TECHNISCHE UNIVERSITÄT MÜNCHEN
Max-Planck-Institut für Biochemie
Abteilung für Molekulare Strukturbiologie

Geometric analysis of macromolecule organization within cryo-electron tomograms

Luis Eugenio Kuhn Cuellar

Vollstandiger Abdruck der von der Fakultät für Chemie der
Technischen Universität München
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Johannes Buchner
Prufer der Dissertation:	1. Hon.-Prof. Dr. W. Baumeister
	2. Prof. Dr. Sevil Weinkauf

Die Dissertation wurde am 05.02.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 13.02.2017 angenommen.

Contents

Zusammenfassung v

Abstract vi

1. Introduction 1
- **1.1 Cryo-Electron Tomography** 1
- **1.2 RuBisCO and Polysomes** 1
- **1.3 Subtomogram analysis in CET and Visual Proteomics** 2
- **1.4 Local Geometric Analysis of Macromolecules** 4
- **1.5 Graphs and Probabilistic Graphical Models** 5
- **1.6 Thesis Outline** 6

2. Background 9
- **2.1 Transmission Electron Microscopy** 9
- **2.2 Cryo-Electron Tomography** 11
  - 2.2.1 Sample Preparation 11
  - 2.2.2 Tilt-Series Acquisition 12
  - 2.2.3 Tomogram Reconstruction 12
- **2.3 Subtomogram Analysis** 13
  - 2.3.1 Macromolecule Localization 14
  - 2.3.2 Subtomogram Alignment and Averaging 16
  - 2.3.3 Subtomogram Classification 16
  - 2.3.4 Resolution Measurement 17
- **2.4 Radial Distribution Functions** 17
- **2.5 Topology Graphs** 18
  - 2.5.1 Markov Random Fields 19
  - 2.5.2 Loopy Belief Propagation 20
- **2.7 Photosynthesis** 21
  - 2.7.1 The Pyrenoid of C. Reinhardtii 22
- **2.8 Protein Synthesis** 23
  - 2.8.1 Polysomes 24

3. Materials and Methods 27
- **3.1 Implementation of algorithms** 27
- **3.2 Dataset Acquisition and Processing** 28
  - 3.2.1 Tomograms of *C. reinhardtii* Pyrenoids 28
  - 3.2.2 Tomograms for Polysome Detection 29
    - 3.2.2.1 Experimental Tomograms of *E. coli* Lysate 29
    - 3.2.2.2 Simulated Tomograms of *E. coli* lysate 29
    - 3.2.2.3 Tomograms of Microsomes Derived from Rough ER 30
    - 3.2.2.4 Subsequent Processing 30

4. Local Geometric Analysis 31
- **4.1 Introduction** 31
- **4.2 Local Geometric Analysis Workflow** 31
- **4.3 Local Extraction of Geometric Information** 33
- **4.4 Analysis of 3D position Vectors** 34
  - 4.4.1 Handling Particle Symmetry in 3D Vector Space 34
  - 4.4.2 Cluster Analysis of 3D Vectors 35
  - 4.4.3 Peak Extraction of 3D Vector Clusters 35
- **4.5 Analysis of Relative Rotations** 35
  - 4.5.1 Handling Particle Symmetry in Relative Rotation Space 36

4.5.2 Cluster Analysis of Rotations ... 36
4.5.3 Peak Extraction of Rotation Clusters ... 37
4.6 Discussion ... 37

5. Local Organization of RuBisCO in the Pyrenoid of *C. reinhardtii* Cells ... 41
5.1 Introduction ... 41
5.2 Tomogram Dataset ... 41
5.3 Localization of RuBisCO Complexes ... 43
5.4 Alignment of RuBisCO Subtomograms ... 45
5.5 Classification of RuBisCO Subtomograms ... 45
5.6 *In situ* Structure of the RuBisCO Complex ... 47
5.7 Local Organization of RuBisCO Complexes in the Pyrenoid ... 48
5.7.1 Radial Distribution of Near Neighbors ... 48
5.7.2 Closest Neighbor Cluster ... 49
5.7.3 Side Neighbor Cluster ... 49
5.7.4 Top Neighbor Cluster ... 54
5.7.5 Corner Neighbor Cluster ... 54
5.7.6 Local Organization Model ... 54
5.8 Discussion ... 57

6. Polysome Detection ... 61
6.1 Introduction ... 61
6.2 Polysome Detection Workflow ... 61
6.3 Probabilistic Model for the Local Geometric Arrangement of Polysomes ... 63
6.4 Graphical Model for Probabilistic Polysome Detection ... 63
6.4.1 Topology Graph of Ribosomes ... 63
6.4.2 Markov Random Field for Maximum-A-Posteriori Classification ... 65
6.4.3 Polysome Clustering ... 66
6.5 Tomogram Datasets ... 67
6.6 Results ... 67
6.6.1 Local Model of Cytosolic Bacterial Polysomes ... 67
6.6.2 Quantitative Evaluation of Bacterial Polysome Detection ... 69
6.6.3 Local Model of ER-Associated Polysomes ... 70
6.6.4 Local Model of Cytosolic Mammalian Polysomes ... 72
6.6.5 Analysis of ER-Associated and Cytosolic Mammalian Polysomes ... 72
6.7 Discussion ... 77

7. Conclusion ... 81

Abbreviations ... 83

Bibliography ... 85

Acknowledgements ... 96

Zusammenfassung

Kryoelektronentomographie (KET) bietet einen noch nie da gewesenen Blick auf die native zelluläre Umgebung bei molekularer Auflösung. Während hochaufgelöste Strukturen abgebildeter molekularer Komplexe mittels Subtomogram Analyse bestimmt werden können, gibt ein Tomogramm zusätzlich Auskunft über die genauen Positionen und Orientierungen dieser Makromoleküle innerhalb der Zelle. Die Analyse geometrischer Beziehungen zwischen benachbarten Makromolekülen, kann strukturelle Einblicke in die molekularen Wechselwirkungen bieten und einheitliche supramolekulare Anordnungen identifizieren. Allerdings müssen rechnergestützte Verfahren für die quantitative Analyse dieser dichten geometrischen Informationen entwickelt werden.

Diese Arbeit stellt eine statistische Methode für die Analyse der unmittelbaren Nachbarschaft von Makromolekülen vor, welche darauf abzielt die 3D-Konfiguration von Makromolekül Paaren zu identifizieren. Diese Methode der lokalen geometrischen Analyse beansprucht allgemeine Gültigkeit; sie kann für jede Art von Makromolekülen angewendet werden und berücksichtigt molekulare Symmetrien. In dieser Arbeit wurde diese Methode genutzt, um die 3D Organisation von RuBisCO Enzymen innerhalb des Pyrenoid von *C. reinhardtii* Zellen zu untersuchen. Mittels Subtomogram Analyse wurden RuBisCO Komplexe innerhalb von Pyrenoid Tomogrammen lokalisiert und es konnte eine Struktur des nativen Komplexes bei 16 Å Auflösung gewonnen werden. Lokale geometrische Analyse von RuBisCO Komplexen lässt auf eine flüssigkeitsähnliche Pyrenoid Matrix schließen. Aus den vorherrschende Konfigurationen von RuBisCO Paaren wurde ein geometrisches Modell der Einheitszelle von RuBisCO Nachbarn entwickelt, das der 3D- Konfiguration dichtester Kugelpackung ähnelt.

Im nächsten Schritt dieser Arbeit wird die lokale geometrische Analyse zur Identifizierung von Strukturen höherer Ordnung verwendet. Flexible supramolekulare Anordnungen von Ribosomen und Messenger-RNA, sogenannte Polysomen, werden mit der hier vorgestellten Methode detektiert. Mittels lokaler geometrischer Analyse wird die bevorzugte 3D-Anordnung benachbarter Ribosomen in Polysomen extrahiert. Diese Vorinformation wird dann in einem nachgeschalteten Detektionsverfahren genutzt, das Ribosomen als Knotenpunkte in einem Graphen dargestellt und durch ein Markov Random Field gruppiert um Polysomen zu lokalisieren. Leistungsbewertung der Methode basierend auf synthetischen und experimentellen Tomogrammen bakterieller Zelllysate zeigt eine 96%ige Vorhersagegenauigkeit. Schließlich wurde das Verfahren angewendet, um cytosolische und membranassoziierte Polysomen in Tomogrammen zu detektieren, die Membranvesikel abgeleitet vom endoplasmatischen Retikulum aus Maus- Myeloma-Zellen abbilden.

Diese Arbeit stellt geometriebasierte Methoden für die Analyse lokaler Organisation von Makromolekülen und die Erfassung von supramolekularen Strukturen in der KET vor. Während sich diese Analyseverfahren vorläufig auf die Verteilung eines einzelnen Makromolekül-Typs beschränken, könnten sie um die Einbeziehung mehrerer Makromolekül-Klassen erweitert werden, um quantitative Analysen für die visuelle Proteomik zu ermöglichen. Somit stellen die hier vorgestellten Verfahren einen Schritt in Richtung der räumlichen Erfassung der gesamten molekularen Soziologie einer Zelle dar.

Abstract

Cryo-electron tomography (CET) provides unprecedented views into the native cellular environment at molecular resolution. While subtomogram analysis yields high-resolution native structures of molecular complexes, it also determines the precise positions and orientations of these macromolecules within the cell. Analyzing the geometric relationships between adjacent macromolecules can offer structural insights into molecular interactions and identify supramolecular ensembles. However, computational tools must be developed for quantitative analysis of this dense geometric information.

This thesis presents a statistical method for analyzing the local neighborhoods around macromolecules, with the aim of identifying 3D configurations of macromolecule pairs. This method of local geometric analysis emphasizes generality; it can be applied to any type of macromolecule and incorporates molecular symmetry. Here, the method was used to study the 3D organization of Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) enzymes within the pyrenoid of *C. reinhardtii* cells. Subtomogram analysis identified RuBisCO complexes within pyrenoid tomograms, producing an *in situ* structure at 16 Å resolution. Local geometric analysis of RuBisCO complexes suggested a fluid-like pyrenoid matrix. Predominant configurations of RuBisCO pairs were identified and combined into a geometric model of the unit cell of RuBisCO neighbors, showing a 3D configuration similar to closely packed spheres.

Next, this thesis progresses from local geometric analysis to identification of higher-order structures by presenting a method to detect polysomes, flexible supramolecular ensembles of ribosomes and messenger RNA. Local geometric analysis extracted the 3D arrangements of neighboring ribosomes in polysomes. This prior information is then used in the detection method, where ribosomes are represented as nodes in a graph and clustered by a Markov random field to reveal polysomes. Performance evaluation on synthetic and experimental tomograms of bacterial lysate indicated a 96% prediction accuracy. Finally, the method was applied to tomograms of rough microsomes derived from the endoplasmic reticulum (ER) of mouse myeloma cells, with the aim of detecting cytosolic and ER-associated polysomes.

This thesis presents geometry-based methods for analyzing the local organization of macromolecules and detecting supramolecular structures in CET. While these methods operate on macromolecule distributions of a single type, they could be expanded to incorporate multiple classes of macromolecules, enabling quantitative analysis for visual proteomics. Thus, they represent a step towards the spatial dissection of the complete macromolecular sociology of cells.

1. Introduction

Molecular studies of biological systems require imaging techniques that produce magnified views of biological samples. While spatial resolution of light microscopy is limited by the wavelength of visible light, techniques such as Nuclear Magnetic Resonance (NMR), X-ray crystallography, and cryo-electron microscopy Single Particle Analysis (SPA) allow studying isolated macromolecular complexes at subnanometer scales. NMR, SPA and X-ray crystallography provide three-dimensional (3D) information of macromolecular structures at resolutions in the Angstrom range [Cheng, 2015]. An enormous limitation is that these methods require a purified, homogenous sample containing many, ideally identical, copies of the macromolecule of interest. Therefore, information about the intracellular context, in which the macromolecule functions, is absent and the purification process can introduce artifacts in the sample, such as induced conformational changes and complex disassociation.

1.1 Cryo-Electron Tomography

Cryo-electron tomography (CET) is an imaging technique that allows the observation of macromolecular complexes in their cellular environment under close-to-native conditions. It provides a 3D electron density map of the biological sample [Lucić et al., 2005]. CET comprises four major steps: (1) a biological sample is placed on a grid and rapidly cooled to cryogenic temperatures, freezing the sample sufficiently fast to avoid the formation of ice crystals, which would otherwise compromise the structural integrity of the biological material. (2) The frozen-hydrated sample is placed in a Transmission Electron Microscope (TEM) to acquire two-dimensional (2D) projections of the sample at different tilt angles (tilt-series). Due to the radiation sensitivity of biological samples, the allowed electron dose applied to the sample must be distributed over the tilt-series, yielding micrographs with significantly lower signal than those acquired by SPA. (3) A reconstruction of the 3D density map (tomogram) of the sample is computed from the acquired 2D projections. (4) The tomogram is subsequently processed by computational means to extract biological information.

Given that CET allows visualization of macromolecules in their physiological environment, it has enabled *in situ* studies of membrane-bound and membrane-embedded macromolecular structures. For example, ribosomes bound to the endoplasmic reticulum (ER) and the ER translocation machinery for co-translational insertion of polypeptides into the membrane and lumen of the ER [Pfeffer et al., 2012, 2014, 2016; Pfeffer, Burbaum, et al., 2015].

1.2 RuBisCO and Polysomes

The Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) enzyme and the ribosome are macromolecular complexes that play pivotal roles in key biological processes, photosynthesis and translation, respectively.

During the light-independent reactions of the photosynthetic process, the RuBisCO enzyme is responsible for incorporating inorganic carbon into biomolecules. However, this enzyme is a very slow catalyst and suffers from an opposing side-reaction that leads to an energy-consuming salvage pathway. The enzymatic limitations of RuBisCO have driven microorganisms to develop carbon concentration mechanisms (CCM) to augment their photosynthetic process. A characteristic feature of CCMs is the packing of large amounts of RuBisCO complexes in micro-compartments, such as carboxysomes in cyanobacteria and the pyrenoid in eukaryotic algae [Meyer et al., 2016].

The RuBisCO structure of the *Chlamydomonas reinhardtii* alga has been determined to atomic resolution by X-ray crystallography [Taylor et al., 2001]. However, the 3D organization of RuBisCO complexes in its pyrenoid is still poorly understood. While a CET study of *C. reinhardtii* pyrenoids observed RuBisCO complexes in configurations similar that of closely packed spheres, resolution limitations precluded detailed analysis of their local geometric organization [Engel et al., 2015].

On the other hand, ribosomes are large macromolecular machines that synthetize proteins by translating messenger RNA (mRNA) into polypeptide chains. Supramolecular ensembles of ribosome particles translating a single mRNA molecule are called polysomes. Interestingly, polysomes have been observed to adopt characteristic structures, which are highly conserved across species and cellular environments. Cryo-electron tomograms of cytosolic polysomes from bacteria and of human cells revealed remarkably similar polysome structures [F. Brandt et al., 2009, 2010], while structural similarities of membrane-bound polysomes, both on the ER surface and in yeast mitochondria, have also been observed in CET studies [Pfeffer et al., 2012; Pfeffer, Woellhaf, et al., 2015].

However, detailed characterization of polysome structures and understanding of their function in the translation process remains elusive, as it requires *in situ* identification of polysome structures. Since detection of these flexible supramolecular structures in tomograms of macromolecule-rich environments is a challenging task, previous CET studies have been restricted to local analysis of polysome structures, i.e., only the relative 3D configuration of neighboring ribosomes in a polysome sequence has been analyzed (e.g. figure 1.2 A).

1.3 Subtomogram analysis in CET and Visual Proteomics

Identifying and localizing macromolecular complexes in CET is typically achieved by template matching [Frangakis et al., 2002; Ortiz et al., 2006]. In template matching, an exhaustive cross-correlation search of a tomogram aims to locate a macromolecular complex by comparing a structural template in different orientations against all same-sized regions within the tomogram. This process yields a cross-correlation function on tomographic space (i.e. the 3D space in a tomogram). By extracting maxima of this function (with respect to orientation and spatial coordinates) it is possible to determine the most probable position and orientation of a target macromolecular complex in a tomogram. It is noteworthy that attaining high specificity and sensitivity in macromolecule detection is a challenging task.

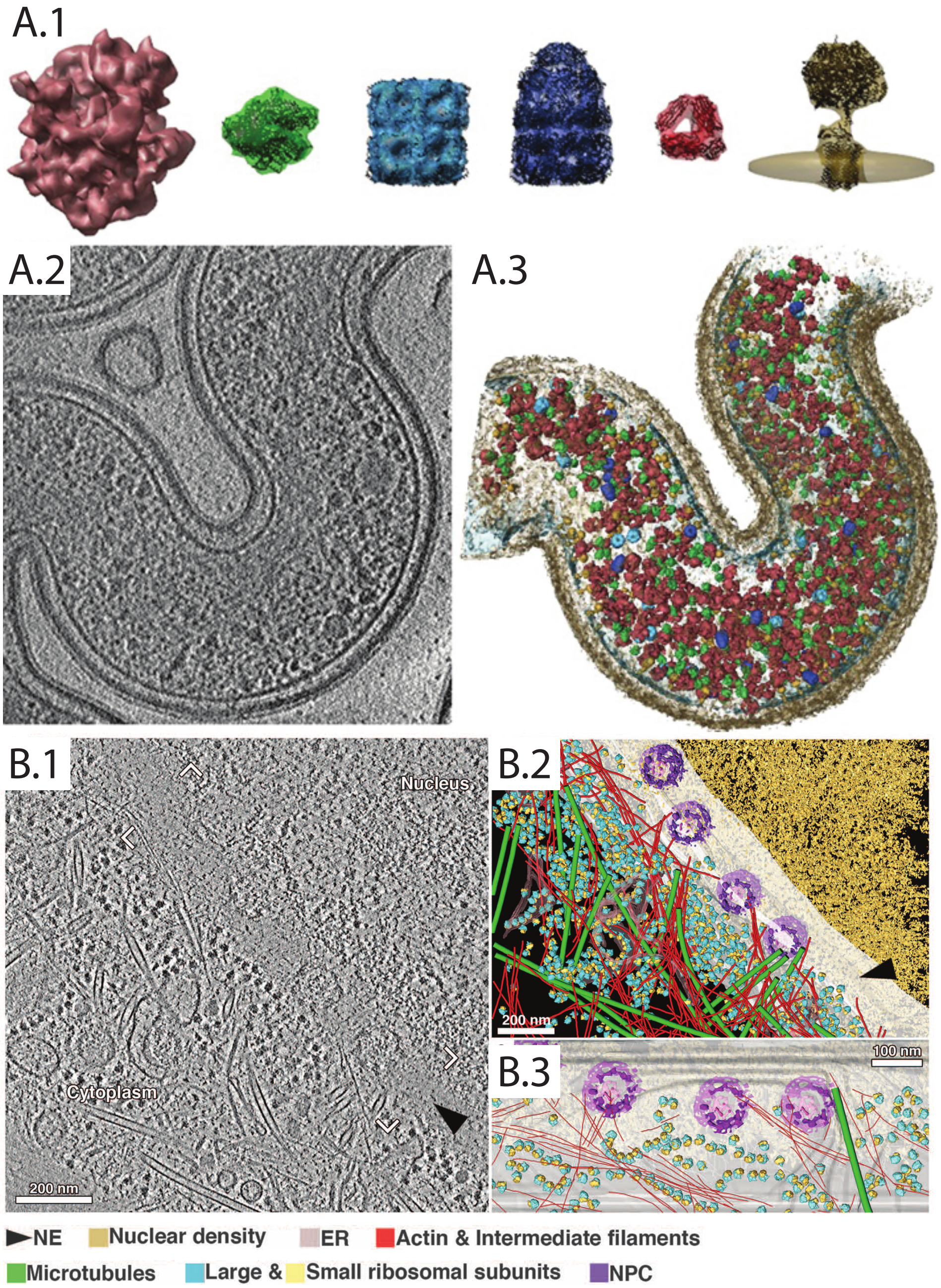

Figure 1.1: Concept of visual proteomics and its application. Visual proteomics aims at detecting the majority of macromolecular complexes within a cell, allowing statistical analysis of their distribution. (A) Visual proteomics of *Leptospira interrogans.* (A.1) Library of structural templates, from right to left: ribosome, RNA polymerase II, GroEL, GroEL-ES, Hsp, and ATP synthase. (A.2) Central slice of a tomogram from a *Leptospira interrogans* cell. (A.3) Structural templates from the library were identified in the tomogram to yield a 3D atlas of molecular complexes. Templates rendered in their computationally determined position and orientation in tomographic space, membranes depicted in blue and the cell wall in brown. (B) The nuclear periphery of a HeLA cell studied using a visual proteomics approach. (B.1) Slice of a tomogram from a region adjacent to the cellular nucleus. (B.2) Rendered tomogram depicting identified macromolecules labeled by color: nuclear envelope (NE), endoplasmic reticulum (ER), nuclear pore complex (NPC). (B.3) Tilted view of the rendered tomogram from a perspective perpendicular to the nuclear envelope. Adapted from [Förster et al., 2010; Mahamid et al., 2016].

