three Mg-chelatase subunits are designated ChlI, ChlD and ChlH, whereas in photosynthetic bacteria they are designated BchI, BchD and BchH. In this thesis, the three subunits are referred to as the I, D and H subunits. The bipartite, chaperone-like complex formed between I and D of Mg-chelatase was the first particle subjected to the methods presented in the previous chapter. During my thesis work cryo-EM data sets of the ID-complex induced with a variety of different nucleotide forms have been used to test the methods described in chapter 2. Mg-chelatase has played the role of our “favourite test-particle” and this chapter serves to illustrate what can be achieved with the methods presented in chapter 2 and the biochemistry of Mg-chelatase will be kept at a introductory level.

The chelation process and the Mg-chelatase subunits

Biochemical studies have shown that the chelation process can be divided into three stages: (1) formation of an activation complex between subunits I and D, which requires ATP and Mg$^{2+}$, (2) charging of the H subunit with PPIX to form an H-PPIX complex, and (3) interaction of the activation complex (Mg:ATP-I:D) with H-PPIX followed by Mg$^{2+}$ insertion, which requires hydrolysis of ATP (Al-Karadaghi et al., 1997; Jensen et al., 1998; Walker and Weinstein, 1991; Walker and Weinstein, 1994; Willows and Beale, 1998; Willows et al., 1996). This chapter is concerned with the structure of the AAA+ activation complex formed between subunits I and D (referred to as the ID-complex). The X-ray structure of the I-subunit (Fodje et al., 2001) demonstrated that it belongs to the AAA+ (ATPases Associated with various cellular Activities) superfamily, which is a family of ATPases with general structural and functional characteristics, described for example in (Iyer et al., 2004; Neuwald et al., 1999; Ogura and Wilkinson, 2001). A common function of AAA+ proteins is to assist cellular processes that require remodelling or modification of target substrates. The ATP-fueled motions are achieved by oligomeric AAA+ assemblies.
Fig. 3.1: Structure and sequence of the I subunit. (A) Dimerical apo-I (pdb: 1G8P) recovered from the crystal packing and with the ATP-binding motifs characteristic for AAA+ proteins highlighted in the 3D arrangement. (B) A generalized topology diagram for the I-subunit as an AAA+ module. Structural elements are numbered according to standard AAA+ conventions. The clade-characteristic Pre-sensor II insert is colored green. This insert results in a translocation of the four-helical bundle relative to the AAA+ core module. In (C) the alignment of BchI sequences from different organisms is shown. Conserved regions are highlighted, while motifs characteristic of AAA+ proteins are boxed.
Insights into the organization of oligomeric I may be obtained from analysis of the packing of the molecules within the crystal lattice, which demonstrates that the surface between two molecules forms an elongated ATP-binding groove, with conserved ATP-binding motifs in close proximity (Fig 3.1). In absence of nucleotide, the I subunit exists in a dynamic equilibrium between monomeric and dimeric states (Lundqvist J., personal communication), whereas in the presence of nucleotide (ADP, ATP or the non-hydrolysable ATP analogue AMPPNP) the I subunit forms hexameric ring structures (Willows et al., 2004). The N-terminal region of subunit D is homologous to I, implying that the D subunit also contains an AAA+ module. However, the D subunit lacks evolutionally conserved ATP-binding motifs and accordingly, no ATPase activity has been reported for the D protein (Jensen et al., 1999; Petersen et al., 1999). Despite its degenerate ATP-binding site, the D-subunit is capable of forming oligomeric ring structures, both in presence and absence of nucleotide, and it has been proposed to act as a stabilizing platform for complex formation with the I-subunit (Axelsson et al., 2006). The D-subunit is unique in the AAA+ superfamily in that it contains an AAA+ module, an integrin-I domain and a proline- and negatively charged residue-rich region in a single polypeptide unit. The structural interplay between these different protein modules in the ID-complex is discussed below.

**Fig. 3.2:** 3D reconstruction of the ADP induced ID-complex at 7.5 Å. (A) View perpendicular to the threefold symmetry axis of the reconstruction. The trimer of D dimers is shown in green and that of I is shown in red. View along the threefold symmetry axis (B) D-side and (C) I-side.
**Fig. 3.3: Quasi-atomic structure of the I trimers of dimers.** (A) EM density of the I hexamer truncated from the complete ID-complex reconstruction. (B) View along the threefold symmetry axis showing the solvent accessible surface of the I-ring. Only the interface between monomers within a homodimer may accommodate a nucleotide. The green atomic surface represents an ATP molecule docked based on the structure of HslU (pdb: 1DO2) and serves to mark the position of the potential nucleotide binding site. (C) Enlarged image of the nucleotide binding site with the ATP molecule in green and the ATP binding motifs characteristic for AAA+ proteins highlighted.
A quasi-atomic model of the ADP induced ID complex

Docking of the I-subunit X-ray structure and a homology-based model of the D-subunit to the 7.5 Å resolution cryo-EM reconstruction of the ID-complex, obtained in the presence of ADP (and referred to as ID$_{ADP}$), completed a quasi-atomic model of the assembly (Paper II). The ID$_{ADP}$ reconstruction is a bipartite, chaperone-like ring structure arranged in trimers of dimers (Fig 3.2). The model exhibits a high degree of structural similarity to other members of the AAA+ superfamily, see for example (Bochtler et al., 2000; Iyer et al., 2004; Pye et al., 2006; Zolkiewski, 2006). The ID$_{ADP}$ reconstruction shows that functional ATP binding sites may only be formed in the interfaces within each homodimer (Fig. 3.3), while the interfaces between I-dimers are stabilized by interactions to subunit D. The two D subunits within each homodimer of the ring structure exist in two distinctly different conformational states. One extended conformation (Fig. 3.4), with the integrin-I domains in contact with the inter-dimer regions of the I-ring, and one contracted conformation (Fig. 3.5) with the proline- and acidic residue-rich region contacting the nucleotide binding interfaces within each I-homodimer. The extended monomer forms an elongated structure that traverses the D-ring from the top, stretching towards the I-ring. Thus, the D-ring is built as a three-layered structure in which the AAA+ modules of the three extended monomers form the top of the ring. The second layer is formed by the proline- and acidic residue-rich domain of the extended monomer and the AAA+ and integrin-I domains of the contracted monomer. The third layer, contacting the I-ring, is formed by the integrin-I domains of the extended D monomers and the proline- and acidic residue-rich regions of the contracted monomers.
**Extended conformation**

**A** View along the threefold symmetry axis. (B) Side-view, with homology-based models of the D domains docked into the EM density. The N-terminal AAA+-like module of the three extended monomers of D is located on the top of the complex, while the integrin I domain is in contact with the I-ring. The AAA+-like module of subunit D is shown in blue (light blue for the AAA+ -like domain and dark blue for the helix bundle domain), the integrin I domain is shown in yellow, the contracted D subunit is shown in khaki and the I subunits are shown in red and orange. The suggested position of the proline- and acidic residue-rich region is highlighted by a circle. Schematic representations of the individual domains of subunit D is shown on the side of the Figure. The khaki-colored spheres with dashed lines represent the adjacent D-subunit, which adopts a contracted conformation (see Fig. 3.5).

Fig. 3.4: Extended conformation of the D subunit within the IDADP-complex. (A) View along the threefold symmetry axis. (B) Side-view, with homology-based models of the D domains docked into the EM density. The N-terminal AAA+-like module of the three extended monomers of D is located on the top of the complex, while the integrin I domain is in contact with the I-ring. The AAA+-like module of subunit D is shown in blue (light blue for the AAA+ -like domain and dark blue for the helix bundle domain), the integrin I domain is shown in yellow, the contracted D subunit is shown in khaki and the I subunits are shown in red and orange. The suggested position of the proline- and acidic residue-rich region is highlighted by a circle. Schematic representations of the individual domains of subunit D is shown on the side of the Figure. The khaki-colored spheres with dashed lines represent the adjacent D-subunit, which adopts a contracted conformation (see Fig. 3.5).
**Contracted conformation**

(A) View along the threefold symmetry axis. (B) Side-view, with homology-based models of the D domains docked into the EM density. The proline- and acidic residue-rich region (purple) of the contracted monomer is associated with the positively charged ATP-binding groove at the interface between two I monomers. This arrangement places the integrin I domain (yellow) close to the proline- and acidic residue-rich sequence of the adjacent extended D subunit. The supposed position of the proline- and acidic residue-rich region is highlighted by a purple circle. The khaki-colored spheres with dashed lines, in the schematic representations, show the adjacent D subunit, which adopts an extended conformation (see also Fig. 3.4). Subunit colors are as in Fig. 3.4.
The integrin-I domain participates in ligand binding through the MIDAS (Metal Ion Dependent Adhesion Site) motif, which is a highly conserved part of the D sequence. The MIDAS motif contacts the negatively charged end of the proline- and negatively charged residue-rich segment of the extended monomer. This interaction is reminiscent of that observed in the crystal structure of the complex formed between an integrin-I domain and a synthetic proline and glutamic acid residue-rich and collagen-like peptide (Fig. 3.6) (Emsley et al., 2000).

**Fig. 3.6: Protein-protein interactions that stabilizes the ID$_{ADP}$-complex.** (A) Position of the superhelical structure within the complex and a potential interaction with the integrin I domain of the contracted D subunit. The integrin-I domain from the extended D monomer associates with the interface between dimers of subunit I, close to the region where the known integrin I-binding RGE-motif is located in the four-helix bundle domain. (B) Side-view of the EM 3D-reconstruction visualized at higher sigma value reveals a super helical shaped structure in the proline-rich region of the extended monomer. (C) Structure of the complex between the integrin I domain and collagen (pdb: 1DZI). The collagen sequence is shown aligned with the sequence of a proline-rich region in subunit D.
**CONCLUDING REMARKS**

The study of Mg-chelatase has motivated much of the methodological work presented in chapter 2. The extraordinarily homogeneous ID\textsubscript{ADP} population was the first experimental data set to provide with robust validation of the power of the RAD algorithm – subsequent refinement on the starting reconstruction gave resolved secondary structure elements, already after five rounds of refinement. Since then, we have been able to generate subnanometer resolved reconstructions of two conformations of another particle from a single heterogeneous population by using the methods described in chapter 2. This particle is the general transcription factor IID and it will be described in the next and last chapter.
CHAPTER 4: THE EUKARYOTIC TRANSCRIPTION MACHINERY

The atomic structure of RNA polymerase II (pol II) was solved in the laboratory of R. Kornberg (Cramer et al., 2001). Pol II is the main target of transcriptional regulation and the molecule alone is capable of unwinding the DNA double helix, copy one strand into RNA, and finally re-unite the two DNA strands while translocating along the DNA. Pol II transcribes all protein-coding genes in eukaryotes and the molecule consists of a 10-polypeptide catalytic core together with the dissociable Rpb4/7 heterodimer. The largest subunit of pol II, Rpb1, has a highly variable C-terminus, which is referred to as the C-terminal domain or CTD. In cells of metazoans, fungi and green plants the CTD is composed of multiple repeats of seven amino acids Tyr1Ser2Pro3Thr4Ser5Pro6Ser7 (Corden et al., 1985; Meinhart et al., 2005). The sequence of these repeats is practically invariant within the eukaryotic kingdom, but the number of repeats varies between species. For example, the species \textit{S. cerevisiae} has 26 repeats, while the human CTD consist of 52 repeats. The CTD has an important regulatory role, and the establishment of a conserved heptad structure has been suggested to mark the evolution of a more complex machinery for the regulation of transcription (Guo and Stiller, 2004; Guo and Stiller, 2005). Initiation of transcription is thought to be the main point of gene regulation. Pol II requires a large number of factors that specifically recruit the enzyme to individual promoters within the genome.