Subvolumes (subtomograms) containing the putative macromolecular complex can be aligned and averaged to reduce noise and yield higher resolution structures than that of individual subtomograms [Hrabe & Förster, 2011]. Furthermore, subtomogram classification can be used to identify large conformational changes and further increase the resolution of individual class averages by producing subtomogram subsets with increased structurally homogeneity.

Applying template matching to cryo-electron tomograms using a library of structural templates permits the identification of specific macromolecular complexes in their native cellular context. The emerging field of visual proteomics aims to combine the geometric information of multiple types of macromolecules, i.e. their position and orientation in a tomogram, to generate 3D atlases within cellular landscapes (figure 1.1), which can visualize the proteome of a cell [Förster et al., 2010; Nickell et al., 2006]. Visual proteomics has the potential of characterizing the proteome of cells spatially, complementing mass spectrometric approaches that lack spatial information [Förster et al., 2010].

Furthermore, cellular maps of identified macromolecular complexes can be used for quantitative analysis of their spatial distribution within cells, cellular contextualization of conformational states, detection of interaction partners and characterization of their 3D geometric arrangement. For example, a CET-based molecular census of 26S proteasomes in intact hippocampal neurons [Asano et al., 2015] allowed statistical characterization of the abundance of predominant molecular species (double and single capped proteasomes) and their major conformational states (substrate-accepting state and substrate-processing state). It provided a 3D cellular atlas of 26S proteasomes depicting their position, orientation and conformational state.

The above-mentioned tasks require quantitative analysis of geometric parameters obtained by template matching and refined by subtomogram alignment and classification. Therefore, the implementation of methodologies for statistical analysis of 3D geometric information is imperative for the development of the visual proteomics field.

1.4 Local Geometric Analysis of Macromolecules

When the positions and orientations of macromolecular complexes in tomographic space have been determined, it is possible to conduct statistical analyses of the geometric relation between adjacent macromolecules. Concrete examples are the CET studies of the 3D organization of polysomes in bacterial lysate and in intact human cells [F. Brandt et al., 2009, 2010], where ribosomes were localized by template matching and their near-neighbor distribution of center-to-center vectors and relative orientations was subjected to cluster analysis. Statistical analysis of these geometric features enabled the identification of predominant geometric configurations between neighboring ribosomes in a polysomic sequence, which once extrapolated produce characteristic polysome topologies. Additionally, identification of candidate ribosomal components for surface interactions between adjacent ribosomes was possible by fitting atomic models into ribosome subtomogram averages arranged as dictated by the mean neighbor configuration derived from cluster analysis of the geometric features. Furthermore, analysis of the above-mentioned geometric features provided structural insights into the mechanics of the translation process: mRNA molecules were predominantly buried inside the polysome structures, mRNA exits and entries of adjacent ribosomes were consistently observed in close proximity, while their peptide tunnel exits were oriented away from each other.

Local geometric analysis of macromolecules in cryo-electron tomograms can elucidate the 3D arrangement of macromolecules in supramolecular ensembles and allow structural characterization of macromolecular interactions. Thus, the development of methodologies for statistical analysis of geometric features (e.g. 3D center-to-center vectors, relative orientations) derived from adjacent macromolecules in tomographic space, is an important step towards understanding the spatial relation between macromolecules in their physiological environment.

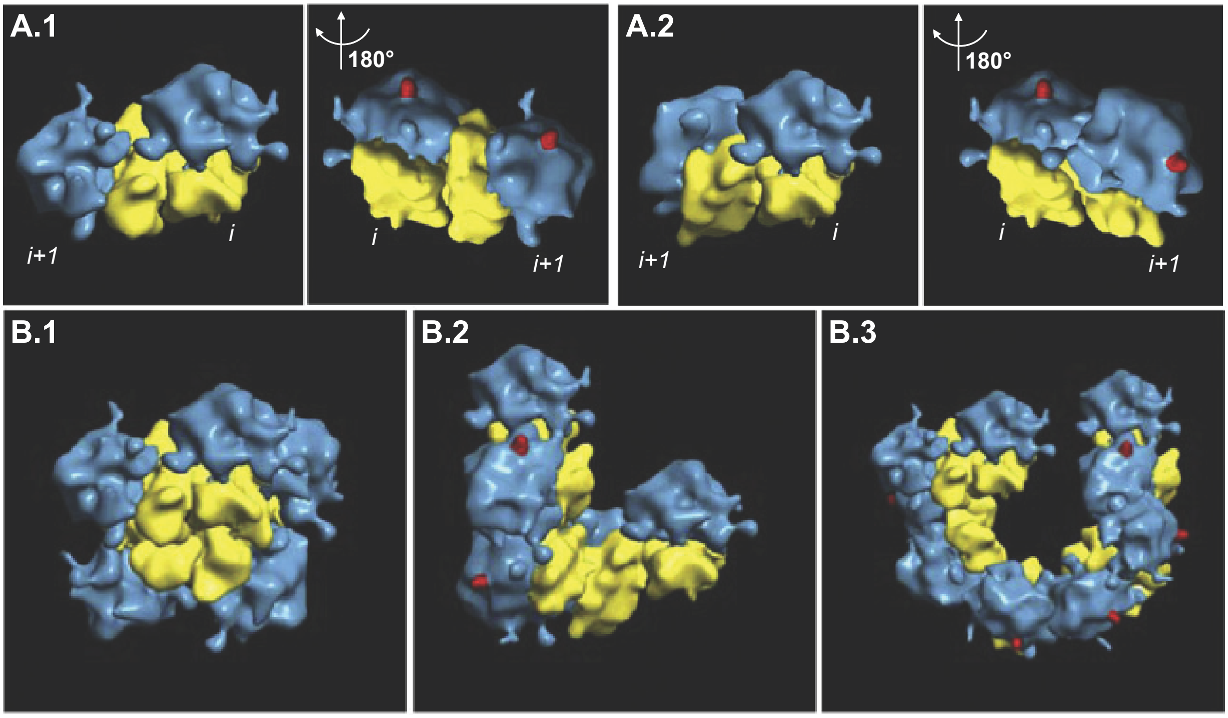

Figure 1.2: Statistical analysis of neighboring 80S ribosomes in intact human cells for geometric characterization of polysomes [F. Brandt et al., 2010]. (A) Cluster analysis of center-to-center vectors and relative orientations of adjacent ribosomes identified predominant neighbor configurations. Neighbors on the 5' side of the i^{th} ribosome in the polysomic sequence are termed *i+1*. The two predominant 3D arrangements (A.1 and A.2) differ by a neighbor rotation of $>$ 90°. (B) Models of polysome structures were generated by extrapolating previously identified configurations of ribosome pairs (A). Extrapolating configuration A.1 yielded a compact helical topology (B.1), while extrapolation of A.2 produced a much loose helical topology (B.2). (B.3) Spiral polysome topologies were produced by 1:1 alternation of configurations A.1 and A.2. Ribosomal large and small subunits in blue and yellow respectively, peptide tunnel density in red. Adapted from [F. Brandt et al., 2010].

1.5 Graphs and Probabilistic Graphical Models

Graphs are mathematical structures composed of nodes and edges used to connect pairs of nodes. Graph theory provides a natural representation for geometric information (e.g. the distribution of points in 3D space) and a large number of algorithms to solve a variety of topological problems [Bollobás, 1979]. Therefore, graphs are especially well suited to deal with visual proteomics tasks, as they can be employed to describe and analyze the distribution of macromolecules in cryo-electron tomograms.

Probabilistic graphical models are graphs that provide a diagrammatic representation of joint probability distributions [Bishop, 2006]. In these frameworks, nodes represent random variables, while edges denote probabilistic relations between these variables. Bayesian networks and Markov random fields are examples of graphical models, which have been successfully used for probabilistic inference in a wide variety of image processing applications. Concrete examples of such applications are image de-noising and segmentation [Bishop, 2006; Blake et al., 2011], where these probabilistic frameworks are used to model the tendency of neighboring image sites to be coherent, i.e., to have the same pixel value or to belong to the same region. Here, each pixel is represented by a node and only nodes corresponding to directly adjacent pixels are connected with an edges. While not only pixels that are direct neighbors are correlated, graphical models are capable of capturing implicit long-range dependencies that arise from short-range connections. Moreover, describing only short-range connections generates sparsely connected graphs, which allow inference algorithms to mine long-range correlations with low computational cost [Blake et al., 2011].

In a similar manner, probabilistic graphical models can be used to analyze the 3D organization of macromolecules in cellular volumes. Once a set of macromolecular complexes have been identified in a tomogram, a probabilistic framework can be used to describe their 3D organization, e.g., by representing each macromolecule with a node and connecting nodes of adjacent macromolecules with an edge. Inference algorithms can then be applied to mine supramolecular information, such as identification of flexible polysome structures.

1.6 Thesis Outline

This thesis describes novel methodologies to analyze the 3D organization RuBisCO complexes and polysomes, as imaged by CET. A general method for local geometric analysis of macromolecular complexes is developed, which is then applied to investigate the local organization of RuBisCO complexes in cryo-electron tomograms of *C. reinhardtii* pyrenoids. Finally, a graph-based method for automated detection of polysomes is presented. It implements a probabilistic graphical model to detect flexible polysome structures in cryo-electron tomograms, using only the positions and orientations of previously localized ribosomes as input. The subsequent chapters of this thesis are organized as follows:

- Chapter 2 provides background on the TEM, CET, the mathematical concepts used for geometry analysis, and a basic description of the biological systems on which the proposed methodologies are applied.

- Chapter 3 describes the materials and methods used in this thesis.

- Chapter 4 presents a strategy for statistical analysis of geometric information, derived from the local organization of previously detected macromolecular complexes in CET data.

- In chapter 5, the method described in chapter 4 is applied to describe the local organization of RuBisCO complex in the crowded environment of the pyrenoid from *C. reinhardtii* cells. A set of predominant arrangements of RuBisCO complex pairs is identified and subsequently used to propose a model of the local organization of RuBisCO complexes.

- In chapter 6, a probabilistic polysome detection method is descried. Statistical models of geometric configurations between neighboring ribosomes in characteristic polysome topologies are constructed using the methodology presented in chapter 4. Within the polysome detection method, these models are regarded as prior knowledge of the local structure of polysomes. The method is evaluated on synthetic and experimental tomograms of bacterial lysate and then applied to tomograms of microsomal preparations derived from rough ER, with the objective of detecting cytosolic and ER-associated polysomes.

- Finally, chapter 7 provides an overarching summary of the thesis, discusses contributions and provides an outlook into future research questions.

2. Background

2.1 Transmission Electron Microscopy

The standard setup of a transmission electron microscope (TEM) has seven major components: (1) The electron gun emits and accelerates electrons to form a spatially and temporally coherent electron beam. Typically, a field emission gun (FEG) is used to generate a highly bright and coherent electron beam. The usual acceleration voltage ranges between 100 kV to 300 kV. (2) The condenser system contains two or three electron lenses, they use electromagnetic coils to focus the electron beam onto a small region of the specimen. Usually it is sufficient to illuminate a specimen area of approximately 1 μm [Reimer, 1984]. A circular metallic mesh (EM grid) of approximately 3 mm in diameter supports the sample, the EM grid is placed on a sample holder system, which is subsequently mounted on (3) the specimen stage. The sample holder and specimen stage permit translation and single-axis rotation of the EM grid. The sample holder is kept at cryogenic temperatures using liquid nitrogen, as it comes in direct contact with the EM grid. (4) The objective lens creates the image of the sample, which is later magnified by (5) the intermediate lens. Finally, (6) the projection lens directs the image into (7) the image detector. The image detector is used to produce a digital image from the signal carried by the electrons. Historically, a TEM was equipped with a Charged Coupled Device (CCD) camera, which uses a scintillator to translate the electron signal to photons, the signal is subsequently transmitted to the CCD chip to generate a digital image. The advent of direct electron detectors significantly improved image quality of the TEM, by avoiding signal degradation from the translation process between electrons and photos, as these devices are capable of recording digital images from an electron signal, dramatically increasing quantum efficiency and image contrast [Faruqi & Mcmullan, 2011]. It is important to point out that the column of the TEM is maintained at high vacuum, allowing the electron beam to interact only with the specimen [J. Frank, 2006; Lucić et al., 2005; Reimer, 1984]. Figure 2.1 depicts a cross-section of a typical TEM.

As the planar wave of electrons passes the sample, electrons interact with the Coulomb potential of atoms in the sample, modifying the path of electrons. Electrons can be elastically or inelastically scattered, while elastic scattering of electrons transfers negligible energy into the sample, inelastic scattering events introduce significant energy, leading to radiation damage in biological samples [Reimer & Kohl, 2008; Williams & Carter, 2009]. The energy introduced by inelastic scattering events heats the sample and can lead to free radicals, by breaking ionic and covalent bonds. Thus, radiation sensitivity of biological material limits sample thickness, as the chance of inelastic scattering events increases in thicker samples. Furthermore, while highly scattered electron are blocked by the objective aperture, an energy filter can be placed between the projection lens and the image detector to filter out inelastically scattered electrons (i.e. lower energy electrons). Energy filters improve contrast and prevent image blurring from chromatic aberrations of the TEM [Lucić et al., 2005].

Micrograph images produced by the TEM are 2D projections of the sample under investigation. Image contrast is the intensity difference between adjacent areas, it is formed by two mechanisms, amplitude and phase contrast [Reimer & Kohl, 2008]. (1) Amplitude contrast is produced by differences in the number of detected electrons between adjacent areas in the image plane. Since highly dense regions in the sample strongly scatter electrons, the objective aperture and an energy filter can be used to block elastically and inelastically scattered electrons, respectively. Therefore, areas on the image plane corresponding to low and high density regions in the sample will register different amounts of electrons. This type of contrast increases as the aperture radius decreases. (2) Phase contrast is produced by constructive and destructive interference of phase-shifted electron waves on the image plane. Electron waves become phase shifted by scattering events as they traverse the sample. Images produced by phase contrast are projections of the electrostatic potential of the sample, convoluted with the inverse Fourier transform of the contrast transfer function (CTF), the CTF is the product of acquisition parameters (e.g. acceleration voltage and defocus value) and imaging conditions in the TEM (e.g. electron beam coherence and spherical aberration) [Lucić et al., 2005]. Furthermore, the image detector records the amplitude of the electron wave function, squared, i.e. the probability of an electron being detected at any particular pixel. It is noteworthy that both amplitude and phase contrast are damped by the modulation transfer function (MTF) of the image detector

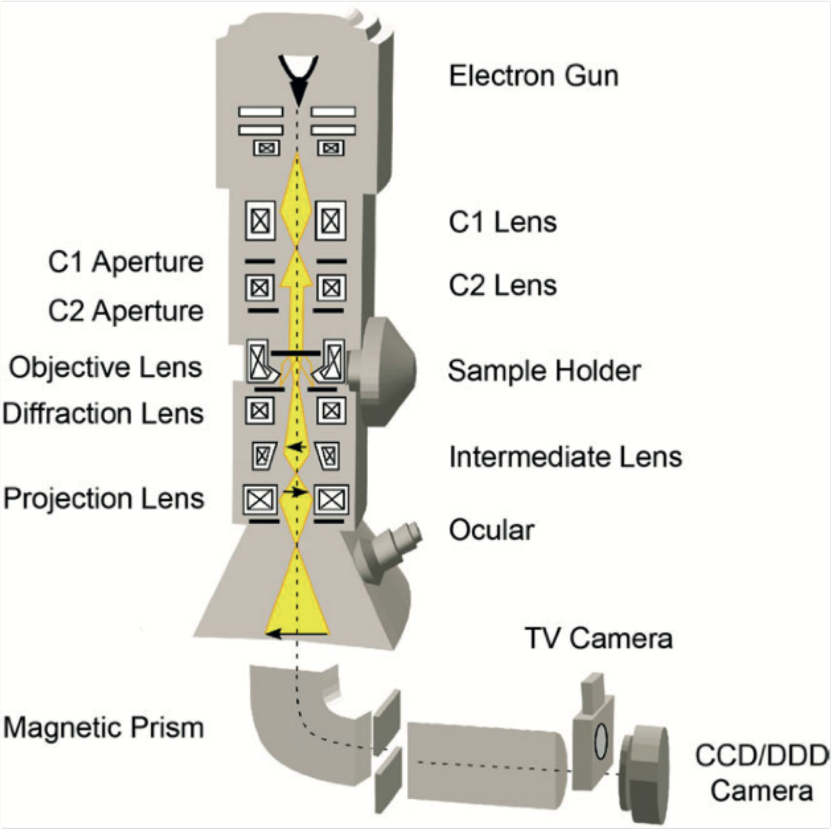

Figure 2.1: Cross-section diagram of a TEM. The diagram shows all major components of a typical TEM, including a two-lens condenser system (C1 and C2) and a magnetic prism acting as an energy filter [Krivanek et al., 1995]. Adapted from [Schweikert, 2004].

Historically, the defocus value used to record micrographs of frozen-hydrated samples established a practical limit on the maximum resolution attainable. Resolution was restricted to the spatial frequency of the first zero crossing of the CTF, since frequency bands at higher resolution are hard to interpret given the alternating contrast produced by CTF oscillation. Extracting structural information at frequencies higher than the first CTF zero crossing (CTF correction) is only possible when the experimental CTF has been accurately estimated. While CTF determination was a challenging task given the low signal-to-noise ratio (SNR) of images collected with CCD cameras, direct electron detectors have significantly increased the SNR of micrographs, allowing reliable estimation of the experimental CTF. Moreover, periodogram averaging can be used to improve CTF determination [Fernández et al., 2006; Zanetti et al., 2009]. In this strategy, a micrograph is divided into tiles and their power spectra is averaged, a theoretical CTF is subsequently fitted to the averaged power spectrum. It is important to point out that micrograph images recorded at small defocus values allow higher frequency information to be available at the cost of reducing contrast. Figure 2.2 shows theoretical CTF at different defocus values.

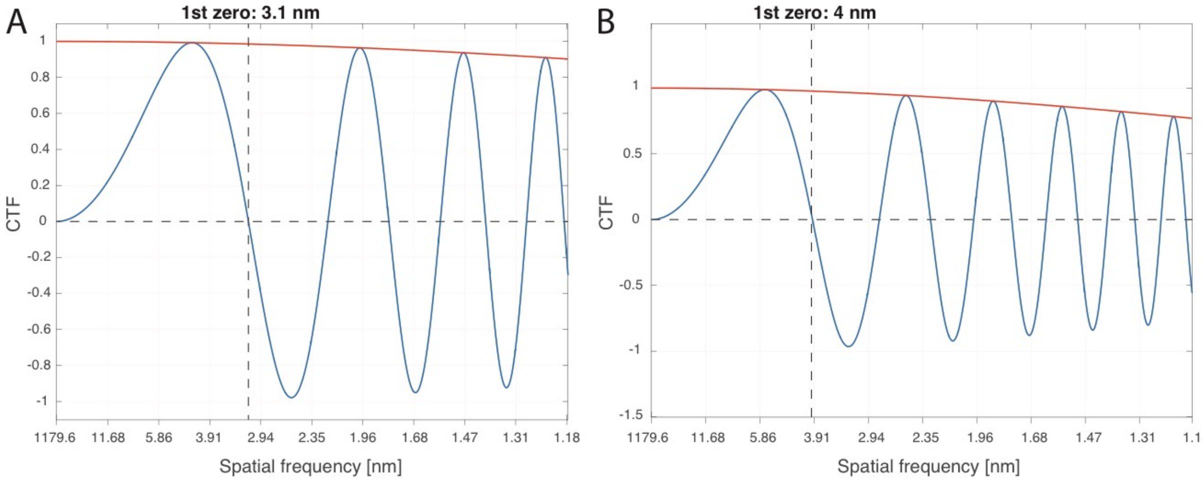

Figure 2.2: Simulated CTFs at different defocus values. CTFs (blue) and envelope functions (red) depicting the first zero crossing at (A) 5 μm and (B) 8 μm defocus values. All CTFs were calculated for a spherical aberration of 2.7 mm, acceleration voltage of 300 kV and a pixel size of 0.288 nm. The values were chosen to agree with the acquisition parameters used for the experimental data presented in the following chapters.