THE GENERAL TRANSCRIPTION MACHINERY

Five general transcription factors (GTFs: TFIIB, -D, -E, -F, -H) assembles together with pol II and Mediator at every promoter before the initiation of transcription. This pre-initiation complex (PIC), with a total mass in excess of 2.5 MDa, recognizes promoter DNA, responds to regulatory signals and synthesizes the first few residues of the RNA transcript. The general transcription factors are responsible for promoter recognition and they are crucial for the formation of a transcription bubble and initiation of RNA synthesis (Conaway and Conaway, 1993). The DNA binding ability of human TFIID was first demonstrated with the Adenovirus major late promoter (AdMLP) (Sawadogo and Roeder, 1985). TFIID is composed of the TATA binding protein (TBP) and a number of conserved and pol II specific TBP-associated factors (TAFs) (Burley and Roeder, 1996). The TBP protein alone is sufficient for
transcription activity in a reconstituted transcription system (Oelgeschlager et al., 1998). Promoter DNA, TBP and monomeric TFIIB together form a complex that induces a sharp DNA bend (Nikolov et al., 1995), which is required for start-site recognition by pol II (Bushnell et al., 2004). The exact function of TFIID in the assembly of a start site recognition competent PIC has not been established, but it has been demonstrated that TFIID assist promoter recognition and PIC nucleation. TFIIF is normally associated with pol II and it is either recruited to the PIC together with the polymerase or directly afterwards, followed by the recruitment of TFIIE. Finally, the multimeric TFIIH complex is recruited, possibly through interactions with TFIIE, and the PIC is completed (illustrated in Fig 4.1 below). The TFIIH complex contains a kinase subunit, cyclin dependent kinase 7 (CDK7), which phosphorylates the fifth serine within each heptad repeat along the CTD (Cadena and Dahmus, 1987; Feaver et al., 1994; O'Brien et al., 1994). This hyperphosphorylation of the CTD is thought to be a molecular switch that causes funnelling of the pol II enzyme onto the start-site of the activated gene, to initiate mRNA synthesis.

Fig. 4.1: Sequential assembly of the pre-initiation complex.
The Mediator complex

The Mediator complex acts as an interface between gene specific regulatory proteins and the basal pol II transcription machinery. It functions as a key regulator of pol II dependent genes in *S. cerevisiae* (Myers and Kornberg, 2000). The stimulatory role of Mediator is demonstrated by depletion of human Mediator from nuclear extracts, which abolishes transcription by pol II (Mittler et al., 2001). The CTD of pol II has an important role for the Mediator function (Kim et al., 1994; Myers et al., 1998), and no fewer than nine SRB genes, encoding for Mediator subunits, were originally identified in a screen for mutants that suppress the cold-sensitive phenotype of a CTD truncation mutant (Nonet and Young, 1989). Purified *S. pombe* Mediator complex contains orthologues to components of the *S. cerevisiae* Mediator head and middle regions but lack orthologues to the proteins found in the *S. cerevisiae* tail region (Spahr et al., 2001) (paper IV). In *S. cerevisiae* and *S. pombe*, the Mediator complex interacts directly with the unphosphorylated CTD and forms a holoenzyme (Myers et al., 1998). The holoenzyme is transient and pol II dissociates from the complex as transcriptional elongation begins (Svejstrup et al., 1997). Evidently, Mediator has a role in stabilizing the catalytic core of the initiation complex by forming a molecular shell that covers almost half of the globular surface of pol II, see for example (Davis et al., 2002) and (paper IV). The recruitment of Mediator to a region nearby the promoter precedes the recruitment of pol II (Cosma et al., 2001) and re-initiation via repeated rounds of pol II binding to a stable Mediator scaffold complex has been observed in vitro (Yudkovsky et al., 2000).

Two forms of the yeast Mediator complex have been identified: the smaller S-Mediator, which has a stimulatory effect on basal transcription in vitro (Myers et al., 1998; Spahr et al., 2003), and the larger L-Mediator, which represses basal transcription (Holstege et al., 1998). The L-Mediator consists of an S-Mediator bound to the cyclin-dependent kinase 8 (CDK8) module (Samuelsen et al., 2003). The CDK8 module binds to the head and middle regions of the *S. pombe* Mediator and it has to be actively displaced from the L-Mediator complex before the interaction with pol II and formation of a holoenzyme (paper IV) (Fig. 4.2).
A model for stimulation of transcription by Mediator

The exact mechanism for the stimulatory effect of Mediator on transcription is not known, but it has been hypothesized that a productive interaction between Mediator and pol II leads to a faster entry of the polymerase to the promoter (Kornberg, 2005; Lewis and Reinberg, 2003). Mediator stimulates transcription in a purified pol II system, even in the absence of a transcriptional activator (Baek et al., 2006; Kim et al., 1994; Mittler et al., 2001). The stimulation of basal transcription by Mediator has been suggested to arise from the embracement of the catalytic core of the PIC (Chadick and Asturias, 2005), perhaps by promoting its formation or maintenance for multiple rounds of transcription (Reeves and Hahn, 2003; Yudkovsky et al., 2000). The molecular consequence of this embracement is not established, due to lack of sufficiently and isotropically resolved 3D data. It has been hypothesized that Mediator-induced conformational rearrangements of the PIC may hold the key to the mechanism for transcription stimulation (Kornberg, 2005). In prokaryotes, repressors prevent polymerase binding to the promoter, whereas activators contact the polymerase and increase its affinity for the promoter or stimulate the transition from a closed to an open polymerase-promoter complex. Perhaps Mediator makes the polymerase structurally ready to initiate transcription by inducing a conformational change in pol II. Further cryo-EM investigations of the dynamic behaviour of pol II upon Mediator binding is required to reveal the mechanism for stimulation of transcription by Mediator.
Fig. 4.2: Model for regulation of the Mediator structure. The model is based on the reconstructions of S-Mediator (yellow), holoenzyme (S-Mediator in complex with pol II (gray)) and L-Mediator (S-Mediator in complex with the repressive CDK8 module (red)) presented in paper IV. S-Mediator was segmented in two domains that were subsequently individually docked into the holoenzyme reconstruction. This act of docking indicates that Mediator has to undergo a jack-knife-like conformational rearrangement in order to embrace the globular polymerase density. It is possible that the compact conformation of Mediator corresponds to a transcriptionally inactive scaffold form and that the elongated conformation is present in an initiation competent PIC that may associate with the polymerase and initiate transcription.
Mediator also plays a key role in activated transcription by bridging DNA-bound activators and the general transcription machinery to the core promoter. Electron microscopic studies of human Mediator have revealed activator specific conformations (Naar et al., 2002; Taatjes et al., 2002; Taatjes et al., 2004). The temporal patterns of transcription initiation has been a source of debate in the transcription field and early investigations of Mediator (Kim et al., 1994; Koleske and Young, 1994) led to the hypothesis that Mediator and pol II may be recruited to the promoter in a single step, as a holoenzyme. However, several lines of evidence support the idea of Mediator recruitment preceding the association of pol II to active PIC:s (Bhoite et al., 2001; Bryant and Ptashne, 2003; Cosma et al., 2001; Kuras et al., 2003), and re-initiation via repeated rounds of pol II binding to a stable scaffold complex, containing Mediator, has also been observed in vitro (Reeves and Hahn, 2003; Yudkovsky et al., 2000).

Genome wide occupancy profiles of Mediator and the repressive CDK8-module has been reported independently from two groups (Andrau et al., 2006; Zhu et al., 2006), and the two complexes appear to exhibit very similar binding patterns. Perhaps the transcriptionally inert L-Mediator, containing the CDK8-module, exists in a dynamic equilibrium with the active S-Mediator form in the scaffold complex. Pol II and the CDK8-module are evidently competing substrates to Mediator (Paper IV) and perhaps one effect of activators is to alter the S-/L-Mediator equilibrium by influencing the dissociation constant of the CDK8-module in the L-Mediator complex or in other ways prevent the formation of an L-Mediator complex. This model (illustrated in Fig 4.2) does not exclude a substantial conformational rearrangement of Mediator upon activator binding. Perhaps the dissociation of the CDK8 module allows for the jack-knife-like conformational rearrangement of Mediator (paper IV) into an open state, ready to interact with the polymerase. The dissociation constant of CDK8 may also be a function of promoter architecture and CDK8 repression has indeed been suggested to be regulated in a promoter-specific manner (Andrau et al., 2006). Fig. 4.3 summarizes the current knowledge about Mediator in the initiation process.
Fig. 4.3: Mediator at initiation. 

a) Recruitment is a part of the activation and occurs at a site upstream from the start-site. Activators and repressors recruit Mediator to specific promoters. Perhaps activators and repressors also influence the dissociation constant of the CDK8 module. 

b) After dissociation of the CDK8 module, the active S-Mediator complex may interact with pol II and form a holoenzyme. A cardinal activity of Mediator is to stimulate phosphorylation of the CTD by transcription factor H. The cyclin dependent kinase of TFIH phosphorylates the CTD. 

c) Mediator is unable to bind to the phosphorylated form of CTD, which leads to dissociation of pol II from the initiation complex. Mediator may remain at the promoter in a scaffold complex together with other components of the general machinery.
The multiprotein TFIID complex is composed of the TATA box binding protein (TBP) and a number of conserved and pol II-specific TBP-associated factors (TAFs) (Burley and Roeder, 1996). In yeast, 13 of the TAFs are essential for cell viability (Yatherajam et al., 2003), which indicates that they play critical and non-redundant roles in the regulation of transcription. The TAFs are not restricted to TFIID, but can also be found in other regulatory complexes, for example the yeast SAGA complex (Grant et al., 1998) and the mammalian PCAF complex (Ogrzyko et al., 1998). The majority of genes in yeast are TFIID dependent (Lee et al., 2000) and many lines of evidence are pointing at the importance of TAFs (Chalkley and Verrijzer, 1999; Oelgeschlager et al., 1996; Sawadogo and Roeder, 1985; Shao et al., 2005) and TBP (Kim et al., 1993; Nikolov et al., 1995) in promoter recognition. The role of individual TAFs in general pol II transcription is however still unclear.

Structural architecture of TFIID

TAF interactions to TBP have only been reported for TAF1 (Liu et al., 1998) and TAF7 (Yatherajam et al., 2003). TBP does not seem to have a crucial role for the integrity of the TFIID complex. Electron microscopic studies of TFIID in negative stain have revealed a horseshoe shaped structure (Andel et al., 1999; Brand et al., 1999; Leurent et al., 2002; Leurent et al., 2004) (paper III), and a conformational breathing in which the two main quaternary domains move away from each other has been reported in a cryo-EM investigation of human TFIID (Grob et al., 2006). The S. pombe TFIID complex also undergoes massive rearrangements. The results presented in paper III of this thesis demonstrates that the two resolved TFIID states correspond not only to different conformations, but also to different morphological states, with one state containing TBP and one state lacking TBP. TFIIA and TFIIIB have been mapped to the each of the two different main quaternary domains that elongates from the central core (Andel et al., 1999), whereas TBP binds in the intersection between these domains (Andel et al., 1999) (paper III). The mapping of TBP in proximity of the central cavity of the structure has led to the suggestion that DNA also binds in the central pore, but structural evidence for this interaction has not been provided. Chromatin immunoprecipitation (ChIP) and DNA crosslinking assays has demonstrated that TBP loading onto the promoter is an important activation step in vivo (Kuras et al., 2000; Li
et al., 1999). Transcriptional activity in yeast strongly correlates with promoter occupancy by general factors such as TBP, TFIIA, and TFIIB, but not with occupancy by TAFs (Kuras et al., 2000).

The *S. cerevisiae* TFIID complex contains two identical copies of scTAF5 (Leurent et al., 2004; Sanders et al., 2002), whereas the *S. pombe* TFIID contains two distinct WD-repeat TAFs. One is TFIID specific (spTAF5) and one (spTAF5b) is shared with the Spf-Ada-Gcn5 acetyl transferase SAGA-like complex (Mitsuzawa and Ishihama, 2004). The importance of the WD repeat domains for the TFIID architecture is illuminated by the interactions reported between the H4-like *S. pombe* TAF6 and the C-terminal WD-repeat containing region of spTAF5b (Mitsuzawa and Ishihama, 2002). X-ray crystallography has provided the detailed structure of hTAF5 (Bhattacharya et al., 2007), and the ability of hTAF5 to form a flexible, extended dimer has been suggested to be of importance for the assembly of human TFIID. Many of the TAFs have histone fold domains (HFD:s) and the interactions between HFD TAFs are believed to be important for the stability of TFIID. A nucleosome-like octameric assembly of HFD TAFs (TAF9-TAF6-TAF12-TAF4)\textsubscript{2} has been proposed to be part of the structure (Selleck et al., 2001), and the crystal structure of the hTAF4-hTAF12 complex (Werten et al., 2002) support this arrangement, rather than the homodimerization suggested by earlier studies (Hoffmann et al., 1996). The EM reconstructions of TFIID (Andel et al., 1999; Brand et al., 1999; Leurent et al., 2002; Leurent et al., 2004) (paper III) contradicts that the structure is able to embrace a compact HFD TAF arrangement similar to the histone octamer.