2.2 Cryo-Electron Tomography

2.2.1 Sample Preparation

An aqueous biological sample is applied on the EM grid and rapidly frozen into a glass-like state, i.e. the sample is vitrified in crystal-free amorphous ice. There are two predominant methods for vitrification, plunge freezing [Dubochet et al., 1988] and high pressure freezing [Studer et al., 2008]. Plunge freezing is used if sample thickness is below 10 μm [Asano et al., 2016], the sample is rapidly submerged into a cryogen, i.e. a liquid at less than −160°C, such as liquid ethane. Sample temperature is reduced to −140°C at an approximate cooling speed of 10^5°C/s, avoiding the formation of ice crystals that would otherwise compromise the structural integrity of the sample. High-pressure freezing is used for samples up to 200 μm in thickness [Asano et al., 2016]. Here rapid cooling is combined with the application of high pressure, acting as a physical cryo-protectant that lowers the melting point of water.

Samples above 1 μm in thickness are virtually intransparent to electrons [Lucić et al., 2013] and require thinning under cryo-conditions. Sample thinning can be achieved by cutting with a cryo-ultramicrotome. However, this methodology usually introduces compression artifacts and crevasse deformations in the sample [Han et al., 2008]. A compression-free strategy for sample thickness reduction is cryo-focused ion beam (cryo-FIB) milling, where a focused beam of gallium ions is used to ablate material at a selected region of the sample, creating electron transparent windows where thickness typically ranges from 100 to 300 nm [Schaffer et al., 2015].

2.2.2 Tilt-Series Acquisition

Once the sample is prepared and loaded into the specimen stage, a set of micrographs is acquired at different orientations, as the sample is tilted by the computer-controlled specimen stage. Physical restrictions of the TEM and large sample thickness at high tilt angles usually restricts tilt range to $\pm 60°$. Since frozen biological samples are very susceptible to radiation damage [Glaeser, 1971], the applied electron dose needs to be restricted to ~ 100 $e^-/\dot{A}^2$, significantly reducing the SNR of each micrograph. Furthermore, sample sensitivity to radiation damage also restricts angular increment of the tilt-series from 2° to 3°. The size (s) of the imaged sample in the direction of the electron beam and the number of projections in the tilt-series (n), determine the maximum isotropic resolution (r) attainable in the resulting tomogram [Crowther et al., 1970]:

$$r = \frac{\pi d}{n} \tag{2.1}$$

Even though automated tilt-series acquisition tools compensate for large displacements caused by mechanical inaccuracies of the specimen holder during tilt movement, the recorded micrograph set needs to be further aligned within a coordinate system common to the complete tilt-series. The aim of tilt-series alignment is to correct shift, rotation and magnification differences before tomographic reconstruction [J. Frank, 2006]. Computational alignment of projections is commonly performed by manual or automatic tracking of high-contrast features (e.g. fiducial gold markes) throughout the tilt-series [S. Brandt et al., 2001a, 2001b]. Typically, a least-squares procedure is used to align the micrograph set using the coordinates of localized features, minimizing an alignment error as a function of lateral shifts, tilt-axis angle and magnification changes [Amat et al., 2010].

2.2.3 Tomogram Reconstruction

Once the 2D projections of tilt-series have been aligned, and if required CTF corrected, a density volume of the imaged sample is reconstructed. The mathematical principles behind tomographic reconstruction base on the central slice theorem, which states that the Fourier transform of a 2D projection from a 3D object corresponds to a central section of the 3D Fourier transform of the imaged object [J. Frank, 2006; Radon, 1917]. However, in CET, the limited tilt range leads to missing information in a wedge-shaped region in Fourier space, causing distortions in the resulting tomogram, such as elongation in the direction of the electron beam. Figure 2.3 illustrates the so-called ‘missing wedge’ problem.

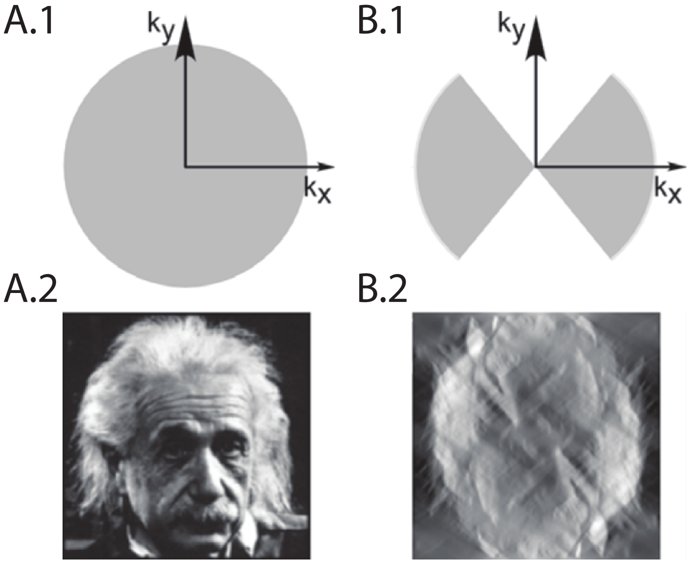

Figure 2.3: Depiction of the missing wedge problem in CET. (A.1, B.1) Sampled Fourier space (gray) and resulting images (A.2, B.2 respectively). The image resulting from a fully sampled Fourier space (A.1) is isotopically resolved in real space (A.2), as opposed to the heavily deformed image (B.2) resulting from a partially sampled Fourier space (B.1). Adapted from [Förster et al., 2008].

Since interpolation in Fourier space is a challenging task, real space reconstruction methods are typically used. The most commonly used real space method in CET is weighted backprojection (WBP), given its computational simplicity: micrographs are projected back into a reconstruction volume, regions where mass is found in the original object are reinforced as back-projected images intersect inside the volume. To obtain a faithful reconstruction of the object, projections need to be weighted in Fourier space to account for uneven sampling, otherwise low frequency information will be artificially enhanced in the resulting tomogram. It is worth mentioning that aside from WBP, there is a variety of real space [Marabini et al., 1998; Penczek et al., 1992; Wan et al., 2011] and Fourier space [Chen & Förster, 2014; Penczek et al., 2004; Sandberg et al., 2003; Zhang et al., 2008] reconstruction methods.

2.3 Subtomogram Analysis

After tomographic reconstruction, the resulting 3D density map can be interpreted by computational analysis of subvolumes containing the macromolecular structure of interest. A typical workflow aimed at structural characterization of macromolecular complexes follows four basic steps: (1) a macromolecular complex of interest is localized in a tomogram, (2) the corresponding subvolumes are aligned and averaged to overcome the low SNR. (3) Aligned subtomograms are subsequently subjected to classification. Here, the objective is the identification of structural heterogeneity, stemming from different conformations of the observed complexes. (4) Finally, the resolution of a structurally homogenous subtomogram average can be measured. Figure 2.4 illustrates a 2D analogy of the typical subtomogram analysis workflow.

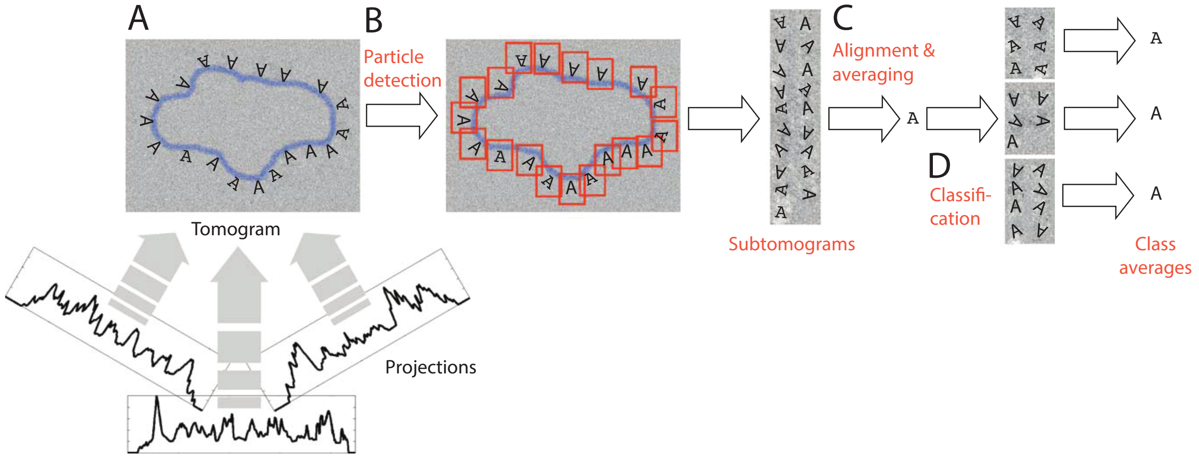

Figure 2.4: 2D simplification of a typical subtomogram analysis workflow. (A) After a tomogram has been reconstructed from a tilt-series, (B) The macromolecule of interest ('A's) is localized in tomographic space and subtomograms containing the target macromolecule are extracted. (C) Subtomogram alignment and averaging is used to produce a higher resolution structure of the target complex. (D) Classification procedures allow the identification of structural heterogeneity (denoted by different fonts), finally, the resolution of class averages is measured, usually exceeding the resolution of the unclassified average. Adapted from [Hrabe & Förster, 2011].

2.3.1 Macromolecule Localization

Localization of macromolecules of interest inside a tomogram is of paramount importance, as it is the basis of the subsequent subtomogram analysis. Template matching is a commonly used method for macromolecule localization in CET; an exhaustive cross-correlation search compares a template of a macromolecular complex in a set of different rotations to each same-sized region inside the tomogram. The template matching procedure has three major steps: (1) Structural template generation, (2) scoring function computation and (3) peak extraction. The template matching pipeline is depicted in figure 2.5 B.

Even though a SPA structure can be used as a template, preparation of structural templates from atomic models derived from X-ray crystallography is a standard practice. The atomic model (i.e. the list of atomic coordinates) of the macromolecule of interest is obtained from the Protein Data Bank (PDB), a repository of experimentally determined structures of biological macromolecules. In order to generate a structural template from an atomic model, the imaging process of CET needs to be faithfully simulated: The electrostatic potential of the macromolecule is calculated from the coordinates and identities of each atom in the atomic model. Subsequently, the resulting electron density map is convolved with a CTF calculated using the experimental parameters used for tomogram acquisition in the TEM. Finally, the template is low-pass filtered according to the first zero crossing of the CTF and binned to the pixel size of the tomogram. This process is illustrated in figure 2.5 A.

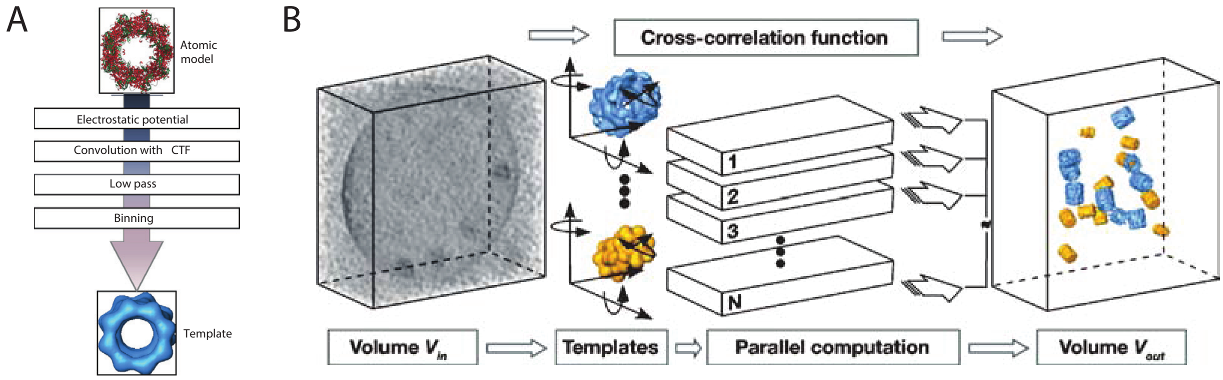

Figure 2.5: Template matching workflow. (A) Diagram for template generation from an atomic model. (B) Depiction of the template matching pipeline. Calculation of the cross-correlation function between the input volume V_{in} and previously prepared structural templates can be accelerated by parallel computation. Peak extraction of the cross-correlation function yields positions and orientations of localized templates in tomographic space V_{out}. Adapted from [Förster et al., 2010; Lucić et al., 2005].

Once a template structure has been prepared, the scoring function describing a similarity measure between the template volume and the tomographic volume can be computed. Conceptually, the scoring function describes the similarity of the template volume at each position in the tomogram volume. Since the orientation of the target macromolecule inside tomographic space is unknown, the orientation space of the template is exhaustively sampled at each position in the tomogram, typically by a set of Euler angles with an angular increment between 5° and 15° (usually 7°). Thus, the scoring function (template matching score) is defined by the orientation that maximizes a similarity measure between the rotated template volume and the subvolume corresponding to each position in the tomogram. The similarity measure used in CET is the locally normalized cross-correlation function [Frangakis et al., 2002], which uses a smoothened binary mask to define the area of interest between the template and tomogram volumes, and applies local normalization before computing the correlation coefficient. Moreover, this measure addresses the missing wedge problem by constraining the correlation calculation to the sampled region in Fourier space.

Since the scoring function provides a quantitative measure of similarity between the template volume and the tomogram, it can be interpreted as the probability of template occurrence in tomographic space. Therefore, scoring function peaks indicate positions in the tomogram where the macromolecule of interest is likely to be found. In practice, the peak extraction process selects coefficient values in descending order, while marking a spherical region around each selected coefficient to keep track of extracted peaks, the spherical radius should approximate that of the template. Each extracted peak indicates the position and orientation of a detected macromolecular complex. Given the low SNR of CET (usually between 0.1 and 0.01) a considerable fraction of the extracted peaks will be false positives. Moreover, estimating the number of peaks to extract is challenging, since the abundance of the target macromolecule inside the tomogram is usually unknown. This issue can be addressed by oversampling the amount of peaks, plotting the histogram corresponding coefficients, and setting a coefficient threshold based on the parameters of a visually estimated Gaussian distribution of true positive peaks [Ortiz et al., 2006]. However, this strategy relies on the assumption that coefficient values of true positives form a

Gaussian distribution, which can be easily segmented from the remaining of the coefficient distribution (i.e. coefficients corresponding to false positives). While this approach has been successful for high-contrast macromolecules such as ribosomes [Ortiz et al., 2006], for other macromolecular complexes the coefficient distributions of false and true positive may significantly overlap, hindering estimation of the true positive Gaussian distribution and appropriate separation of these two classes by coefficient thresholding.

2.3.2 Subtomogram Alignment and Averaging

After a macromolecule of interest has been located inside a tomogram, the corresponding subtomograms can be extracted, aligned and averaged to procedure a higher resolution structure. Assuming that each subtomogram represents the sum of structural signal from the macromolecule of interest plus additive noise, averaging aligned subtomograms will linearly increase the SNR with the number of subtomograms [Hrabe & Förster, 2011], therefore increasing resolution. Moreover, since instances of the target macromolecular complex have different orientations with respect to the missing wedge, isotropic resolution tends to increases. Typical alignment procedures follow and iterative approach: align each subtomogram against a structural reference, average aligned subtomogram to create a new reference, and use the new reference structure for next iteration. Subtomogram alignment not yields higher resolution structures, it also refines the positions and orientations parameters of detected macromolecular complexes, which can be subsequently used to analyze the
spatial distribution of the target complex in tomographic space.

At each alignment iteration, subtomograms need to be aligned against a reference structure. This is achieved by maximizing a similarity score between the reference structure and all subtomogram, where the similarity score is a function of unknown shifts and rotations. Constrained cross-correlation is used as a similarity measure, to account for the missing wedge problem [Förster & Hegerl, 2007]. Expectation-maximization is a commonly used algorithm for subtomogram alignment, as presented in [Hrabe et al., 2012], where shifts are efficiently computed by a fast correlation search based on Fourier transforms [Roseman, 2003]. However, a time-consuming rotation search is performed based on explicit angular sampling in real space (RS). To address this issue, [Chen et al., 2013] proposed an alignment algorithm based on the generalized convolution theorem, which uses spherical harmonics to dramatically accelerate rotation search without sacrificing accuracy. Fast rotation matching (FRM) allows the efficient implementation of a reference free protocol to avoid reference bias, where an initial alignment reference is calculated by randomly rotating and averaging all subtomogram [Scheres et al., 2009]. It is worth mentioning that average resolution is not only restricted by the number of subtomograms, but also by structural flexibility and the intrinsic SNR of the source tomogram.

2.3.3 Subtomogram Classification

Structural heterogeneity, stemming from different conformational states of the imaged macromolecule of interest, is a limiting factor for subtomogram average resolution. To address this issue, subtomogram classification aims at creating structurally homogenous subtomogram sets, thus identifying structural differences and yielding higher resolution class averages. Subtomogram classification also aids in the identification of false positive results from the template matching procedure, which can be subsequently discarded to improve the resolution of the macromolecular structure yielded by subtomogram averaging.

A commonly-used classification approach is constrained principal component analysis (CPCA), where a constrained cross-correlation measure (to account for the missing wedge) between aligned pairs of subtomograms is used to compute a similarity matrix. The similarity matrix is subjected to principal component analysis for dimensionality reduction, and k-means or hierarchical clustering is subsequently applied [Förster et al., 2008]. Alternative approaches based on maximum likelihood [Scheres et al., 2009; Stölken et al., 2011] and multi-reference alignment and classification [G. Frank et al., 2012][G. Frank et al., 2012] have also been developed. Another methodology is auto-focused classification (AC3D)[Chen et al., 2014], it is an iterative multi-reference optimization method, similar to k-means clustering, capable of automatically detecting regions of statistically significant structural differences between class averages, as opposed to CPCA classification, which uses a static mask to focus cross-correlation calculation to a region of interest. At each iteration, masks are computed by calculating a standard deviation map between pairs of class averages, a standard deviation threshold is subsequently used to binarize the map (typically $> 3\sigma$). Masks are then used to focus the calculation of the constrained cross-correlation measure in the following iteration, thereby directly influencing class assignment.

2.3.4 Resolution Measurement

A commonly used method of estimating the resolution of subtomogram averages is Fourier shell correlation (FSC) [Saxton & Baumeister, 1982]. Here, a correlation function between two subtomogram averages is computed in Fourier space; correlation coefficients between volumetric shells of corresponding spatial resolution are computed, yielding a FSC curve. Subsequently, resolution is estimated by the intersection of the FSC curve at specific cutoff values [Penczek, 2010]. In cases where averages are simply computed by splitting the set of collectively aligned subtomograms in half, a cutoff vale of 0.5 is typically used. However, when 'gold-standard' alignment is applied, a cutoff value of 0.143 is commonly used. In gold-standard alignment, half sets are treated separately during the alignment process, avoiding enhancement of noise-based features, which can lead to over-estimating resolution. Cross-resolution FSC estimates the resolution of a subtomogram average by comparing it with an external, high-resolution reference, typically derived from SPA or X-ray crystallography, often using a cutoff value of 0.3.

2.4 Radial Distribution Functions

A particle in the context of CET refers to a mathematical object describing the position and orientation of a detected macromolecular complex in a tomogram. Once a set of particles has been determined, characterizing their local geometric organization can offer valuable insights into the associated biological. The radial distribution function (RDF) is widely used for statistical analysis of the structure of materials. Given a set of N particles in a volume V, the RDF $\rho(r)$, describes how density changes as a function of radial distance from a reference particle, and represents the probability of finding the center of a particle at a distance r from another particle. The RDF can be expressed as [Takeshi & Billinge, 2012]:

$$\rho(r) = \rho_0 g(r) = \frac{1}{N4\pi r^2} \sum_i \sum_j \delta(r - r_{ij}) \tag{2.2}$$

where δ is the Dirac delta function, r_{ij} the center-to-center distance between the i^{th} and j^{th} particles, and $\rho_0 = N/V$ is the number density of the system. The function $g(r)$ is called the pair distribution function. The atomic RDF of a material can be experimentally determined by radiation diffraction measurements (e.g. X-ray crystallography), it is connected by a Fourier transform to the total scattering structure function, also known as the structure factor [Takeshi & Billinge, 2012]. Figure 2.6 shows examples of RDFs for amorphous and crystalline materials.