**Histone targeting and modification**

TAF1 is a bipartite protein kinase (Dikstein et al., 1996) with multiple functions within TFIID. It possesses histone acetyl transferase (HAT) activity (Mizzen et al., 1996) and it has been shown to activate transcription by phosphorylation of serine 33 in histone H2B (Maile et al., 2004). TAF7 interacts with TAF1 (Gegonne et al., 2006) and thereby inhibits its intrinsic acetyl transferase activity. Upon PIC assembly TAF7 has been reported to dissociate from the TFIID complex. A function attributed to hTAF1 is the selective binding to multiply acetylated histone H4 peptides via the two tandem bromodomain modules (Jacobson et al., 2000), which suggests that TFIID may be targeted to specific chromatin bound promoters and play a role in chromatin
recognition. In yeast, TAF1 is smaller than in the *Drosophila* or human systems due to lack of the two tandem bromodomain modules and a C-terminal acidic domain (Matangkasombut et al., 2004). A candidate for this missing piece is the bromodomain factor 1 (Bdf1), which has been shown to associate with yeast TFIID (Matangkasombut et al., 2000).

**Promoter recognition**

Recognition of the core promoter by TFIID is coordinated with the assembly of an active PIC (Hahn et al., 1989; Horikoshi et al., 1989; Peterson et al., 1990). Sequence elements found in core promoters of pol II transcribing genes that are recognized by TFIID include the TATA element, the initiator element (Inr) and the downstream promoter element (DPE) (Smale, 2001). The recognition element composition is variable among promoters and no one element is essential for promoter function. Altogether the promoter structure serves to orient the transcription machinery to support unidirectional transcription. Synergies between TAFs contribute to the DNA specificity of TFIID. The HFD TAFs may play a role in the recognition process, since HFD mediated interactions in the TAF6-TAF9 and TAF4-TAF12 pairs have been shown to enhance the DNA binding activity and stimulate protein sequence specific binding to DNA by TAF6 and TAF9 (Shao et al., 2005). Another example of a synergy between TAFs is the TAF1-TAF2 complex. The individual TAF1 and TAF2 subunits are incapable of specific DNA recognition, but together with TBP the TAF1-TAF2 complex binds selectively to sequences that match the Initiator (Inr) consensus at the appropriate distance from the TATA element (Chalkley and Verrijzer, 1999).

A dramatic bend of the DNA is required for a stable TBP interaction, which makes the formation of a TBP-DNA complex intrinsically slow. The extent to which DNA is bent has been reported to correlate with the stability of the TBP-DNA complex (Starr et al., 1995). It is likely that this step is the rate-limiting step in the initiation process (Hoopes et al., 1998). Early investigations raised the possibility of a nucleosome-like DNA wrapping by TFIID (Hoffmann et al., 1996; Oelgeschlager et al., 1996; Xie et al., 1996). The following EM reconstructions of TFIID reported a compact nucleosome-like HFD TAF arrangement to be unlikely (Andel et al., 1999; Brand et al., 1999; Leurent et al., 2002; Leurent et al., 2004). The amino acid residues responsible for the DNA interactions in the nucleosome structure (Luger et al., 1997) are not conserved in
the TAFs. It is possible that the high degree of conservation between certain TAFs and histone proteins is related to another function than binding of DNA. However, TFIID has been shown to bind DNA both specifically and non-specifically and a TFIID-induced promoter kink would provide a more productive interaction to TBP regardless if the promoter contains a TATA sequence or if the DNA interaction is not specifically dependent upon TATA box. Regardless of the exact mechanism, a stable distortion of the DNA is required for wrapping of the promoter around TFIIB and pol II, to provide the architecture necessary for start-site recognition (Bushnell et al., 2004).

In the solution structure of the complex formed between TBP and the N-terminal region of TAF1 (Liu et al., 1998), TAF1_{11-77} is mimicking the minor groove surface of the partially unwounded TATA-box, which provides the structural basis to the mechanism by which TAF1 negatively controls the TATA-box binding activity within *Drosophila* TFIID. This regulatory pathway may explain how different promoter architectures discriminate between TFIID and SAGA, since TATA-containing promoters tend to be TFIID-independent and instead prefer to load TBP via the SAGA-complex pathway (Huisinga and Pugh, 2004).

**Activated transcription**

In the mid nineties, *in vitro* interactions between activators and isolated TAFs were reported (Burley and Roeder, 1996; Goodrich and Tjian, 1994; Tjian and Maniatis, 1994), which lead to the suggestion of TAFs being direct and obligatory activator targets, thought to communicate regulatory signals to the transcription machinery. This suggestion was reassessed when activated transcription was reported not to be generally dependent on TAFs (Apone et al., 1996; Moqtaderi et al., 1996), and the recognition and selection of core promoters by TAF1 was shown not to involve upstream activators (Shen and Green, 1997). Although activated transcription is not generally dependent of TAFs in the yeast system, activator specific recruitment of TFIID is important for the regulation of ribosomal protein genes (Mencia et al., 2002) and the derepression of DNA damage regulated genes (Li and Reese, 2000). *In vivo*, the Mediator complex is the main target of activators (Myers and Kornberg, 2000). Mediator is of central importance also for basal transcription (Myers et al., 1998; Spahr et al., 2003), and it fulfils the definition of both a co-activator and a general transcription factor (Takagi and Kornberg, 2006).
TBP interacts with remarkably few TAFs within the TFIID complex, but interactions to TAF7 have been reported (Yatherajam et al., 2003). TAF7 has also been reported to interact with TAF1 (Gegonne et al., 2006; Yatherajam et al., 2003), and a three-way interaction between TAF1, TAF7 and TBP may occur. The observed dissociation of TAF7 upon PIC formation (Gegonne et al., 2006) also suggests that conformational rearrangements in this region occur, possibly to aid loading of TBP onto the promoter. TFIID protects the DNA-binding surface of TBP via interactions with the N-terminal region of TAF1 (Liu et al., 1998) (paper III). This autoinhibitory domain acts as an activation domain when fused to the GAL4 activator (Kotani et al., 2000). TAF1 has also been reported to compete with the VP16 activator in binding to the TATA binding domain of TBP (Nishikawa et al., 1997). This regulatory pathway appears to be important and apart from promoting TBP-DNA binding, TFIIA also promotes binding of TFIID to DNA through this autoinhibition mechanism in an activator dependent manner (Chi et al., 1995; Kokubo et al., 1998). Altogether these observations suggest a plausible role of transcriptional activators as antirepressors of the autoinhibition exhibited by TAF1.