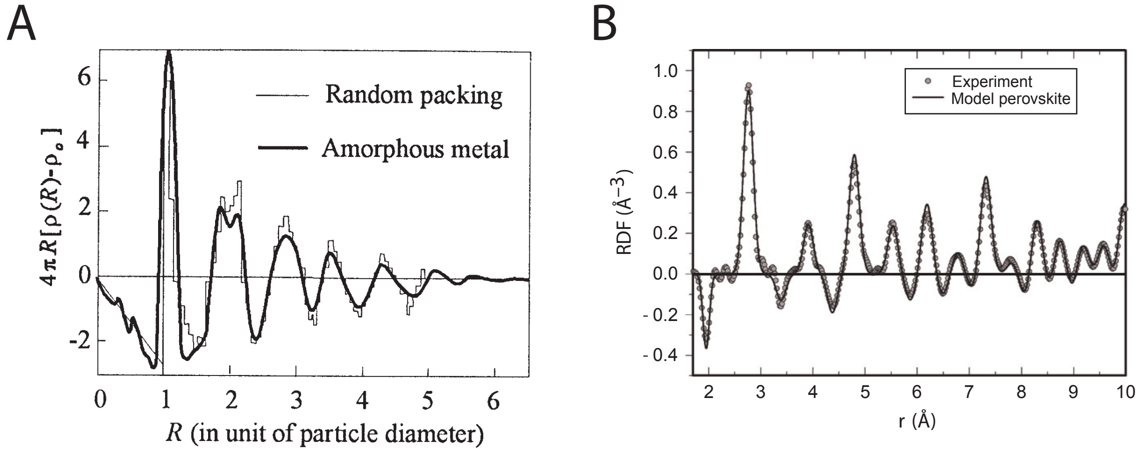

Figure 2.6: Radial distribution functions of crystalline and amorphous materials. (A) RDF of amorphous metals compared to a theoretical RDF of random close packing of particles. (B) Theoretical model and experimentally acquired perovskite RDF. Experimental RDFs were determined by radiation diffraction experiments. Adapted from [Cargill III, 1975; Louca & Takeshi, 1999].

2.5 Topology Graphs

While RDFs are useful to describe the average local organization of a set of particles, they fail to capture many geometric features from underlying topologies of 3D particle distributions. Therefore, it is necessary to use other mathematical tools to describe sets of particles in 3D space. Graph structures are well suited to describe 3D geometry and are widely used for geometry processing [Berner et al., 2008; Mitra et al., 2013; Tevs et al., 2009]. A graph structure is a mathematical abstraction used to define pairwise relations within a set of objects [Bollobás, 1979]. A graph $G = (V, E)$, has a set of nodes, or vertices $V = \{1, \dots, n\}$ and a set of edges E. An edge $(i, j) \in E$ denotes a connection between vertices i and j. In an undirected graph, edges (i, j) and (j, i) are equivalent. However, in a directed graph, an edge tuple also defines directionality, i.e. a directed graph can only be traversed from vertex j to vertex i if there is an edge $(j, i) \in E$. Furthermore, weighted graphs assign a 'weight' value $w_{i,j}$ to each edge (j, i), representing a distance or similarity measure between the connected nodes. Figure 2.7 depicts examples of simple graphs.

Here, the term topology graph is used to denote a graph that describes the local organization of objects in 3D space. Given a set of vectors $\{\mathbf{t}_1, \dots, \mathbf{t}_\mathrm{n}\}$, where each vector $\mathbf{t}_\mathrm{i} = (\mathrm{x_i}, \mathrm{y_i}, \mathrm{z_i})$ represents a 3D point, a topology graph aims to model the spatial neighborhoods of the 3D point set. Local neighborhoods can be modeled by defining a graph with a vertex set $V = \{1, \dots, \mathrm{n}\}$, and connecting vertices corresponding to spatially adjacent 3D vector pairs. There are two commonly used approaches to create this type of graphs [von Luxburg, 2007]: (1) The ϵ-neighborhood graph, where nodes are connected if their pairwise distance is smaller ϵ, i.e. $(i, j) \in E$ if $\left\|\mathbf{t}_\mathrm{i} - \mathbf{t}_\mathrm{j}\right\| < \epsilon$. (2) The k-nearest neighbor graph connects vertex i with vertex j, if point $\mathbf{t}_\mathrm{j}$ is one of the k-nearest neighbors of $\mathbf{t}_\mathrm{i}$. Both strategies allow the creation of directed and undirected graphs.

For the purpose of describing the organization of detected macromolecular complexes in tomographic space, this work defines a topology graph as a ϵ-neighborhood graph. Spatial neighborhoods of ϵ-nm radius for each $\mathbf{t}_\mathrm{i}$, a 3D point associated with the position of a detected complex, can be easily detected using a k-dimensional tree. k-dimensional trees are space partitioning structures [Bentley, 1975] capable of performing efficient near-neighbor queries in sets of points, when the number of points n is much larger than 2^k, where k indicates dimensionality [Indyk et al., 2004]. Given that the amount of detected macromolecules in a tomogram is usually well above 2^3, k-dimensional trees are well suited for this particular application.

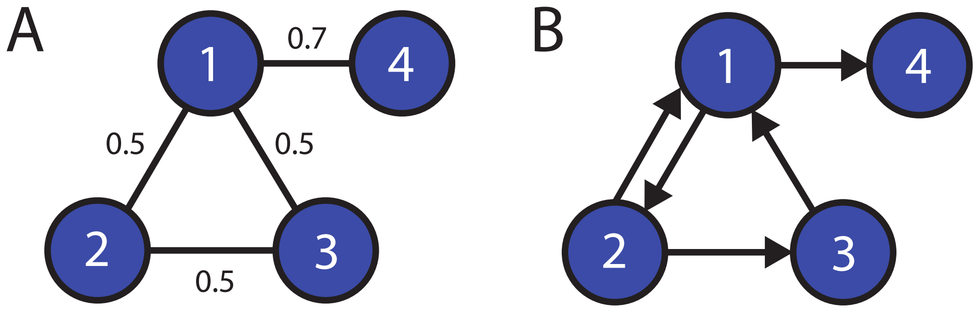

Figure 2.7: Depiction of simple graphs. (A) A weighted, undirected graph $G = (\{1,2,3,4\}, \{(1,2), (1,3), (2,3), (1,4)\})$, with edge weights $w_{1,2} = w_{1,3} = w_{2,3} = 0.5, w_{1,4} = 0.7$. (B) An unweighted, directed graph $G = (\{1,2,3,4\}, \ \{(1,2), (2,1), (2,3), (3,1), (1,4)\})$.

2.5.1 Markov Random Fields

Probabilistic graphical models are useful tools for visualizing the structure of probabilistic models, detecting statistical dependence, and graphically modifying a model with implicit mathematical correspondence [Bishop, 2006]. Graphical models are graph structures where each vertex represents a random variable, and edges denote probabilistic relations between variable pairs. The graph describes how the joint distribution over all random variables can be factorized into factors that depend only subsets of random variables. The two prominent types of graphical modes are Bayesian networks and Markov random fields (MRF), while Bayesian networks are used to express causal relations, MRF are capable of describing soft constrains between random variables [Bishop, 2006]. MRF are undirected graphs where edges indicate probabilistic dependence, stated specifically, the conditional probability of a random variable is dependent only on the variables in its Markov blanket, i.e. the set of neighboring nodes directly connected by an edge.

The notion of a maximal clique is useful to appropriately define the joint probability distribution of a MRF. A clique is a subset of vertices in a graph, such that all pairs of vertices are connected by an edge, i.e. a fully connected subgraph. Furthermore, a maximal clique is a clique, which cannot include anymore nodes without ceasing to be a clique. By defining a maximal clique in a MRF $G = (V, E)$ as $C \subseteq V$, the set of random variables in the clique as $\mathbf{x}_c$, and an associated potential function ψ_c over $\mathbf{x}_c$, the Hammersley-Clifford theorem [Clifford, 1990] can be used to express the joint probability distribution $p(\mathbf{x})$ of the MRF, as the product of clique potential functions [Bishop, 2006]:

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_c(\mathbf{x}_c) \tag{2.3}$$

$$Z = \sum_{\mathbf{x}}\prod_{C}\psi_c(\mathbf{x}_c) \tag{2.4}$$

where Z is the so-called partition function, ensuring that $p(\mathbf{x})$ is correctly normalized. Potential functions $\psi_c(\mathbf{x}_c)$ should be strictly positive. They are usually expressed as Gibbs distributions [Bishop, 2006]:

$$\psi_c(\mathbf{x}_c) = \exp(-E(\mathbf{x}_c)) \tag{2.5}$$

Here $E(\mathbf{x}_c)$ is the energy function associated with the set of random variables of clique C. Potential functions do not have specific probabilistic interpretations, while this provides a large degree of flexibility for modeling problems, they require appropriate motivation, often in terms of compatible configurations of neighboring variables [Bishop, 2006; Blake et al., 2011].

2.5.2 Loopy Belief Propagation

Belief propagation is an algorithm for probabilistic inference, where nodes in a graphical model pass local messages along connecting edges. The two main variants of belief propagation, the sum-product and max-product algorithms, aim to marginalize the joint probability distribution and estimate a maximum a posteriori (MAP) solution, respectively [Bishop, 2006; Blake et al., 2011]. This algorithm was initially proposed for tree structures, i.e. graphs without cycles, where it was proven to converge to the correct solution [Pearl, 1988]. However, the application of this algorithm to graphs with loops was proposed by [B. Frey & MacKay, 1998]. This was mainly possible since the message-passing rules of the algorithm operate locally within the graph. However, the introduction of cycles may allow information to circulate indefinitely inside the graph, leading to oscillation, and preventing the algorithm to converge to a stable solution [Bishop, 2006; Pearl, 1988]. Despite theoretical reservations, in particular, the well-known NP-hardness of probabilistic inference in arbitrary graphs [Cooper, 1990; Shimony, 1994], loopy belief propagation has been applied successfully on a variety of computer-vision problems [Blake et al., 2011; Freeman et al., 2000; B. J. Frey et al., 2001]. Moreover, loopy belief propagation has also been shown to provide exceedingly good empirical results in error-correcting codes [Berrou et al., 1993; Kschischang & Frey, 1998; McEliece et al., 1998]. It is worth mentioning that the optimality of loopy belief propagation and the structural characteristics of the underlying graph that allow convergence to reasonable approximations are still being investigated [Murphy et al., 1999; Weiss & Freeman, 2001a].

2.7 Photosynthesis

Photosynthesis is the biological mechanism responsible for the production of chemical energy and biomass from atmospheric CO_2, solar energy and water. This process can be classified into light-dependent, and light-independent reactions. While light-dependent reactions generate chemical energy in the form of ATP and NADPH molecules and O_2 from H_2O and light, the subsequent light-independent reactions use this chemical energy to fix inorganic CO_2 into organic carbon in the Calvin-Benson-Bassham cycle [Andersson & Backlund, 2008; Miziorko & Lorimer, 1983]. The enzyme Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) has a central role in the Calvin-Benson-Bassham pathway because it binds CO_2 to ribulose-1,5-bisphosphate (RuBP) to yield two molecules of 3-phosphoglycerate (3PG), the following reactions use ATP and NADPH to produce glyceraldehyde-3-phosphate (G3P), ADP, $NADP^+$, and inorganic phosphate (P_i). Finally, 3PG can be used to regenerate RuBP in an ATP-driven process, or for the production of biomolecules, e.g. sugars and amino acids. However, RuBisCO is a notoriously inefficient enzyme, with a low catalytic rate of $\sim 3 - 10$ CO_2 molecules per second [Ellis, 2010], and a competing oxygenase reaction that produces 2-phosphoglycolate. To compensate for the oxygenase activity of RuBisCO, the energy-consuming pathway known as photorespiration recycles 2-phosphoglycolate back to 3PG, leading to the loss of previously fixed carbon in the from CO_2 [Hartman & Harpel, 1994]. The most common form of RuBisCO found in algae, cyanobacteria and plants has a hexadecameric structure. It consists of eight large ($\sim$50 kDa) and eight small subunits ($\sim$15 kDa), forming a $\sim$520 kDa holoenzyme with a diameter of $\sim$11 nm [Andersson & Backlund, 2008; Hauser et al., 2015].

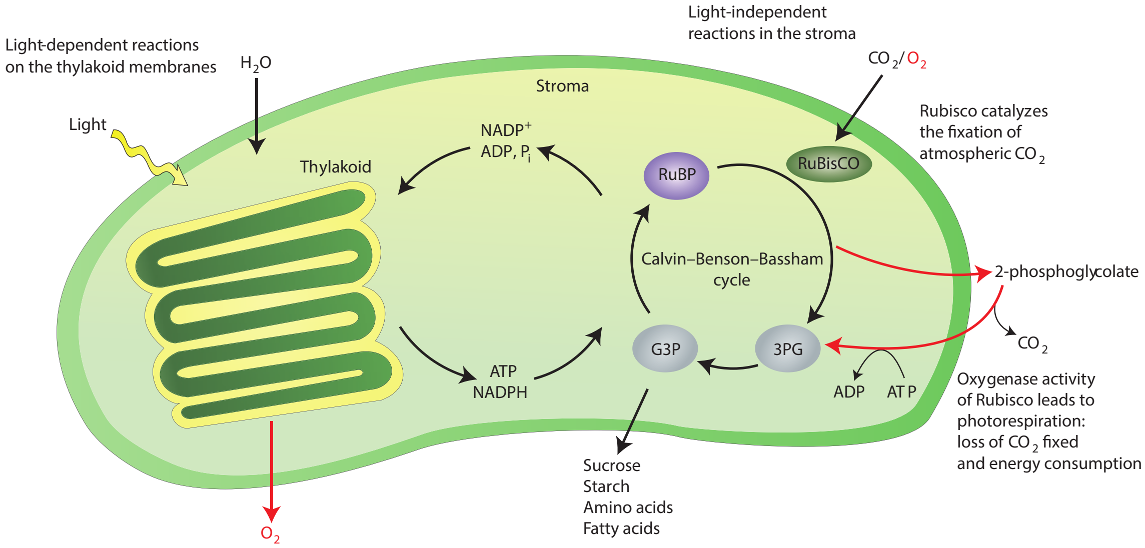

Figure 2.8: Diagram of the photosynthetic process. Overview of the light-dependent and light-independent reactions, depicting the relation between the Calvin-Benson-Bassham cycle and the photorespiration pathway. Adapted from [Hauser et al., 2015].

2.7.1 The Pyrenoid of C. Reinhardtii

Photosynthetic organisms have developed strategies to cope with the enzymatic limitation of RuBisCO. While most organism produce large amounts of RuBisCO to increase the total enzymatic activity [Ellis, 1979], microorganisms have augmented their photosynthetic process by biophysical carbon concentration mechanisms (CCM). A salient feature of CCMs is the packing of RuBisCO in micro-compartments, such as the pyrenoid in eukaryotic algae and carboxysomes in cyanobacteria [Meyer et al., 2016]. The pyrenoid of the unicellular green alga *C. reinhardtii* is a spherical organelle in the chloroplast stroma, which usually contains $\sim 90 - 95\%$ of the cellular RuBisCO. The pyrenoid matrix is surrounded by starch sheaths and traversed by tubules of thylakoid membranes [Engel et al., 2015; Meyer et al., 2016]. Mass spectrometry analysis of pyrenoid proteins after inducing the CCM by reduction of CO_2 in the medium indicate that the large majority of the protein found in the pyrenoid matrix consists of RuBisCO large and small subunits, RuBisCO activase, and a $\sim$33 kDa protein labeled 'Essential Pyrenoid Component 1' (EPYC1) because it is necessary for pyrenoid formation in *C. reinhardtii* cells [Mackinder et al., 2016].

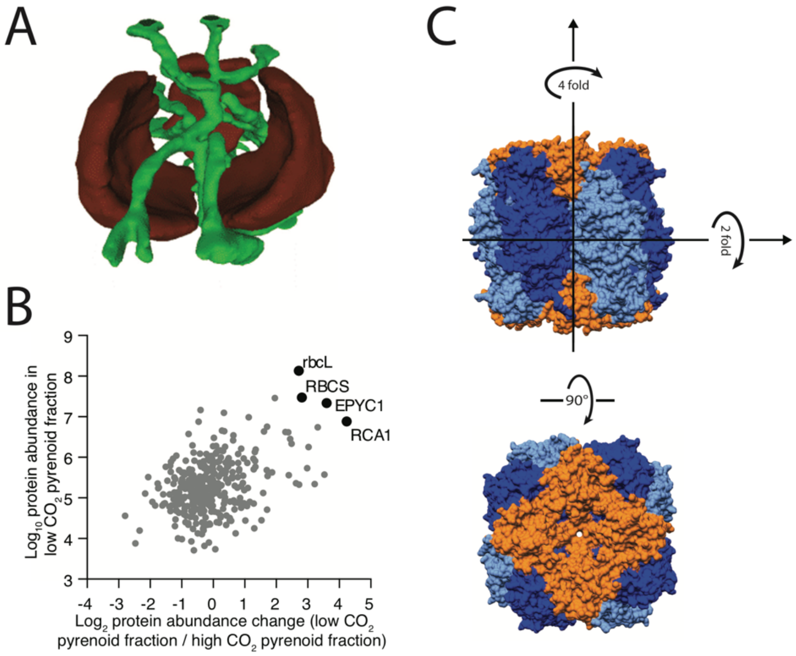
A
C
4 fold
2 fold
90°
B
rbcL
RBCS
EPYC1
RCA1
9
8
7
6
5
4
3
-4 -3 -2 -1 0 1 2 3 4 5
Log_{10} protein abundance in low CO_2 pyrenoid fraction
Log_2 protein abundance change (low CO_2 pyrenoid fraction / high CO_2 pyrenoid fraction)

Figure 2.9: Structure and protein content of the *C. reinhardtii* pyrenoid. (A) Scanning-electron-micrograph reconstruction of a pyrenoid, depicting starch sheaths (red) and thylakoid tubules (green). (B) Mass spectrometry analysis of pyrenoid proteins from *C. reinhardtii* cells after CCM induction, showing a significantly higher protein abundance of RuBisCO large (rbcL) and small (RBCS) subunits, RuBisCO activase (RCA1), and EPYC1 protein. (C) Surface model of the RuBisCO holoenzyme from *C. reinhardtii*, showing the D_4 symmetry of the hexadecameric arrangement of large (blue and light blue) and small (orange) subunits, model derived from [Taylor et al., 2001]. Adapted from [Mackinder et al., 2016; Meyer et al., 2016].

2.8 Protein Synthesis

The ribosome is the macromolecule responsible of translating mRNA into polypeptide chains. Ribosomes are large macromolecular machines, with a molecular mass of 2.5 MDa and 4 MDa in bacteria and eukaryotes respectively. The ribosome mass is approximately one third protein and two thirds RNA [Tissières, 1974]. Ribosomes are composed of one large and one small subunit. The large subunit is responsible for the enzymatic function (addition of amino acids to the nascent peptide chain), while the small subunit ensures fidelity in the mRNA translation process. In bacteria, the large subunit is referred to as 50S and the small subunit is labeled 30S. In the case of eukaryotic ribosomes, the large and small subunits of are known as 60S and 40S, respectively [Voorhees & Ramakrishnan, 2013]. The large and small subunits assemble on the mRNA to form the bacterial 70S and eukaryotic 80S ribosome. During translation, a ribosome creates a protein sequence by binding amino acids from transfer RNAs (tRNAs) in the order dictated by the mRNA molecule, in an iterative subprocess called elongation. tRNA molecules bind to the ribosome and sequentially move from the aminoacyl (A) site to the peptidyl (P) site, and finally to the exit (E) site before detaching from the ribosome complex.

The translation process has four main phases: (1) Initiation; the ribosome positions itself over the start codon of the mRNA sequence. (2) Elongation; ribosome sequentially adds amino acids to the polypeptide chain. (3) Release; once the ribosome reaches the mRNA stop codon, the nascent peptide is detached from the ribosome. (4) Recycling; the ribosome subunits disassociate and prepare for another protein synthesis cycle. Thus, the core of the translation process happens during the elongation stage, a highly conserved cycle between bacterial and eukaryotic organisms [Voorhees & Ramakrishnan, 2013].

The elongation cycle has three major steps [Voorhees & Ramakrishnan, 2013]: After initiation, the ribosomal A site is empty and the P site is occupied. In the subsequent (1) decoding step, an aminoacyl tRNA (aa-tRNA) is delivered to the A site by the elongation factor Tu (EF-Tu). Once the cognate tRNA binds to the ribosome, the EF-Tu factor disassociates and peptide bond formation occurs, adding a new amino acid to the nascent protein, this step is known as (2) accommodation. Finally, in the (3) translocation step, tRNAs move from the A to the P site, and from the P site to the E site. Subsequently, the elongation factor G (EF-G) triggers a single-codon shift of the mRNA molecule. Finally, EF-G dissociates and the ribosome is ready for another iteration of the elongation cycle.

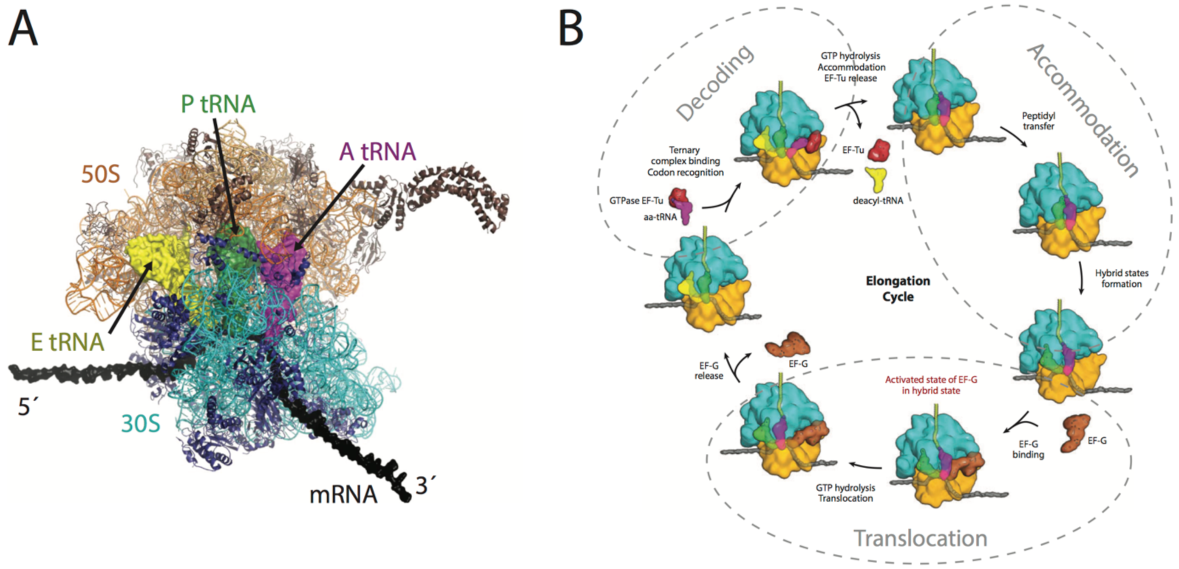

Figure 2.10: Structure of the bacterial ribosome and overview of the elongation cycle. (A) Top view of the 70S ribosome, depicting the mRNA (black), and the A (purple), P (green) and E (yellow) tRNAs. (B) Depiction of the elongation cycle, marking decoding, accommodation and translocation steps. Adapted from [Schmeing & Ramakrishnan, 2009].

2.8.1 Polysomes

Typically, an mRNA molecule is simultaneously translated by many ribosomes. Ensembles of various ribosomes translating a single mRNA molecule are called Polysomes. Ribosomes within a polysome are expected to be actively translating an mRNA sequence, and thus be engaged in the elongation cycle [Rich et al., 1963]. Furthermore, ribosomes in polysomic ensembles have been observed to adopt specific supramolecular arrangements, yielding characteristic polysome structures. In lysates of *E. coli* cells polysome displayed characteristic planar and helical organizations [F. Brandt et al., 2009]. Helical, planar and spiral-like topologies were also observed in the cytosol of intact human cells [F. Brandt et al., 2010]. The remarkably conserved arrangements of cytosolic polysome in bacteria and humans suggest a universal evolutionary pressure on polysome topology [F. Brandt et al., 2010]. It has been suggested that polysome topologies contribute to the prevention of non-native interactions between nascent polypeptides [F. Brandt et al., 2009], and the optimization of the folding process by providing ample space for the folding machinery to access nascent proteins and shield them from the crowded environment of the cytosol [Hartl & Hayer-Hartl, 2009]. Another interesting topological feature of cytosolic polysomes is the sequestering of the mRNA into the core of the supramolecular arrangement, possibly shielding the mRNA molecule from RNases during elongation slowdown [Buchan & Stansfield, 2007; Sivan et al., 2007].

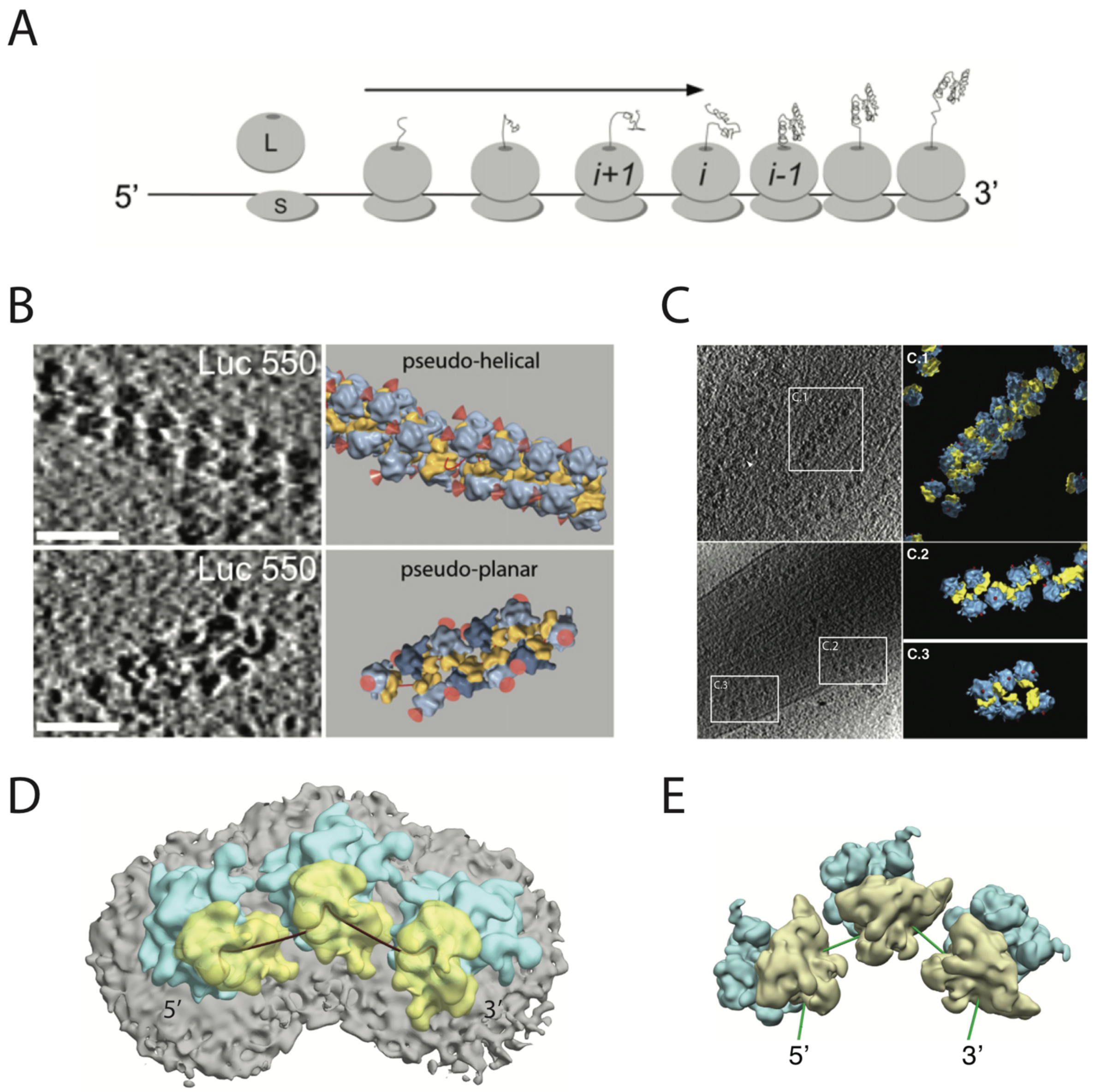

Figure 2.11: Polysome topologies observed by CET. (A) Diagram of a polysome structure, small (s) and large (L) ribosomal subunits assemble on the mRNA, translating from the 5’ end to the 3’end of the mRNA molecule. Neighbors of the i^{th} ribosome in the polysomic sequence are termed *i+1* and *i-1* at position on the 5’ and 3’ sides, respectively. (B) Planar and helical topologies of cytosolic bacterial polysomes, tomographic cross-sections (left) and surface models (right) of observed polysomes, large subunits in blue and light blue, small subunits in yellow, red cones point to the peptide exits, scale bars: 50 nm. (C) Helical (C.1) and planar (C.2,

C.3) topologies of cytosolic polysomes in human hells, tomographic cross-sections (left) and surface models (right) of observed polysomes, large subunits in blue, small subunits in yellow, red cones point to the peptide exits. (D) Model of the local organization of ER-associated polysomes, putative mRNA path in red. (E) Model of the local structure of membrane-bound mitochondrial polysomes, putative mRNA path in green. Adapted from [F. Brandt et al., 2009, 2010; Pfeffer et al., 2012; Pfeffer, Woellhaf, et al., 2015].

On the surface of membranes extracted from ER of canine pancreatic cells, polysomes preferred curved and spiral-like topologies [Pfeffer et al., 2012]. Membrane-bound polysomes of yeast mitoribosomes display similar topologies, further supporting the hypothesis of a universal evolutionary pressure on polysome arguments, as both ER-associated and membrane-bound mitochondrial polysomes face the same topological requirements for threading an mRNA molecule between polysomic neighbors on the membrane surface, while simultaneously translocating the nascent polypeptide through the membrane [Pfeffer et al., 2015].

3. Materials and Methods

3.1 Implementation of algorithms

Tomogram reconstruction: All tomograms were reconstructed using weighted backprojection. Two software packages were used for tomographic reconstruction, the Matlab-based TOM toolbox for tomography [Nickell et al., 2005] and IMOD [Kremer et al., 1996].

Tomogram visualization and segmentation: Visualization of tomograms was performed with the UCSF Chimera package [Pettersen et al., 2004]. This software was also used for fitting atomic models into subtomogram averages. The Amira software (FEI visualization Science Group) was used for manual segmentation of tomographic volumes.

Template Matching: Template generation from atomic models was performed as described in section 2.3.1, specifically, the protocol described in [Förster et al., 2010] using the TOM toolbox [Nickell et al., 2005]. The template matching procedure was applied on binned tomograms using the PyTom software [Chen et al., 2012; Hrabe et al., 2012] a tomography toolbox implemented in the C++ and python.

Subtomogram Alignment and Classification: Subtomograms were aligned in PyTom, using the FRM method [Chen et al., 2013] (in Fourier space) and the RS expectation-maximization method [Hrabe et al., 2012]. Both methods have a built-in option for gold-standard alignment. Classification of subtomograms by CPCA used PyTom to calculate similarity matrices, and Matlab for principal component analysis, k-means clustering and hierarchical clustering. On the other hand, AC3D classification is fully implemented in PyTom. PyTom provides a parallelized implementation of alignment and classification algorithms using the python-based message passing interface, substantially reducing computation time. PyTom-based alignment and classification tasks were performed on a computer cluster, each task used a number of computer nodes ranging from 64 to 256, each with 16 CPUs and 516 GB of RAM.

Three-Dimensional Range Queries: Range queries on sets of 3D points for the identification of particle neighborhoods (chapter 5) and the construction of topology graphs (chapter 6) used the SciPy (www.scipy.org) implementation of the k-dimensional tree [Bentley, 1975]. SciPy is a widely used python-based library for scientific analysis, which uses NumPy (www.numpy.org) for efficient linear algebra calculations.

Graph structures and algorithms: The NetworkX library (networkx.github.io) was used to generate graph structures and incorporate graph-based algorithms (e.g. Dijkstra's shortest path algorithm [Dijkstra, 1959]) into the polysome detection method. NetworkX is a python library for graph theory, it provides data structures for different types of graphs and implements a large number of algorithms.

Gaussian mixture models: Fitting Gaussian mixture models was performed by expectation-maximization using the scikit-learn library (scikit-learn.org). Scikit-learn is a python library for machine learning with an extensive clustering module. 3D Gaussian components were fitted with unconstrained covariance matrices.

Clustering 3D vector distributions was performed by fitting Gaussian mixture models. First, a mixture model was fitted to a vector distribution, the cluster label for each data point was then obtained by selecting the Gaussian component with the largest likelihood value.

Quaternion analysis: The clustering module of the scikit-learn library was used for spectral clustering [von Luxburg, 2007] of quaternions, with a python implementation of algorithm 5.1 to calculate similarity matrices. Additionally, the Bingham Statistics Library (github.com/sebastianriedel/bingham) was used to fit Bingham distributions, this library contains C and Matlab implementations of the Bingham distribution operations described in [Glover & Kaelbling, 2013]. The C implementation of the Bingham Statistics Library was accessed as a python module by using function wrappers.

Loopy belief propagation: An implementation of a Markov random field with functionality for loopy belief propagation was written in python. This module used NetworkX, PyTom, SciPy (for a k-dimensional tree) and NumPy as dependencies.

3.2 Dataset Acquisition and Processing

3.2.1 Tomograms of *C. reinhardtii* Pyrenoids

Acquisition: Pyrenoid lamellas of plunge-frozen *C. reinhardtii* cells were milled using a dual-beam instrument for FIB milling and scanning electron microscopy (Quanta 3D FEG, FEI). Lamellas were milled as described in [Rigort et al., 2012] and subjected to CET. A dataset of nine lamella tomograms was recorded on a Titan Krios microscope (FEI, Eindhoven, NL) equipped with an energy filter and a K2 direct electron detector (Gatan, Pleasanton, USA). A tilt range of -60° to 60° with an increment of 2° was used, defocus ranging from 5 to 6 µm, and a pixel size of 0.34 nm. Tilt-series alignment was performed by patch tracking with the IMOD software [Kremer et al., 1996; Mastronarde, 1997]. IMOD was also used for tomographic reconstruction.

Subtomogram analysis: For template matching, a structural template was derived from a *C. reinhardtii* RuBisCO structure obtained by X-ray crystallography [Taylor et al., 2001] (section 5.3). For each tomogram, the cross-correlation function produced by template matching was filtered using a binary mask, which covered the region in the tomogram where the lamella was found, removing cross-correlation values in the periphery. These binary masks were generated manually. Extraction of RuBisCO particles from the filtered cross-correlation function was performed by peak extraction with a mask radius of 8.2 nm. To accomplish exhaustive peak extraction, the parameter indicating the number of peaks to extract was set to a large number, allowing the peak extraction procedure to finish only when the cross-correlation function was completely set to zero.

The IMOD software [Kremer et al., 1996] was used for CTF correction of subtomograms. FRM alignment of data binned to a pixel size of 0.68 nm (section 5.4) did not impose D_4 symmetry and the maximal number of iterations was set to 5. It considered a maximal spatial resolution of 2.12 nm to reduce the influence of noise and provide an initial, coarse-grained refinement of position and orientation parameters. AC3D classification (section 5.5) was restricted to a maximal spatial resolution of 3.83 nm, number of classes was set to 15, number of iterations was set to 10 and the standard deviation threshold for difference maps was set to 3σ. After subtomogram classification, alignment of unbinned data (section 5.6) was separated in two steps. The initial FRM alignment step was performed with the maximal number of iterations set to 5, it was restricted to a spatial resolution of 2.12 nm in the first iteration, to provide a global solution for particle orientations without significant noise bias. In the following RS alignment step, the number of iterations was set to 10, the initial angular increment to 3° and angular shells to 3. Both FRM and RS alignment of unbinned data imposed D_4 symmetry.

Tomogram handedness: Control tomograms from adjacent, ribosomes-rich regions were acquired and reconstructed using the same parameters. Tomogram handedness was tested by applying template matching with mirrored templates of ribosome structures. The distribution of template matching scores from extracted peaks was compared with the distribution yielded with the unmirrored template, since scores computed using a template with incorrect handedness tends to display significantly reduced values [Ortiz et al., 2006].

Local geometric analysis: Range queries for computation of the RDF (section 5.7.1) used a radius parameter of 40 nm, while the radius parameter for inspection of the first near-neighbor shell was set to 21 nm. Once range queries identified particle neighborhoods, fine-grained dissection of the first near-neighbor shell was performed by filtering the distribution of center-to-center vectors by vector magnitude, to allow selection of radial shells within the 21 nm radius.

Gaussian mixture model fitting on the distribution of center-to-center vectors (sections 5.7.2 – 5.7.5) was performed as follows: D_4 symmetry was applied to the 3D vector distribution of the chosen radial shell and the number of Gaussian components was estimated by visual inspection. The Gaussian mixture model was fitted and symmetrically redundant clusters were merged. Likelihood cutoffs for Gaussian components were set to values corresponding to 1σ distance from the mean.

For rotation clustering, D_4 symmetry was applied to the set of quaternions and the corresponding distribution of rotated 4-fold axes (figure 5.7 B) was plotted to estimate the number of clusters. After spectral clustering [von Luxburg, 2007], symmetrically equivalent clusters were merged by inspecting rotated RuBisCO templates, as dictated by the mode of the corresponding Bingham fit. Likelihood cutoffs for Bingham distributions were set within a range of 70 to 80% the likelihood value of the quaternion corresponding to the Bingham mode.

3.2.2 Tomograms for Polysome Detection

3.2.2.1 Experimental Tomograms of *E. coli* Lysate

Acquisition: Six tomograms of *E. coli* spheroplast lysates from the dataset published in [F. Brandt et al., 2009] were used for detection of cytosolic polysomes. Each tomogram was recorded using a tilt range of -60° to 60° and an increment of 3°. The tilt series was acquired with a defocus of 3 µm and a pixel size of 0.28 nm. The TOM toolbox [Nickell et al., 2005] was used for tilt-series alignment (by tracking gold markers) and for tomographic reconstruction.

Subtomogram analysis: This dataset was processed by [F. Brandt et al., 2009]. For template matching, tomograms were binned to a pixel size of 2.24 nm and the template was generated from an atomic model of the 70S ribosome (PDB entries 2AW7 and 2AWB [Schuwirth, 2005]), which was lowpass filtered to 4 nm resolution. Positions and orientations from 1,500 peaks of the cross-correlation functions were extracted. A large number of peaks was extracted to fully cover the true positive class of ribosomes. While a large number of false positives might have been introduced, high sensitivity was required to obtain true polysome structures. This dataset was binned to a pixel size of 2.24 nm and 1.12 nm for subsequent processing (details in [F. Brandt et al., 2009]). This analysis produced a set of ribosome particles, from which polysomes in characteristic pseudo-helical and pseudo-planar arrangements were manually identified. The manually identified polysome were later used as ground truth knowledge to evaluate the performance of the polysome detection method.

3.2.2.2 Simulated Tomograms of *E. coli* lysate

Simulation: Five tomograms of 512x512x256 voxels were simulated, each containing three pseudo-helical and three pseudo-planar polysomes. Polysome arrangements were simulated by successively adding 70S ribosomes densities (simulated from PDB entries 2AW7 and 2AWB [Schuwirth, 2005]) in t-t or t-b configurations [F. Brandt et al., 2009], up to a randomly chosen length between 10 and 15 ribosomes. Additionally, 3 to 5 gold beads were added to each volume, as well as 500 single ribosome densities in random positions and orientations (monosomes). Simulation of electron tomography followed the method described in [Beck et al., 2009; Förster et al., 2008] using the same acquisition parameters as in the analogous experimental dataset (section 3.2.2.1): Once simulated electron densities of 70S ribosomes and gold beads were added to a volume with pixel size of 0.28 nm, the volume was low-passed filtered and projected to create a tilt series of -60° to 60°, with 3° angular increments. Gaussian noise was added to each projection, a simulated 2D CTF function with 3 µm defocus was applied, and MTF noise was added. Tomographic reconstruction was performed with the TOM toolbox [Nickell et al., 2005].

Subtomogram analysis: Tomograms were subjected to template matching with a 70S ribosome template derived from [Schuwirth, 2005] and an angular increment of 7°. For each tomogram, 1,000 template matching peaks were extracted to achieve high sensitivity for true positive ribosome particles. Based on ground truth knowledge, ribosome particles were then assigned to polysome sets for subsequent evaluation of the polysome detection method.

3.2.2.3 Tomograms of Microsomes Derived from Rough ER

Acquisition: A dataset of 18 tomograms of a preparation of rough microsomes derived from ER of mouse myeloma cells was collected at 8 µm defocus and a pixel size of 0.288 nm, using a tilt range of -60° to 60° and an increment of 3°. Tilt-series alignment was performed with the TOM toolbox [Nickell et al., 2005] by manual detection of fiducial markes. The TOM toolbox was also used for tomographic reconstruction.