**Structure of TFIID (paper III)**

TFIID is a dynamic molecule that predominantly exists in two conformational states, with and without TBP bound (paper III). The conformational transition is coupled to TBP association and dissociation, which involves a rearrangement of the HFD TAF assembly. The HFD TAF assembly of TFIID is clearly different from that of the compact nucleosome core, and represents a novel HFD arrangement that may provide the plasticity required for allosterically controlled TFIID movements. The high degree of conservation between especially histone H4 and TAF6, but also between other histone proteins and HFD TAFs, suggest a functional interconnection between TFIID and the nucleosome. As discussed above, the DNA binding function is unlikely to be conserved and perhaps it is the conformational plasticity and ability to efficiently rearrange that explains the high degree of conservation. Eukaryotic transcription initiation requires reverse of the repressive chromatin structure to make the promoter accessible. The investigation presented in paper III of this thesis offers an example of how the assembly of multiple protein sequences provides mean for regulating the structure and function of complex structures by binding of relatively small molecules.
The study also points at the importance of methods for resolving heterogeneous population of molecules, to gain dynamic structural information and to overcome one of the major barriers towards reaching high resolution using single-particle cryo-EM – structural heterogeneity.

**DNA binding to TFIID requires removal of TBP**

TBP exists in at least two forms in vivo; one TAF containing form corresponding to the TFIID complex and one form lacking the TAFs (Kuras et al., 2000). TAF complexes lacking TBP have been identified in both yeast (Sanders et al., 2002) and human (Wieczorek et al., 1998) systems. It is therefore not surprising that the two resolved conformational states of *S. pombe* TFIID presented in paper III correspond to a TBP-bound and a TBP-lacking state. An elongated promoter in complex with TBP was modelled based on the X-ray structure (Kim et al., 1993) and docked to the reconstruction according to the position of TBP. The model fits nicely to the reconstruction, with promoter DNA residing in a groove along the head, which elongates from the HFD arrangement of the back and towards the TBP binding site. In order to prove this model, several attempts to reconstruct a TFIID-TBP-DNA complex were made. TBP was naively thought to be a major determinant of DNA association to TFIID. Excess recombinant TBP and promoter DNA was therefore added to the TFIID preparation and the TBP containing structure was used as a template for the alignment, but no extra density could be identified in the reconstruction. The assumption of TBP being a major determinant of DNA association to TFIID was wrong. When performing a supervised classification on the DNA-bound popoulation, using the TFIID and TFIID\(\Delta\)TBP reconstructions as templates, a large fraction of particles mapped to the TFIID\(\Delta\)TBP state. The bent DNA structure appeared in the reconstruction calculated from particles mapped to TFIID\(\Delta\)TBP and no DNA binding was observed for the TBP-containing structure (Elmlund H., work in progress). These preliminary results recieve support from the previously shown autoinhibitory role of the N-terminus of TAF1 (Burley and Roeder, 1998; Liu et al., 1998; Mal et al., 2004), which apparently prevents DNA binding to TBP-associated TFIID. The inability of the TBP-containing complex to bind DNA also fits well with the docking presented for TBP (paper III), where TBP closes the entry to the DNA-binding groove of the head, with its DNA-binding surface pointing towards the core of the structure. Moreover, these results are
in agreement with a stepwise assembly model that starts with TFIID\(\Delta\)TBP associating to promoter DNA and inducing the bent DNA structure. Subsequent binding of TBP results in a TFIID-promoter complex. TBP and the histone-fold TAFs are located in opposite sides of the TFIID complex and according to the promoter binding model presented here (paper III), promoter DNA binding to TFIID is not critically dependent on the histone-fold TAFs. It is possible that association of a longer DNA segment to the structure may include interactions also to the HFD TAFs. Another possibility is that DNA binding to the HFD TAFs has a different role than aiding loading of TBP onto the promoter. Perhaps ligand interactions to the HFD TAFs can induce the conformational rearrangement required for TBP-release. This would be of high functional relevance, since DNA binding to TFIID requires the removal of TBP.

Fig. 4.4: Promoter binding by TFIID. Reconstruction of a TFIID\(\Delta\)TBP-promoter complex with the DNA (red) colored based on difference mapping. No DNA binding was observed to the TBP-containing complex.

TFIID as “mediator of the scaffold complex”?

Mediator is responsible for orchestrating the molecular signals required for the formation of an active PIC that may associate with the polymerase and funnel it onto the start-site of an activated gene. Recent findings show that Mediator, TFIID, TFIIA, TFIIIE, TFIIG and TFIIH, but not TFIIB or TFIIF, can form a scaffold complex that lacks the polymerase and remains at the promoter after initiation (Yudkovsky et al., 2000). The transition from an inactive scaffold into an initiation competent PIC is a
most likely a crucial regulatory event, since it determines at which rate pol II is reloaded onto an activated gene. According to the current knowledge about the structure of the PIC (see (Bushnell et al., 2004; Chen et al., 2007) and references therein) and the TFIIDΔTBP-DNA reconstruction presented in Fig. 4.4, the TAFs occupy the space where pol II associates with promoter DNA in the PIC. It is therefore tempting to speculate that the TAFs must leave the promoter, at least in part, before entry of pol II into the PIC.

The TATA consensus does not appear to be a major determinant of TBP binding in yeast (Kim and Iyer, 2004). Only approximately 20% of the genes in yeast contain a TATA box, and these genes are associated with responses to stress, they are highly regulated, and they preferentially utilize SAGA before TFIID to load TBP (Basehoar et al., 2004). One possibility is therefore that TFIIDΔTBP plays an especially important role in DNA bending and TBP-loading at promoters lacking a TATA consensus. Another possibility is that of transcription without TBP, which would require stabilization of the bent promoter structure by general factors other than TBP to accomplish initiation. The latter possibility agrees with what has been observed for the TBP-lacking human TAF complex, TFTC (Wieczorek et al., 1998). I see no reason to exclude any of the two pathways (Fig. 4.5).