Subtomogram analysis: In preparation for template matching, the tilt-series was binned to a pixel size of 2.304 nm and subsequently reconstructed. A SPA reconstruction of a canine 80S ribosome [Chandramouli et al., 2008] was used as a template, lowpass filtered to 4 nm resolution. A large number of peaks was extracted (approximately 2,000 per tomogram) to ensure high sensitivity, since false negatives would hinder polysome detection. A total of 32,487 peaks were extracted from 18 tomograms. Subtomograms binned to a pixel size of 0.576 nm were subjected to RS alignment, where the number of iterations was set to 5, initial angular increment to 3° and angular shells to 3. For CPCA classification, subtomograms were lowpass filtered to 2.9 nm, 5 eigenvectors were used and the number of classes was set to 320. The 320 classes were then merged by hierarchical clustering, using constrained cross-correlation as distance measure, yielding 3 classes: 'noise', 'putatively active' and 'putatively inactive' ribosomes. Finally, the 'putatively active' and 'putatively inactive' classes were merged into a positive class containing 25,683 ribosome particles, the remaining 6,804 particles were labeled as false positives.

3.2.2.4 Subsequent Processing

Tomogram handedness: Tomogram handedness was tested by subjecting tomograms to template matching with mirrored ribosome templates [Ortiz et al., 2006], as described in section 3.2.1.

Local models of polysomes: After Gaussian mixture models were fitted to distributions of mRNA exit-to-entry vectors, likelihood cutoffs for Gaussian components were set to values corresponding to 2σ distance from the mean, allowing extraction of relative rotations corresponding to the 5' cluster.

Polysome detection: The radial parameter r_{max} for range queries used to construct topological graphs (section 6.4.1) was set to approximately twice the diameter of the ribosome. For topological graphs of 70S ribosomes this parameter was set to 40 nm, while 80S ribosome graphs used an r_{max} of 60 nm. Furthermore, the number of iterations for loopy belief propagation was set to 5.

Subtomogram analysis of ER-associated and cytosolic mammalian polysomes: CPCA classification of the monosome class (figure 6.10) was applied with a mask covering only the ER membrane, subtomograms were lowpass filtered to 5.23 nm resolution, 10 eigenvectors were used and the number of classes was set to 2.

4. Local Geometric Analysis

4.1 Introduction

Once a macromolecular complex of interest has been located in tomographic space, the set of positions and orientations of detected complex particles can be refined by subtomogram analysis, precise determination of these geometric parameters enables statistical analysis of the local organization of complexes particles. Characterization of the local spatial distribution of macromolecular complexes can offer mechanistic insights into biological processes, a major objective in the visual proteomics field [Förster et al., 2010; Nickell et al., 2006]. Statistical analysis of relative positioning and orientation of adjacent mammalian ribosomes in cryo-tomograms of intact cells and rough ER microsomal preparations enabled geometric characterization of pairwise arrangements from neighboring ribosomes, responsible for the characteristic supramolecular organization of polysomes. Moreover, it allowed for a quantitative description of variability in polysome topologies [F. Brandt et al., 2010; Pfeffer et al., 2012].

Local geometric analysis of localized macromolecular complexes can be used for descriptive or predictive purposes. A model of the geometric configuration of macromolecular neighborhoods can be created based on a set of characteristic arrangements of particle pairs. Statistical analysis of relative position vectors and relative orientations (geometric features) allows the proposed models to be intrinsically quantitative. Moreover, supramolecular structures of biological interest can be identified by detecting extrapolations of characteristic particle pair configurations, since these local arrangements can be understood as geometric templates. Statistical description of geometric feature clusters associated with characteristic configurations of particle pairs, allows the design of probabilistic methods for supramolecular structure detection, capable of coping with varying degrees of local flexibility.

In this chapter, a strategy is presented for geometric analysis of local neighborhoods from the spatial distribution of detected particles in tomographic space: First, the local neighborhood of each particle is identified, followed by geometric feature calculation, describing the 3D organization of neighboring particles. Finally, the distribution of local geometric features is subjected to cluster analysis, and statistical models are used to describe the distribution. The main objective is to identify characteristic geometric arrangements of particle pairs.

4.2 Local Geometric Analysis Workflow

Here, the proposed strategy for local geometric analysis is formally described. Given a set of particles $P = \{p_1, \ldots, p_n\}$, of a macromolecular complex, as localized by template matching, where each particle $p_i = (\mathbf{t_i}, \mathrm{R_i})$ is described by a position $\mathbf{t_i} \in \mathbb{R}^3$ (3D vector) and orientation $\mathrm{R_i} \in SO(3)$ (rotation matrix) of the detected macromolecular complex in the tomographic coordinate system. The objective is to identify predominant geometric configurations of spatially adjacent pairs of particles, by analyzing the distribution of relative (pairwise) 3D position vectors, and relative rotations (geometric features) derived from the local neighborhood of each particle in tomographic space.

The proposed local geometric analysis is performed by the following steps: (1) Identification of particle neighborhoods and geometric feature computation. (2) Statistical cluster analysis and extraction of mode peaks from the 3D position vector distribution. (3) A position peak is selected for further inspection of the associated distribution of relative rotations, finally, rotation clustering and cluster peak extraction is performed. Once a rotation mode has been extracted from the geometric feature distribution of a previously selected 3D position peak, the corresponding geometric configuration of particle pairs can be fully characterized. Figure 4.1 exemplifies the proposed workflow for a set of RuBisCO complex particles, detected in a pyrenoid tomogram of a *C. reinhardtii* cell.

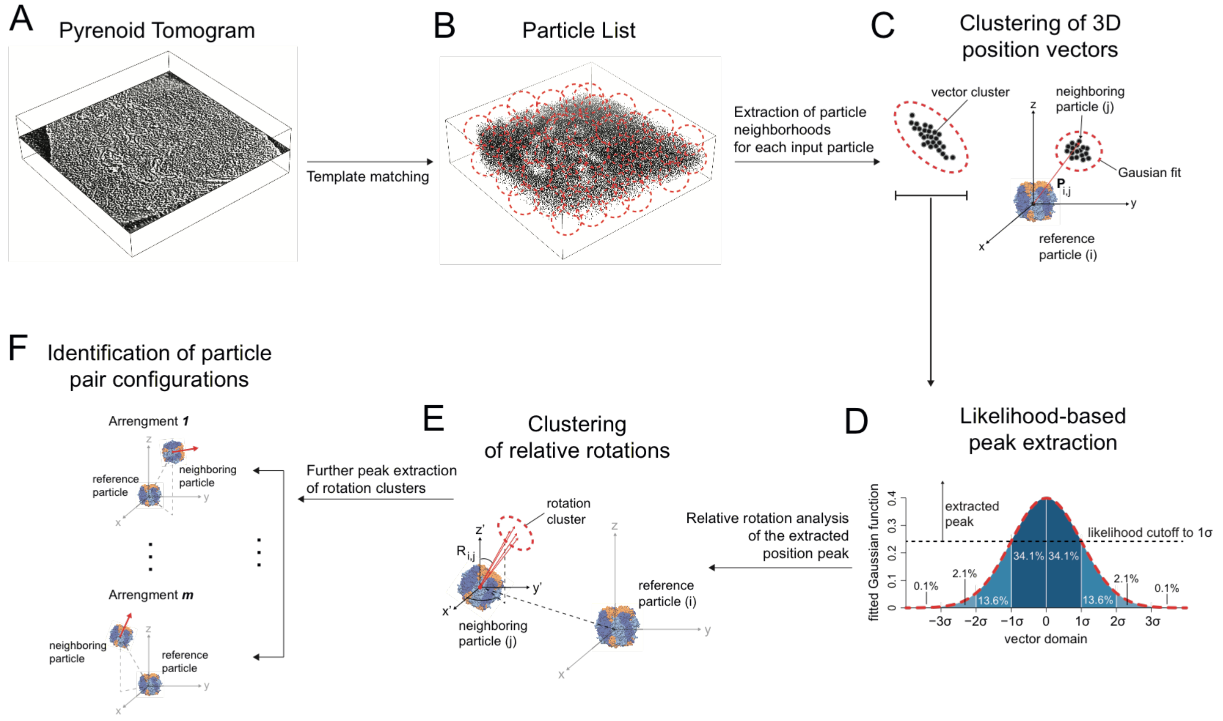

Figure 4.1: Local geometric analysis workflow, exemplified on an input set of RuBisCO complex particles detected in a pyrenoid tomogram of a *C. reinhardtii* cell. (A) Source pyrenoid tomogram of a *C. reinhardtii* cell. (B) A set of RuBisCO complex particles yielded by template matching on the source tomogram (3D positions rendered as black spheres), detected particle neighborhoods are depicted as red circles. (C-D) Cluster analysis and peak extraction of 3D position vectors. (D) 1D Gaussian analogy of the peak extraction strategy. Since the depicted likelihood cutoff is set to a value corresponding to 1σ distance from the mean, the extracted peak approaches 68% of the cluster. (E) Cluster analysis of relative rotation distributions associated to position peaks. (F) Peak extraction of rotation clusters yields subpopulations of geometric features describing specific particle pair configurations. RuBisCO complex model from [Taylor et al., 2001], large subunits: blue and light blue, small subunit: orange.

(1) Identification of particle neighborhoods and geometric feature computation. Initially, the set of input particle positions $\{\mathbf{t_1},\ldots,\mathbf{t_n}\}$ is used to detect the radial neighborhood around each particle $p_i \in P$. Once particle neighborhoods have been identified, relative 3D position vectors $\mathbf{p_{i,j}} \in \mathbb{R}^{\mathbf{3}}$ and relative rotations $\mathrm{R_{i,j}} \in SO(3)$ of adjacent particle pairs $(p_i, p_j \in P)$ are computed to generate the geometric feature distribution of input set *P*.

(2) **Cluster analysis and peak extraction of 3D position vectors.** The feature distribution of 3D position vectors is clustered in the 3D Cartesian space defined by the template. If the reference structure displays symmetry, the input dataset of 3D vectors can be mapped onto a symmetrized space. A Gaussian mixture model is fitted to the 3D vector distribution, theoretically speaking, each component distribution $\mathcal{N}(\boldsymbol{\mu}_\mathrm{x}, \Sigma_\mathrm{x})$ represents a position cluster $\mathrm{C_x}$, Bayesian interpretation of the estimated model parameters $\boldsymbol{\mu}_\mathbf{x}$ and Σ_x (3D mean vector and covariance matrix, respectively) is subsequently used to extract the mode peak of each $\mathrm{C_x}$ cluster: a likelihood cutoff parameter is used to discriminate the subpopulation of data points enclosed in the 3D region of the cluster mode, i.e. the domain space where the $\mathcal{N}(\mathbf{p}|\boldsymbol{\mu}_\mathrm{x}, \Sigma_\mathrm{x})$ likelihood function exceeds the cutoff (figure 4.1 D), where **p** represents a 3D random vector.

(3) **Cluster analysis and peak extraction of relative rotations.** Once peaks of 3D position clusters have been identified, a position peak of interest is selected for quaternion-based cluster analysis of the associated relative rotation distribution. The proposed clustering approach can be applied in symmetrized space. Finally, a Bingham distribution function is fitted to each quaternion cluster $\mathrm{C_y}$, and likelihood-based peak extraction is done as described in step (2). Modes of 3D position and relative rotation clusters, describe specific geometric configurations of particle pairs.

Estimated statistical parameters of Gaussian and Bingham fits, provide a quantitative description of 3D position and rotation clusters, useful for structural interpretation of statistically significant particle pair configurations. Likelihood-based peak extraction allows for systematic segmentation of the geometric feature distribution. Moreover, Bayesian interpretation of fitted Gaussian and Bingham models can be subsequently used in predictive methodologies for structural characterization of supramolecular architectures.

4.3 Local Extraction of Geometric Information

Given a set of particles *P*, the initial step of the proposed analysis is the detection of radial neighborhoods from the distribution of particle positions $\{\mathbf{t_1}, \dots, \mathbf{t_n}\}$. The radial neighborhood of a particle $p_i \in P$, denoted as $N_r(i)$, is defined as the set of neighboring particles within a *r* nm radius, i.e. $N_r(i) = \{ p_j \in P \colon \mathrm{r} > \|\mathbf{t_i} - \mathbf{t_j}\| \}$. A k-dimensional tree [Bentley, 1975] of particle position vectors is used to efficiently calculate the 3D range queries required for detecting large amounts of neighborhood sets N_r.

Two types of geometric features are calculated for every $N_r(i), p_i \in P$, relative 3D position vectors $\mathbf{p_{i,j}}$ and relative rotations $\mathrm{R_{i,j}}$, for all neighboring particles: The relative 3D position vector $\mathbf{p_{i,j}}$ between reference particle p_i and a neighboring particle $p_j \in N_r(i)$, can be defined as a vector between centers of mass:

$$\mathbf{p_{i,j}} = R_i^T[\mathbf{t_j} - \mathbf{t_i}] \qquad (4.1)$$

The center-to-center vector $(\mathbf{t_j} - \mathbf{t_i})$ in tomographic space is mapped to the reference coordinate system by the inverse $\mathrm{R_i}$ transformation. Similarly, a 3D vector can also be derived from structural features of biological relevance (e.g. mRNA exit-to-entry vector of adjacent Ribosome particles, as presented in chapter 6). In this approach, the relative rotation matrix is defined as follows:

$$\mathrm{R_{i,j}} = \mathrm{R_i^T R_j} \tag{4.2}$$

Describing the relative orientation of particle *j* with respect to the reference coordinate system of particle *i*. In this manner, each neighborhood N_r contributes a set of position vectors and relative rotations to the overall distribution of geometric features associated with *P*, the distribution of detected particles in a single tomogram. This process can be iteratively applied to individual tomograms in large datasets, to generate geometric feature distributions from a series biological samples.

4.4 Analysis of 3D position Vectors

4.4.1 Handling Particle Symmetry in 3D Vector Space

In cases where the reference structure is symmetric, it is necessary to map the distribution of 3D position vectors to a symmetrized space before cluster analysis. The point group symmetry of the reference particle defines a set of rotation operations, which can be used to produce a series of symmetrically equivalent vectors for each $\mathbf{p_{i,j}}$, e.g. C_6 symmetry would imply 6 equivalent vectors, while D_4 implies 8. Each symmetry-associated rotation $\mathrm{R}_{\mathbf{a},\alpha} \in SO(3)$ can be fully characterized by two parameters: (1) a symmetry axis $\mathbf{a} = (\mathrm{x_a}, \mathrm{y_a}, \mathrm{z_a})$, defining a rotation axis in the orthonormal basis of the reference structure, and (2) a rotation angle $\alpha = \varkappa * \left(\frac{360°}{\mathcal{F}}\right)$, where $\mathcal{F}$ is the associated fold number and $0 \leq \varkappa < \mathcal{F}, \varkappa \in \mathbb{Z}$. Using these parameters, a rotation matrix can be directly computed:

$$\mathrm{R}_{\mathbf{a},\alpha} = \begin{bmatrix} \cos(\alpha) + x_a^2[1-\cos(\alpha)] & x_a y_a[1-\cos(\alpha)] - z_a \sin(\alpha) & x_a z_a[1-\cos(\alpha)] + y_a \sin(\alpha) \\ x_a y_a[1-\cos(\alpha)] + z_a \sin(\alpha) & \cos(\alpha) + y_a^2[1-\cos(\alpha)] & y_a z_a[1-\cos(\alpha)] - x_a \sin(\alpha) \\ x_a z_a[1-\cos(\alpha)] - y_a \sin(\alpha) & y_a z_a[1-\cos(\alpha)] + x_a \sin(\alpha) & \cos(\alpha) + z_a^2[1-\cos(\alpha)] \end{bmatrix} \tag{4.3}$$

By exhaustive sampling of $\varkappa$ factors and symmetry axes **a**, followed by multiplication of the resulting rotation matrices $\mathrm{R}_{\mathrm{a},\alpha}$ with a position vector $\mathbf{p_{i,j}}$, i.e. $\mathrm{R}_{\mathbf{a},\alpha}\mathbf{p_{i,j}}$, all symmetrically equivalent vectors can be generated.

Once symmetrically equivalent representations of the position vector distribution have been generated, it is possible to define a suitable space for cluster analysis. One approach is to use all symmetrically equivalent representations of the vector distribution, and subsequently adjust clustering parameters, in particular parameters related to the amount of expected clusters in the distribution. This approach will generate redundant clusters, which need to be merged in a post-processing step.

4.4.2 Cluster Analysis of 3D Vectors

Since the proposed methodology aims to segment the position distribution in a similarly manner as k-means clustering [Lloyd, 1982], and subsequently use a 3D Gaussian function to represent each cluster, it is theoretically equivalent to directly fit a 3D Gaussian mixture model (GMM) to the position distribution. The probability density function of a Gaussian mixture model is:

$$\mathrm{P}_{\mathrm{GMM}}(\mathbf{p}) = \sum_{x=1}^{k} \mathrm{w}_{\mathrm{x}}\, \mathcal{N}(\mathbf{p}; \boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathrm{x}}) \tag{4.4}$$

where $\mathbf{p}$ is a 3D random vector, the vector mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance matrix Σ_{x} are the statistical parameters for x^{th} Gaussian component, corresponding to vector cluster C_{x}, while w_{x} denotes component weight. A GMM is fitted to the distribution of 3D position vectors using an expectation-maximization procedure [Dempster et al., 1977].

4.4.3 Peak Extraction of 3D Vector Clusters

Once a GMM has been fitted to the 3D position vector distribution, likelihood functions of GMM components are used to derive a subset $\mathrm{C}_{\mathrm{x}}^{\mathrm{p}}$ from each cluster C_{x}, by extracting position vectors with likelihood values above a cutoff ϵ_x:

$$\mathrm{C}_{\mathrm{x}}^{\mathrm{p}} = \left\{\mathbf{p}_{\mathbf{i,j}} \in \mathrm{C}_{\mathrm{x}} \colon \mathcal{N}\left(\mathbf{p}_{\mathbf{i,j}} \middle| \boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathrm{x}}\right) \geq \epsilon_x \right\} \tag{4.5}$$

Appropriate values for likelihood cutoffs can be easily estimated by visual inspection of 3D vector distributions from extracted peaks $\mathrm{C}_{\mathrm{x}}^{\mathrm{p}}$. However, geometric interpretation of covariance matrices Σ_{x} can produce statistical relevant values, e.g. translation of the mean $\boldsymbol{\mu}_{\mathbf{x}}$ by an eigenvector of Σ_{x}, with vector magnitude set to the square root of the corresponding eigenvalue, will yield a point at one standard deviation distance from $\boldsymbol{\mu}_{\mathbf{x}}$, the likelihood function can be evaluated at this point, and the resulting value set as a cutoff. In this manner, likelihood cutoff values corresponding to factors of a standard deviation can be calculated.

4.5 Analysis of Relative Rotations

In the proposed approach, relative rotation analysis is limited to the feature distribution of a specific 3D position cluster. The aim is to restrict geometric feature space by relative 3D position, and subsequently dissect the associated distribution of relative rotations. As an iterative process, once the relative rotation distribution of a position cluster peak $\mathrm{C}_{\mathrm{x}}^{\mathrm{p}}$ has been analyzed, another position peak can be chosen for inspection, until the rotation distribution of all position peaks of interest have been scrutinized.

4.5.1 Handling Particle Symmetry in Relative Rotation Space

Symmetry of the reference structure presents similar challenges for relative rotation analysis as for 3D position vectors. Each relative rotation $\mathrm{R}_{\mathrm{i,j}}$, has a set of symmetrically equivalent rotations, defined by the point group symmetry, specifically, by the set of symmetry-associated rotation $\mathrm{R}_{\mathbf{a},\alpha}$. A symmetrically equivalent rotation for $\mathrm{R}_{\mathrm{i,j}}$ can be calculated by the rotation concatenating operation $\mathrm{R}_{\mathrm{i,j}}\mathrm{R}_{\mathbf{a},\alpha} \in SO(3)$, applying this operation for all valid parameters **a** and α, under the specific point group symmetry of the reference structure, will generate the complete set of equivalent rotations for $\mathrm{R}_{\mathrm{i,j}}$.

As in the case of 3D vectors, the relative rotation distribution needs to be mapped to a suitable space before cluster analysis can be performed. Analogous to the symmetrization of the position vector distribution, all symmetrically equivalent rotations for each $\mathrm{R}_{\mathrm{i,j}}$ in the distribution are generated, a clustering procedure is used to segment the symmetrized distribution, and redundant clusters are subsequently merged.