The TAFs are involved both in activation and repression of transcription (Burley and Roeder, 1998; Guermah et al., 1998; Guermah et al., 2001). They inhibit TBP-mediated basal transcription in the absence of activator and they stimulate activated transcription in synergy with the Mediator complex. This thesis provides a part of structural basis for the TAF duality, even if much more work needs to be done. The TBP-less TAF complex binds and bends promoter DNA. No DNA binding is observed to the TBP-containing TFIID complex. Moreover, the identification and docking of TBP within TFIID (paper III) provides structural evidence for the inhibition of promoter-TAF and promoter-TBP interactions when TBP is bound to the TAFs.
Fig. 4.5: A model for TFIIID function based on our current structural knowledge about TFIIID. TFIIID (brown) exist in a dynamic equilibrium between TBP-bound and TBP-lacking states and this equilibrium may be affected by increasing the relative molar amount of TBP (yellow). This observation is consistent with a dynamic association of TBP to TFIIID. Promoter (red) binding occurs only to the TBP-lacking state, which is in agreement with the previously suggested autoinhibitory role of the N-terminus of TAFI and docking of TBP to our TFIIID reconstruction. TFIIIDΔTBP induces the bent promoter structure required for initiation and in TBP-dependent transcription (a) it may function by presenting a protein-DNA surface for TBP to interact with. At promoters that are not dependent upon TBP (b) general factors other than TBP must stabilize the promoter kink.
Concluding remarks

Eukaryotic transcription initiation contains many chemical events that may be targeted for regulation. Before interaction with pol II, Mediator has to undergo a jack-knife-like conformational rearrangement and unfold to be able to embrace globular polymerase density. Moreover, polymerase association with Mediator requires the removal of the repressive CDK8 module (paper IV).

The conformational behaviour of TFIID (paper III) may also have a central regulatory role. It is possible that the release of TBP from the TFIID structure rather represents the release of TAFs from DNA, leaving a set of general factors bound to promoter DNA for Mediator to arrange into an initiation competent PIC. How this is achieved at a molecular level and how general factors and/or activators affect the conformational behaviour of TFIID are questions that remains to be answered, but the fact that TBP-binding is coupled to a rearrangement of the TFIID structure is an important observation and a cornerstone for further studies of higher order initiation complexes.

The mechanism for unzipping the DNA double helix and the transport of a single DNA strand into the active site of pol II is an enigma in both bacterial and eukaryotic transcription. Only carefully designed biochemical experiments combined with structural biology investigations will provide the answers. Despite the exact mechanism for promoter escape it is likely that the polymerization of the first few nucleotides is coupled to additional conformational rearrangements in the PIC. A more detailed description of how Mediator affects the conformation of pol II, how promoter escape is achieved, and how that affects the other factors involved, have to await cryo-EM reconstructions to higher resolution and of larger portions of the PIC. Large structural rearrangements undoubtedly accompany transcription initiation and cryo-EM in combination with single-particle reconstruction from heterogeneous populations of molecules has an important role in revealing the structure and dynamics of this sophisticated regulatory machinery.
CHAPTER 5: SUMMARY AND PERSPECTIVES

PAPER I

A new cryo-EM single-particle *ab initio* reconstruction method for structural characterization of multiprotein complexes to subnanometer resolution is described. Major results are:

- The RAD algorithm – Reference-free Alignment in a Discrete angular space
- Algorithm for sorting particle images according to a “free” parameter, based on the spectral common line correlation
- The VolRec algorithm – Volume reconstruction based on inverted interpolation in Fourier space
- The first structural evidence of a complex formed between BchI and BchD of Mg-chelatase

**Main conclusion:** The newly implemented methods in combination with common line correlation based refinement over all five degrees of freedom in Strul (Lindahl, 2001) yields reconstructions to subnanometer resolution, directly from large homogeneous sets of vitrified, unstained and randomly oriented low-symmetry single-particle images.

**My contributions:** Design of research (together with ML), specimen preparation for electron microscopy (together with JL), data collection, development of RAD and method for sorting particle images, analysis of results (together with JL and ML), writing of paper (together with ML, JL and SA).

PAPER II

Structural characterization of components in the *R. capsulatus* Mg-chelatase reaction cycle, using the new image processing methods described in paper I and III, is presented. Major results are:

- Quasi-atomic model of the ADP induced complex formed between BchI and BchD
- Four conformational states resolved from the ATP induced population of the complex formed between BchI and BchD
Main conclusion: An ATP-fueled and Integrin-I mediated conformational transition of the *R. capsulatus* Mg-chelatase AAA+ activation complex is required for holoenzyme assembly.

My contributions: Design of research (together with JL, SA and MH), specimen preparation for electron microscopy (together with JL), data collection (together with DK), image processing (together with JL and ML), analysis of results (together with JL, ML and SA), writing of paper (together with JL and SA).

**PAPER III**

Structural characterization of the general transcription factor IID from *S. pombe* using the method presented in paper I and a novel method for resolving heterogeneous populations of vitrified, unstained and randomly oriented single-particle images is presented. Major results are:

- Two conformational/morphological states of TFIID to subnanometer resolution
- Immunolabelling of TAF4
- TBP affects the equilibrium between the two TFIID states

Main conclusion: TBP-binding is coupled to a rearrangement of the HFD structure of TFIID.

My contributions: Design of research (together with CMG), specimen preparation for electron microscopy, data collection, image processing, development of method for resolving heterogeneous single-particle populations, analysis of results, writing of paper (together with CMG).
PAPER IV

Low resolution structural characterization of three forms of the *S. pombe* Mediator by negative stain single-particle electron microscopy and existing image processing methods are presented. Major results are:

- Reconstruction of the core Mediator (S-Mediator)
- Reconstruction of the core Mediator in complex with the repressive Cyclin C dependent Kinase 8 (CDK8) module (L-Mediator)
- Reconstruction of the core Mediator in complex with RNA polymerase II (holoenzyme)

**Major conclusion:** The CDK8 module sterically blocks RNA polymerase II interactions with Mediator.

**My contributions:** Design of research (together with CMG), specimen preparation for electron microscopy, data collection, image processing (together with ML), analysis of results (together with CMG and ML), writing of paper (together with CMG, PK, ML, HH).

**FUTURE PERSPECTIVES**

Cryo-EM is and will be an essential tool for studying large macromolecular complexes, but further technical evolution is necessary. The methods presented in this thesis offers an easily expandable platform of software in which future extensions may include approaches for fully automatic initial reconstruction generation and refinement from heterogeneous single-particle populations, approaches for extending the resolution and studying small or membrane associated macromolecular structures. It is my belief that such approaches will play an essential role in the technical development of cryo-EM, and provide a platform for high-resolution electron microscopy studies.
ACKNOWLEDGEMENTS

Thank you: Martin Lindahl, for letting me work with the excellently formulated Strul code and follow up the “common line heritage”, dating back to Klug, Crowther and van Heel, among others. I would also like to thank you Martin for your sound scepticism against both my wild ideas and existing single-particle methods, and for your ability to formulate methodological concepts and scientific texts stringently and concisely, which has been a source of inspiration. Your supervision has driven my work forward in the best possible direction. Claes M. Gustafsson, for being my academic mentor and always supporting me in tough times, scientifically and financially. Your positive energy, sharp intellect, well organized group and impressive scientific network over the globe has made our common scientific efforts a true pleasure. Hans Hebert, for giving me the freedom and trust I needed to perform science in a very untraditional environment, and for always having complete control over all technical equipment. Without your technical skills, and without the impressive machine park at Novum, this thesis would not exist. Philip Köck, for opening my eyes to the single-particle method, and for the always interesting and inspiring discussions. If it was not for you I would probably be a theoretical chemist, and in retrospect I must admit that the macromolecular world is more colourful than the quantum mechanical world. Joakim Lundqvist, for believing more in me than myself, and for putting your beloved Mg-chelatase molecule in the hands of novel and never tested methods. I will miss the work we performed together during the few, but very intense, months in the “monkey cage” for the rest of my life. Gustav Nordlund, for showing me that Spider can perform tasks beyond processing single-particle images, and for our lively and outrageous discussions. Perhaps one day the enigma of the oxidative phosphorylation will be solved… Dominika Kasperska, for your enthusiasm over a barely comprehensible method, and your ability to learn complicated stuff quickly. Thank you also for all high-quality data you have collected on both Mg-chelatase and Mediator preparations. Mats Lindstam, for teaching me Linux and for showing me the beauty of computer programming. Too bad we could not share office for longer. Gary Gippert, for introducing me to C. Peter Holm, for introducing me to the art of visualization, and for teaching me that “if equipped with a ‘tjockt pannben’ you are able to achieve a lot in the sciences.” Anders Liljas, for your encouraging words at tough times and for your
enthusiasm about ribosome structure and function as well as science in general. Roger D. Kornberg, for your interest in my work and for communicating paper IV of this thesis. Dave Bushnell, for teaching me that when it comes to transcription things are not always as simple as they might seem. Thank you also for providing me with a bed and nice company during my visit in Stanford. Joachim Frank, for the discussions about heterogeneity during your visit in Lund, inspiring me in the development of the methods for resolving conformational states. Steve Ludtke, for always patiently answering my stupid questions about EMAN, and for the nifty python script that came in handy in paper IV. Salam Al-Karadaghi, for giving Joakim and me the opportunity to explore the single-particle method via Mg-chelatase. Mats Hansson and the “Mg-chelatase crew” at the Biochemistry Department in Lund, for providing solid knowledge and nice preparations of our “favourite test particle”. All past and present members of the Department of Laboratory Medicine at KI, Novum, for your contributions to paper III-IV, especially: Vera, for the always high-quality protein preparations and Tomas, for the interesting discussions, and for your very nice thesis that I have stolen a lot from, thank you! All past and present members of the electron microscopy group at the Center for Structural Biochemistry, KI, Novum, for all your help with the specimen preparation, the microscope and the scanning, especially: Pasi, for always being there when I was panicking over the microscope, Kimberley, for always letting me borrow your stuff, and Ronny, for the interesting methodological discussions about multivariate statistics. All past and present members of the Molecular Biophysics Department in Lund, especially: Ullis, for all the fun we had during the work we performed together on the Frataxin molecule, and for our crazy trip to Il Coccio (I’ll never forget that one…), Sebban, for the interesting discussions about the translation mechanism, Ray, for answering a true novice’s questions about molecular biology, Anna, for actually believing in me being capable of solving the phase problem (that is the finest compliment I have ever had), I’ll do my best... All past and present members at the Biochemistry Department in Lund for making the coffee breaks a social event, sometimes even resembling a professional stand-up comedy show. Pappa och Karin, för att jag alltid varit välkommen att besöka er. Skogen och våra fantastiska måltider (Pappa) och löprundor (Karin) tillsammans har varit balsam för en sönderstressed doktorandsjäl. Markus Nilsson, för att du alltid lyssnat intresserat på mina galna idéer och kommit med kvalificerade programmeringsrelaterade synpunkter. Fia och Jon, för att ni tog emot mig när jag
vilje förverkliga min dröm om att flytta till södern för att läsa teoretisk kemi. Det var så
det började allthop... Anna Åsén, för att du så tålmodigt lyssnade på godnattsagorna
om NMR-spektrosopi... Cissi, för att du tränade min högerhjärnhalva. En vacker dag
kanske jag målar klart den där tavlan... Storasyster, Beppe och barnen, för att ni alltid
fanns där. Syrran, dags att sy upp klännningen... Ne och Jonis, för att man inte kan
vara olycklig hamrandes på en bongotrumma mitt ute bland kossorna skrålandes en
"Klausöns blues". Tomas och Susanne, för er gästvänlighet och vänskap och Tomas,
för våra studtals rätt djuplodande diskussioner om Feynman och kvantmekanikens
grunder. Merike och Jonas, för att er soffa har varit så mycket trevligare än ett sunkigt
hotellrum. Kristina Bergholm, för att mina resor och min studiegång alltid fungerat på
bästa möjliga vis, väl anpassat efter min förvirrade personlighet.

Hans Elmlund, Lund 2008
REFERENCES