4.5.2 Cluster Analysis of Rotations

Clustering 3D rotations requires a distance measure which accurately describes similarity between two rotations $R_{i,j}, R_{k,l} \in SO(3)$. In this approach, a quaternion-based metric is applied for clustering relative rotations. Each rotation matrix $R_{i,j}$ is represented by an analogous unit quaternion $\mathbf{q}_{i,j} = (\mathrm{w}_{\mathrm{i,j}}, \mathrm{x}_{\mathrm{i,j}}, \mathrm{y}_{\mathrm{i,j}}, \mathrm{z}_{\mathrm{i,j}}) \in \mathbb{R}^{\mathbf{4}}$. The proposed metric is the inner product of quaternion vectors ($\mathbf{q}_{i,j} \cdot \mathbf{q}_{k,l} = \mathrm{w}_{\mathrm{i,j}}\mathrm{w}_{\mathrm{k,l}} + \mathrm{x}_{\mathrm{i,j}}\mathrm{x}_{\mathrm{k,l}} + \mathrm{y}_{\mathrm{i,j}}\mathrm{y}_{\mathrm{k,l}} + \mathrm{z}_{\mathrm{i,j}}\mathrm{z}_{\mathrm{k,l}}$), a measure proportional to the length of the geodesic path connecting the vectors on the 4D unit sphere [Kuffner, 2004]. Computation of the metric is described in algorithm 4.1, this algorithm accounts for the antipodally symmetric nature of quaternion space, i.e. polar opposite quaternions **q** and **– q** represent the same rotation**.**

Procedure for computing a quaternion similarity metric:

Input: Two unit quaternions $\mathbf{q_1}$ and $\mathbf{q_1}$
Output: Rotation similarity in the range [0, 1]

```
$\lambda = \mathbf{q_1} \cdot \mathbf{q_2}$
if $\lambda < 0$ then
        $\lambda = -\mathbf{q_1} \cdot \mathbf{q_2}$
return $\lambda$
```

Algorithm 5.1: Algorithm for calculating a quaternion-based similarity measure for elements of the *SO(3)* rotation set, adapted from [Kuffner, 2004]. This metric is proportional to the length of the geodesic path between quaternion vectors on the surface of the quaternion 3-sphere. It accounts for antipodal symmetry by evaluating the inner product of quaternion vectors.

A spectral clustering algorithm is used to cluster quaternions based on the metric described in algorithm 4.1. The input parameters for this procedure are a similarity matrix and the number of clusters *k*. The similarity matrix is subjected to dimensionality reduction, and subsequent k-means clustering of matrix eigenvectors yields a set of quaternion clusters C_{y}

[von Luxburg, 2007]. This clustering method was chosen given its capacity to operate in non-flat manifolds.

4.5.3 Peak Extraction of Rotation Clusters

Once relative rotations have been clustered, a Bingham distribution is fitted to each cluster set $\mathrm{C_y}$ and cluster peaks are extracted by likelihood cutoffs. The antipodally symmetric Bingham distribution on quaternion space is a zero-mean Gaussian function conditioned to lie on the surface of the unit 3-sphere $\mathbb{S}^3 \subset \mathbb{R}^4$ [Glover et al., 2011], the probability density function takes the form:

$$\mathcal{B}(\mathbf{q};\ \mathbf{\Lambda}, \mathrm{V}) = \frac{1}{F} \exp\left\{\sum_{\mathrm{i}=1}^{3} \lambda_{\mathrm{i}} (\mathrm{v_i^T} \mathbf{q})^2\right\} \tag{4.6}$$

where $\mathbf{q}$ is a random 4D vector in unit quaternion space, F the normalization constant, $\mathbf{\Lambda}$ a vector of 3 concentration parameters λ_{i}, and the columns of the 4x3 matrix V are orthogonal unit vectors $\mathrm{v_i}$. $\mathbf{\Lambda}$ and V are defined so that $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq 0$, where a small-magnitude λ_{i} indicates that the distribution is spread out along direction $\mathrm{v_i}$, and a large-magnitude λ_{i} indicates it is highly peaked along $\mathrm{v_i}$ [Glover & Kaelbling, 2013]. The parameters $\mathbf{\Lambda_y}$ and $\mathrm{V_y}$ of Bingham functions fitted to clusters $\mathrm{C_y}$ are estimated using the maximum-likelihood procedure described in [Glover et al., 2011]. Once Bingham functions have been fitted, likelihood peaks $\mathrm{C_y^p}$ are extracted from each cluster $\mathrm{C_y}$, by selecting quaternions with likelihood values above a cutoff ϵ_y:

$$\mathrm{C_y^p} = \left\{\mathbf{q_{i,j}} \in \mathrm{C_y}: \mathcal{B}\left(\mathbf{q_{i,j}} \middle| \mathbf{\Lambda_y}, \mathrm{V_y}\right) \geq \epsilon_y\ \right\} \tag{4.7}$$

Since geometric features were filtered, initially by relative 3D position and subsequently by relative rotation, particle pairs (p_i, p_j) corresponding to extracted features $\mathbf{q_{i,j}} \in \mathrm{C_y^p}$, have geometric configurations averaging the rigid transformation described by the modes of the distribution functions fitted to the associated clusters of position vectors and relative rotations, $\mathcal{N}(\mathbf{p};\ \boldsymbol{\mu}_\mathbf{x}, \Sigma_\mathrm{x})$ and $\mathcal{B}(\mathbf{q};\ \mathbf{\Lambda_y}, \mathrm{V_y})$ respectively. Thus describing a characteristic geometric arrangement for the subpopulation of particle pairs associated with $\mathrm{C_y^p}$.

4.6 Discussion

In this chapter, a systematic approach was presented to analyze local geometric information from the spatial distribution of detected macromolecular complexes in tomographic space. It aims at identifying characteristic geometric configurations of particle pairs by a three step approach: (1) detect spatial neighborhoods of particles and compute relative 3D position vectors and relative rotations. (2) Cluster analysis of 3D position vectors, followed by (3) cluster analysis of the relative rotation distribution associated with a previously selected position cluster. The proposed cluster analysis provides a statistical description of the geometric feature distribution by fitting probability distribution functions to clusters in 3D position and quaternion space. Once statistical models have been fitted, the use of a likelihood cutoff was proposed for cluster peak extraction, allowing identification of particle pairs associated with local modes in the geometric feature distribution.

Similar geometry-based analyses have been previously applied to describe the relative arrangement of neighboring ribosomes within polysomes. In [F. Brandt et al., 2009], characterization of the geometric arrangement between ribosome particles in cytosolic bacterial polysomes used k-means clustering of center-to-center vectors to identify clusters of 3' and 5' polysomic neighbors. Subsequently, the distribution of relative rotations associated with each center-to-center vector clusters was dissected by k-means clustering in quaternion space. Theoretically speaking, k-means clustering is not a suitable method for rotation clustering in quaternion space, as it assumes Euclidean distance between data points. This assumption does not hold in quaternion space, since this space is a manifold defined on the surface of the 4D unit hypersphere. In practice however, this methodology successfully identified characteristic 'top-to-top' and 'top-to-bottom' ribosome pair configurations of polysomic neighbors. A similar study of cytosolic polysomes in human cells [F. Brandt et al., 2010] applied hierarchical clustering of geometric features with a distance metric that combined a center-to-center vector distance and a quaternion-based distance by using a weighted sum model. This approach aimed at 3D vector and rotation clustering in a single step, as opposed to the proposed methodology, which conceptually decouples 3D vector and rotation. It is important to point out that this type of analyses rely on previous detection of macromolecular complexes, therefore sensitivity and specificity of template matching results are issues that need to be addressed beforehand.

A salient characteristic of the above-mentioned analysis is the efficient detection of particle neighborhoods. Particle neighborhoods are detected by near-neighbor queries in a k-dimensional tree structure, constructed from the 3D positions of *n* input particles. This allows efficient identification of local neighborhoods from large and densely clouds of points. Near-neighbor queries in a 3D k-dimensional tree have an average time of *O(log n)* [Freidman et al., 1977] and a worst-case time of $O(3n^{1-\frac{1}{3}})$ [D. T. Lee & Wong, 1977], significantly outperforming the *O(3n)* running time of a naïve linear search.

The two major contributions of this analysis are quaternion-based clustering and statistical description of rotation distributions. (1) This analysis uses a quaternion-based metric for rotation clustering. This metric accurately describes similarity between elements of the *SO(3)* group, since it is proportional to the geodesic distance between quaternions on the surface of the unit hypersphere. Accordingly, spectral clustering was chosen for its ability to operate in non-flat manifolds [von Luxburg, 2007]. (2) Bingham functions are used to describe rotation distributions in quaternion space. The Bingham function accurately represents the antipodal symmetry of this space, and is the maximum entropy distribution on the quaternion hypersphere that matches the sample inertia matrix $E[\mathbf{q}\mathbf{q}^{\mathrm{T}}]$, where $\mathbf{q}$ is a unit quaternion [Mardia, 1975], thus it may be better suited to represent a noisy quaternion distribution than other models [Glover & Kaelbling, 2013].

It is important to mention that global fitting of distribution functions to clusters (i.e. subsets of the distribution) can lead to artificially sharpened distribution fits. This issue can significantly impact parameter estimation when clusters do not fully capture the corresponding local mode, or when a significant amount of the data is uniformly distributed. Artificial sharpening of fitted distributions can be addressed by a local fitting procedure, i.e. setting boundary conditions on random variables in order to restrict the optimization procedure to cluster boundaries. Alternatively, a mixture model can be directly fitted to the distribution. In order to avoid artificial sharpening of Bingham fits, a natural extension of the proposed rotation analysis would be the inclusion of a Bingham mixture model fit, which could use the above-described quaternion clustering procedure as an initialization strategy for maximum-likelihood estimation of model parameters. This extension would allow the inclusion of a uniform component into the mixture model, which can be expressed as a Bingham component with all its concentration parameters λ_i set to zero [Glover et al., 2011].

5. Local Organization of RuBisCO in the Pyrenoid of *C. reinhardtii* Cells

5.1 Introduction

Structural characterization of cellular regions associated with the CCM, such as the pyrenoid of the unicellular *C. reinhardtii* algae, might offer mechanistic insights in the cellular processes that evolved to overcome RuBisCO's enzymatic limitations. RuBisCO complexes in *C. reinhardtii* pyrenoids have been observed to be densely packed in supramolecular organizations similar to spheres in a close-packing configuration, specifically a hexagonal close-packing arrangement [Engel et al., 2015]. However, resolution limitations of the CCD-acquired tomograms analyzed in this previous study, precluded identification of RuBisCO complex orientations, thus limiting the geometric information available for construction of models describing the 3D arrangement of RuBisCO complexes in the pyrenoid.

Here, we employed local geometric analysis to characterize the local arrangement of RuBisCO complexes in high-resolution tomograms of *C. reinhardtii* pyrenoids. First, tomograms of cryo-FIB milled pyrenoid lamellas were subjected to subtomogram analysis: Template matching was used to localize RuBisCO complexes in tomographic space (as described in chapter 3), yielding an initial set of positions and orientations. Subtomogram alignment was subsequently used to refine the geometric parameters of the detected RuBisCO particles. Once detected particles were aligned, subtomogram classification yielded a highly specific subset of RuBisCO particles. Nominal resolution of the resulting symmetrized average was measured and the associated particles were used as input for local geometric analysis (chapter 4).

Geometric analysis of the 3D distribution of RuBisCO complexes aimed at identifying predominant arrangements of RuBisCO particle dimers, which could be subsequently used to characterize the local organization of RuBisCO complexes. The local distributions of relative positions and rotations were analyzed and clustered in D_4 symmetrized space. Statistical peak extraction was used to test sample predominant geometric clusters and calculate averages of the corresponding RuBisCO dimer arrangements. Finally, the previously identified RuBisCO dimer configurations were used to construct a geometric model of complexes within the first near-neighbor (NN) shell.

5.2 Tomogram Dataset

Tomogram acquisition with a direct electron detector (section 3.2.1) greatly improved tomogram resolution, enabling direct visualization of the *in situ* RuBisCO structure. This allowed the expansion of the preliminary analysis of the organization of pyrenoid RuBisCO reported in [Engel et al., 2015]. Figure 5.1 depicts a tomogram of a pyrenoid lamella.

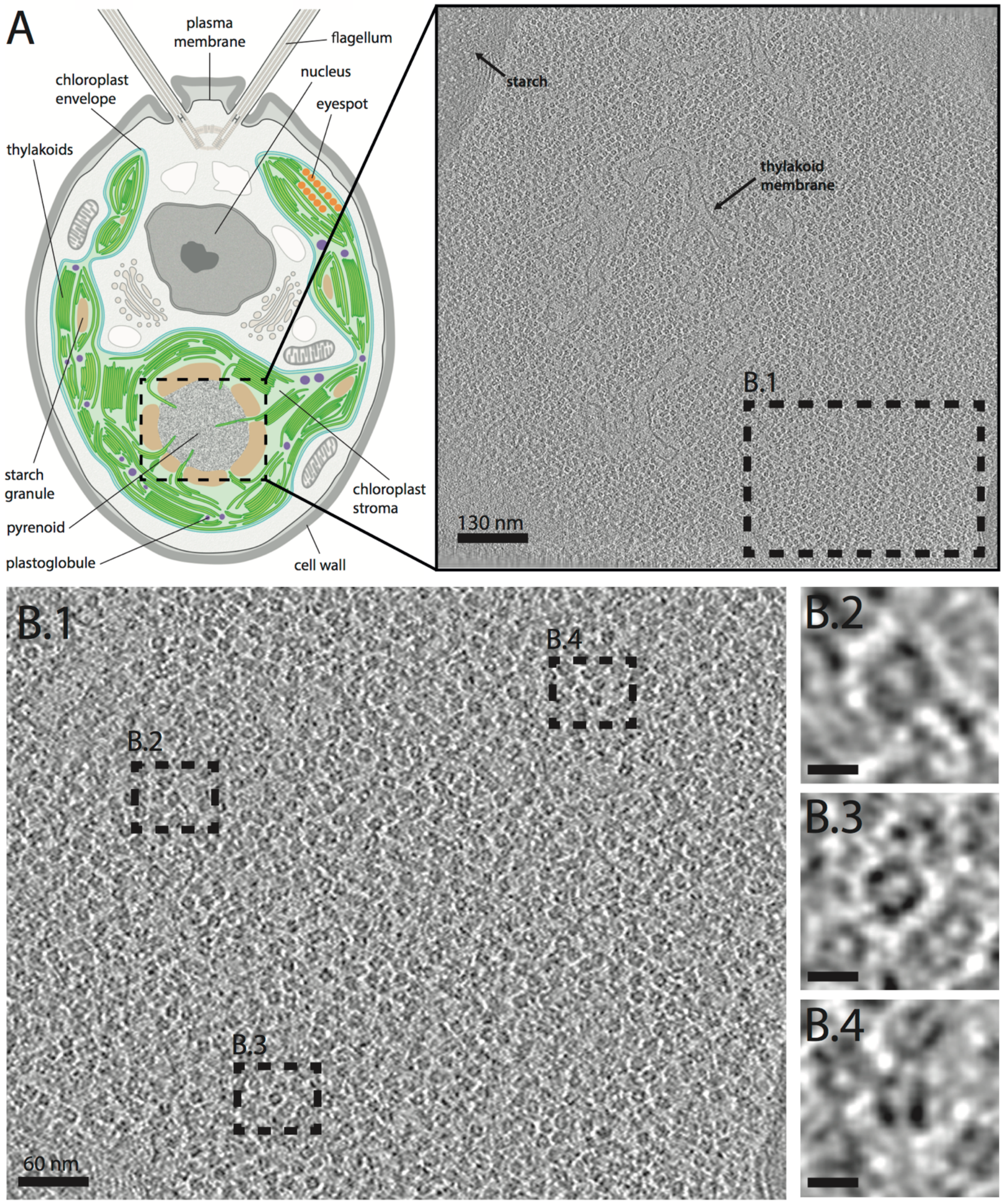

Figure 5.1: Tomogram of pyrenoid lamellas from plunge-frozen *C. reinhardtii* cells. (A) Diagram of a *C. reinhardtii* cell, adapted from [Engel et al., 2015] next to a cross-section from a tomogram of a pyrenoid. (B.1) Region of the tomographic cross-section showing the high density of RuBisCO complexes. (B.2-4) Magnified examples of densities with structural features that are consistent with the cage-like shape of the RuBisCO complex at different orientations (rounded cube shape with a lack of density in the middle). Scale bars: 8.5 nm.

5.3 Localization of RuBisCO Complexes

For template matching, tomograms were binned to a pixel size of 1.36 nm, and a template of the *C. reinhardtii* RuBisCO complex was generated from a crystal structure of the same species (PDB entry 1GK8 [Taylor et al., 2001]). To thoroughly sample RuBisCO complexes from the densely packed pyrenoid, peaks were exhaustively extracted and then the expected Gaussian distribution of true positive particles (figure 5.2) was approximated by visual inspection of binned subtomograms. Specifically, 20 random subtomograms were evaluated for different template matching scores. Once true positive Gaussians were estimated for all nine tomograms, a cutoff value was set to one standard deviation towards the low-valued tail of each Gaussian, all particles with scores below the cutoff value were discarded (figure 5.3 A). Approximately 200,000 RuBisCO particles were extracted from each tomogram.

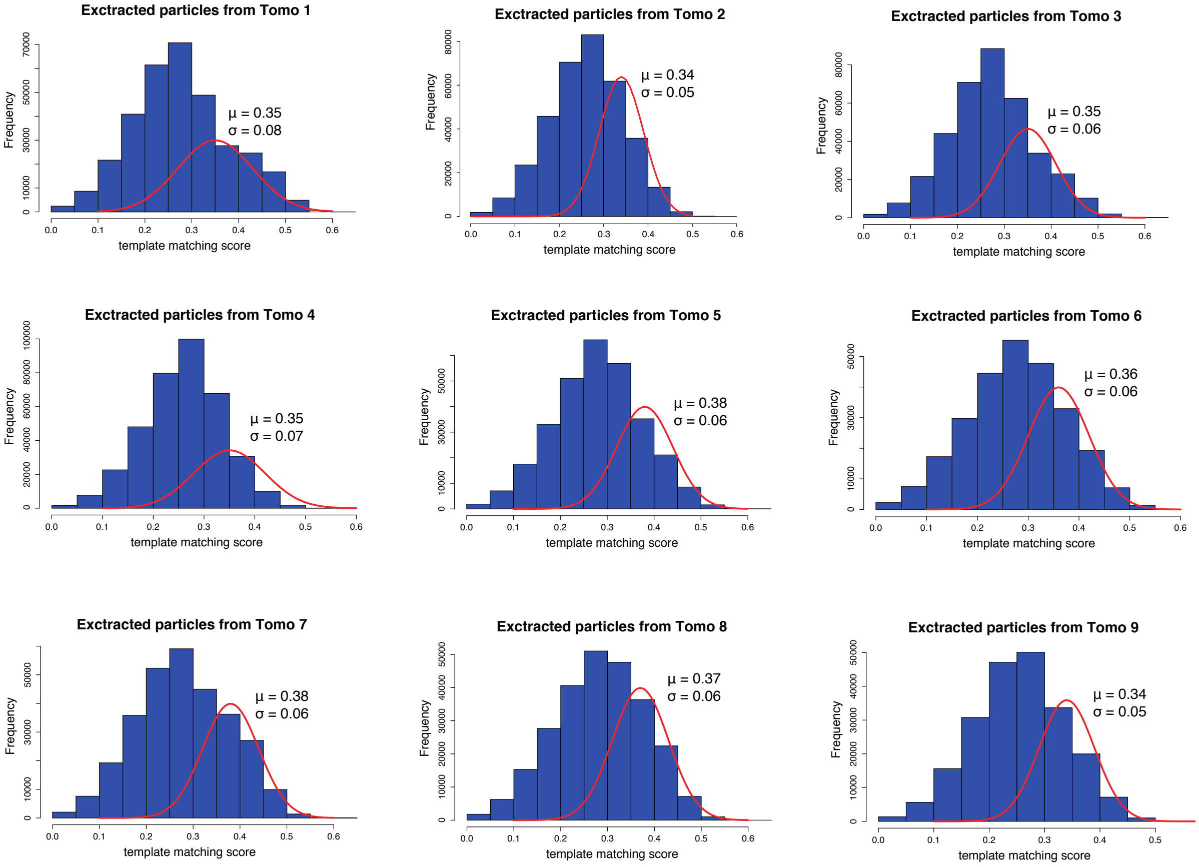

Figure 5.2: Histograms of template matching scores from exhaustive peak extraction of RuBisCO particles. Each histogram shows a single tomogram's estimated parameters for the Gaussian distribution of true positives particles (red).

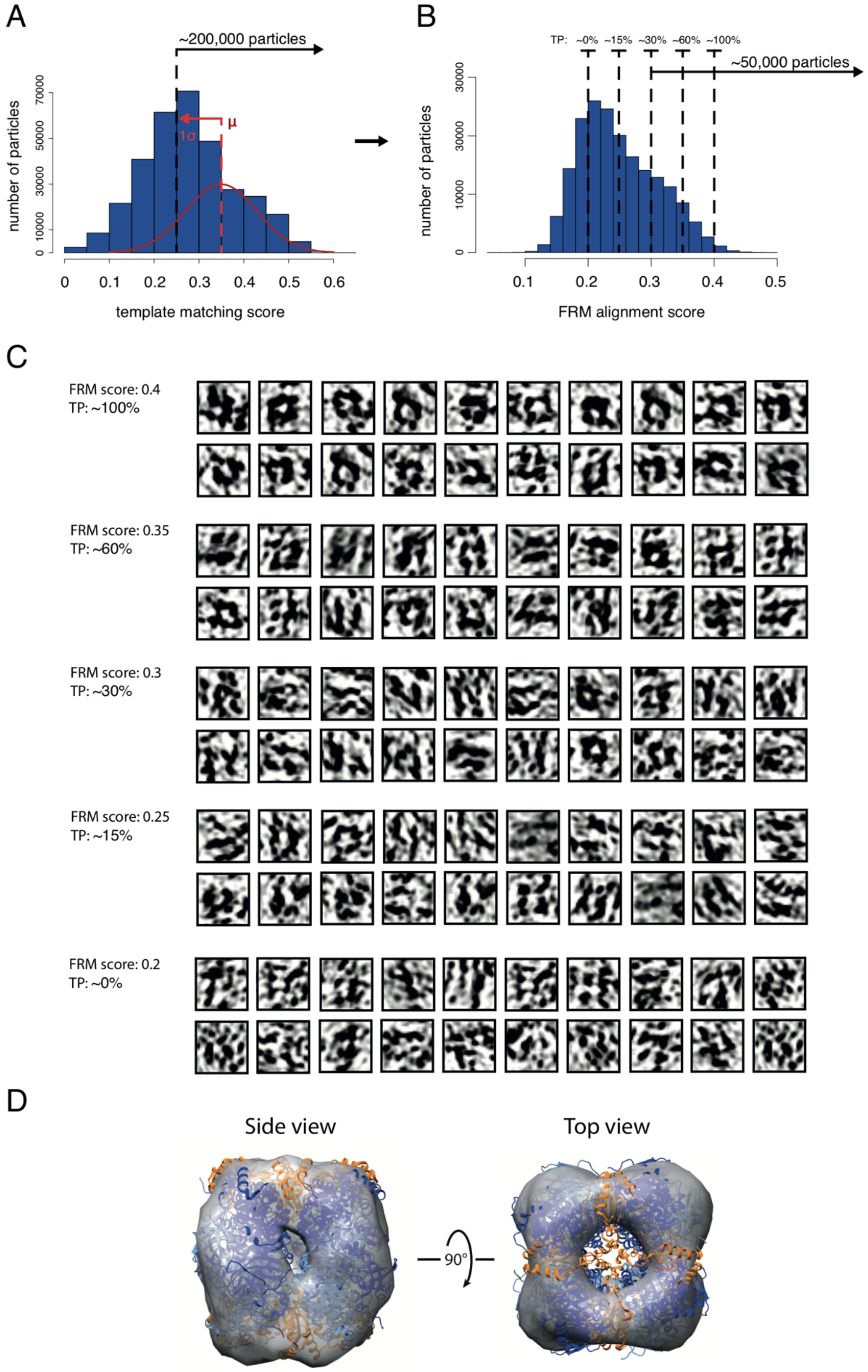

Figure 5.3: Subtomogram alignment of RuBisCO complexes. (A) Template matching results of a sample tomogram, depicting the score cutoff. (B) FRM alignment results for the sampled tomogram, showing

approximate true positive (TP) rates at different score values and the FRM score cutoff. (C) Visual inspection of aligned subtomograms at different FRM score values (20 subtomograms sampled per value); Projecting the sampled subtomograms along the 4-fold symmetry axis of the RuBisCO complex facilitates the estimation of TP rates. (D) Side and top views of an aligned average of ~50,000 RuBisCO complex particles, fitted to the crystal structure [Taylor et al., 2001]. Large subunits: blue and light blue, small subunits: orange.

5.4 Alignment of RuBisCO Subtomograms

Subtomograms of the extracted RuBisCO particles were binned to a pixel size of 0.68 nm for subtomogram alignment. FRM alignment was employed to process the large number of subtomograms (9 tomograms × ~200,000 particles = ~1,800,000 subtomograms). After alignment, reduction of the dataset was necessary to both remove large quantities of false positive particles and significantly reduce the computational complexity of processing the dataset. Further visual inspection of aligned subtomograms was used to estimate a FRM score cutoff value for each tomogram such that approximately 30% of the sampled particles were true positives (figure 5.3 B, C). By discarding particles with FRM scores below the cutoff values, the amount of particles was further reduced to ~50,000 particles per tomogram.

5.5 Classification of RuBisCO Subtomograms

The reduced dataset of aligned RuBisCO subtomograms (binned to a pixel size of 0.68 nm) was subjected to AC3D classification. The objective of this step was to identify a structurally homogenous subset of subtomograms corresponding to true positive RuBisCO particles, and in parallel detect sets of false positive particles which can be subsequently discarded, thereby increasing the structural quality of the dataset. Since the aim was to classify subtomograms on the basis of overall shape, the classification procedure was restricted to low spatial frequencies (< 38.3 Å resolution information). The number of classes was oversampled to allow for fine-grained segmentation of the dataset. Visual inspection of the resulting class averages was used to merge classes into structurally homogeneous subpopulations.

A significant percentage of the subtomogram dataset (~40%) belonged to a subpopulation of 'RuBisCO-like' particles, which were similar to the RuBisCO complex but displayed missing densities (figure 5.4). AC3D classification yielded a variety of classes showing 'RuBisCO-like' averages lacking densities in different regions (figure 5.4 C). Classes of false positive particles were grouped into a 'noise' class, and subsequently discarded. Mapping the classified particles back into tomographic space, showed that particles belonging to both the 'positive' and 'RuBisCO-like' class appeared to be evenly distributed within the pyrenoid (figure 5.5).

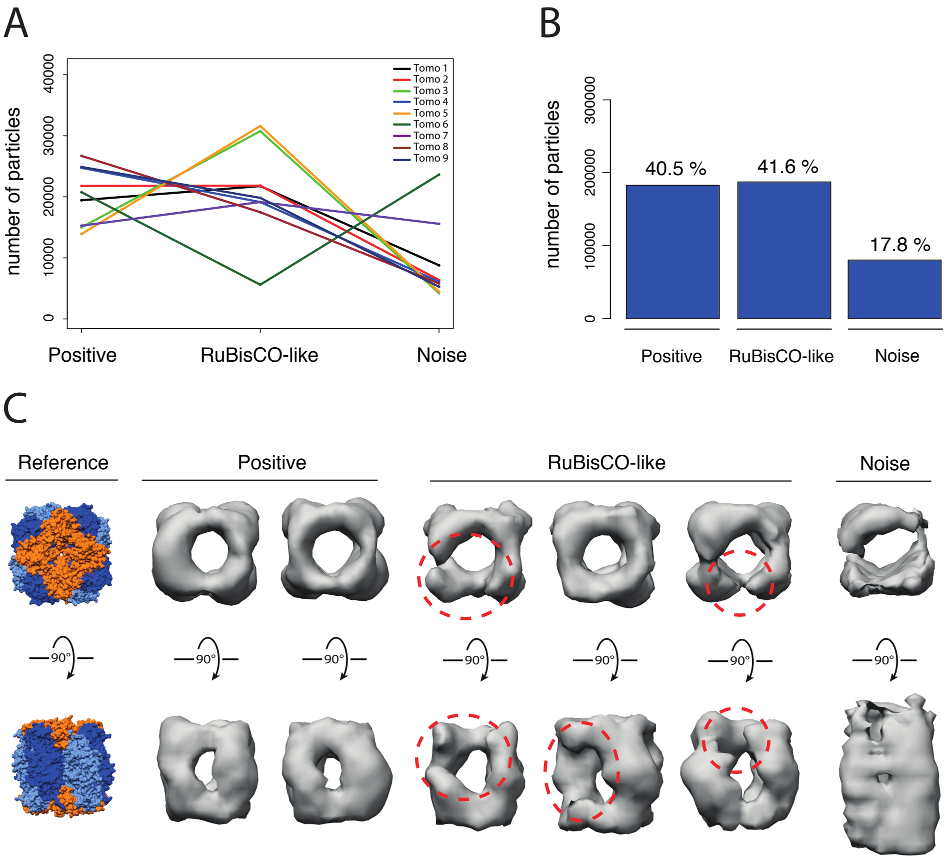

Figure 5.4: AC3D classification of detected RuBisCO complexes. Classification results by tomogram (A) and for the complete dataset (B), where the 'positive' class refers to a positive set of RuBisCO complexes (aiming for high specificity) and the 'RuBisCO-like' class denotes a set of particles similar to the RuBisCO complex but with missing densities. Classes of false positive subtomograms were labeled as 'noise'. (C) Surface model of the RuBisCO complex, as structural reference (large subunits: blue and light blue, small subunits: orange) next to sample averages of 'positive', 'RuBisCO-like' (dashed red lines marking regions with missing densities), and 'noise' subclasses. Pixel size: 0.68 nm.

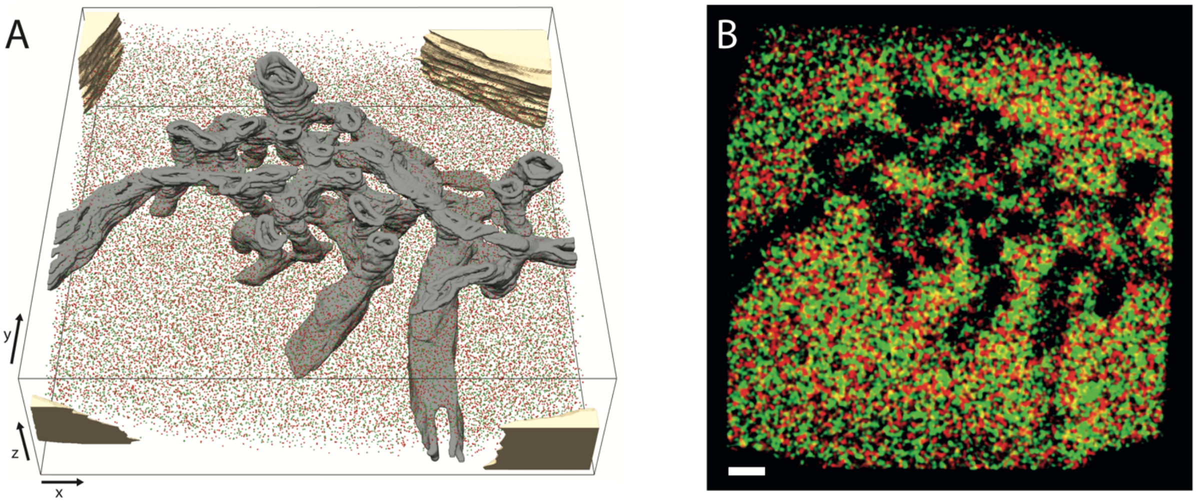

Figure 5.5: Distribution of 'positive' and 'RuBisCO-like' classes in a sample tomogram (Tomo1). (A) Rendered tomographic space depicting thylakoid membranes (gray), surrounding starch (yellow), and positions of

‘positive’ and ‘RuBisCO-like’ particles as green and red spheres respectively. (B) Projection along the **z** axis, showing ‘positive’ (green) and ‘RuBisCO-like’ (red) positions, and regions where classes overlap in the projected xy plane (yellow). Scale bar: 110 nm.

5.6 *In situ* Structure of the RuBisCO Complex

The ‘positive’ class was subsequently subjected to a final, two-tier subtomogram alignment step, producing final averages of the RuBisCO complex. In the first step, subtomogram were unbinned to their original 0.342 nm pixel size and realigned using FRM. Once all nine tomograms were FRM aligned, a second RS alignment step was used to refine shift and rotation parameters. Both alignment procedures enforced D_4 symmetry. Cross-resolution estimates per tomogram ranged from $FSC_{0.3} = 20$ Å to $FSC_{0.3} = 24$ Å, while the average produced by merging subtomogram sets from the top two performing tomograms yielded a resolution of $FSC_{0.3} = 16$ Å (figure 5.6). These top two tomograms were selected by visual inspection of the resulting averages, and cross-resolution FSC curves. The reference atomic model used for estimating resolution was derived from PDB entry 1GK8 [Taylor et al., 2001].

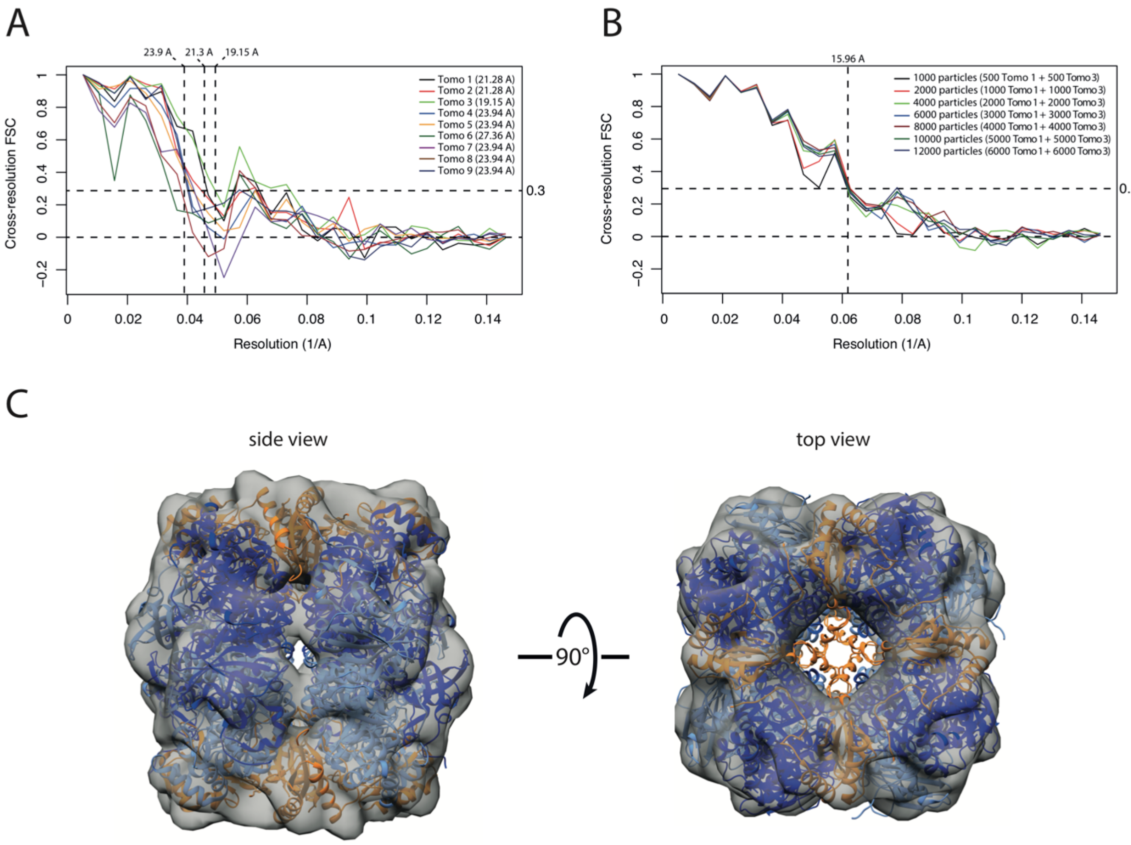

Figure 5.6: *In situ* structure of the RuBisCO complex resolved to 16 Å resolution. (A) Cross-resolution FSC curves (PDB entry 1GK8) of D_4 symmetrized averages from all nine tomograms (within a defocus range of 5 to 6 μm), and corresponding resolution estimates using the 0.3 criterion. All averages were computed using gold standard alignment. (B) Cross-resolution FSC curves for D_4 symmetrized averages computed by merging sets of particles from the top two performing tomograms, at varying amounts of particles. (C) Top and side views of the 12000 particle average (pixel size: 0.34 nm) from the top 2 tomograms (Tomo 1 and Tomo 3) with a fitted crystal structure [Taylor et al., 2001], large subunit: blue and light blue, small subunits: orange.

5.7 Local Organization of RuBisCO Complexes in the Pyrenoid

Next, the aligned subtomograms were leveraged to perform a fine-grained analysis of the local organization of RuBisCO complexes in pyrenoid tomograms. Unbinned alignment of the 'positive' class provided refined position and orientation parameters for this set of particles, allowing for geometric characterization of their local environment. Since the 'positive' particle set is comprised of AC3D classes that were labeled as true positives by visual inspection of class averages, specificity of this particle set is expected to be high. Furthermore, by visual inspection of unbinned averages and cross-resolution FSC curves, only six tomograms were selected for geometric analysis, each tomogram contributed ~20,000 particles.

The analysis presented here follows the methodology described in chapter 4. Peak extraction based on likelihood cutoffs was applied with the objective of test sampling specific modes of the 3D position vector and relative rotation distributions, in order to provide a geometric description of the local neighborhood. Extracting large fractions of distribution modes to compute statistically significant averages of RuBisCO dimers is beyond the scope of this chapter. However, RuBisCO dimer averages were computed as a control of correspondence between test samples of the geometry distributions and the observed electron densities in subtomograms.

5.7.1 Radial Distribution of Near Neighbors

The radial distribution of 3D center-to-center vectors from neighboring RuBisCO particles (figure 5.7 A) was analyzed by incremental segmentation of radial shells. The radial distribution function of NN positions reveled a large peak ranging from ~11 to ~16 nm radial distance from the center of the reference particle, corresponding to the first NN shell (figure 5.8 A); the 3D distribution of the associated vectors is depicted in figure 5.8 B. Fine-grained dissection of the NN position distribution was performed by incrementally selecting 1 nm radial shells, and plotting the associated 3D vectors on the unit sphere (figure 5.8 C.0-C.10).

Analysis of the distribution of NN positions revealed specific clusters. In this study, the following four predominant position clusters were sampled by 3D vector clustering: (1) the closest cluster, spanning a radial range of ~10 − 12 nm, (2) top and (3) side clusters within the ~11 − 14 nm radial range, and (4) the corner cluster, found within the ~15 − 18 nm range.

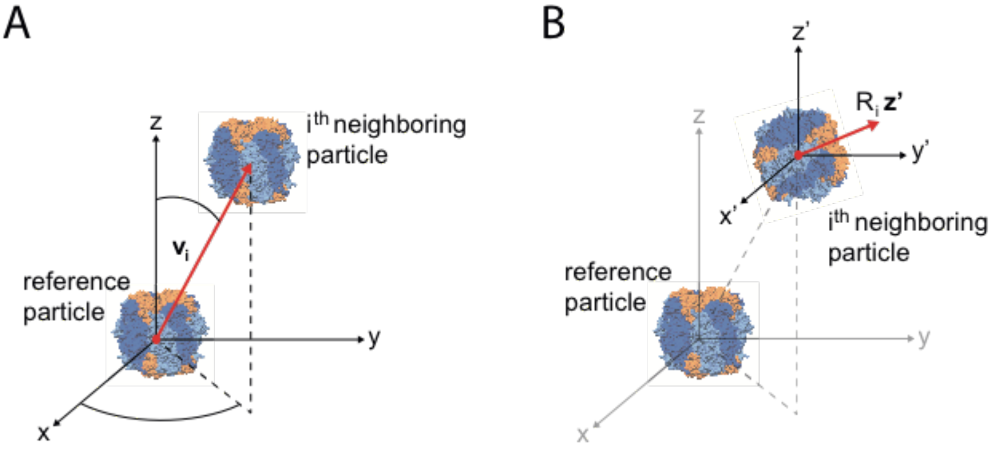

Figure 5.7: Vectors describing relative positions and rotations of neighboring RuBisCO particles. (A) Relative position with respect to the reference particle, depicting a center-to-center 3D vector $\mathbf{v_i}$ (red). (B) Relative rotation R_i of a neighboring particle *i*, represented as the 3D vector (red) produced by the rotation of the 4-fold symmetry axis **z'** from the neighboring particle. RuBisCO models derived from [Taylor et al., 2001], large subunit: blue and light blue, small subunits: orange.

5.7.2 Closest Neighbor Cluster

The closest cluster was sampled from a radial shell ranging from 9.9 nm to 12.0 nm (figure 5.9 A), with a cluster center located at a radial distance of 11.0 nm from the center of the reference particle. The source shell contained a total of 11,812 particle dimers, from which a cluster peak sample of 573 dimers was extracted; the resulting average is shown in figure 5.9 B.1-B.2. The distribution of relative rotations of particle dimers associated with the sampled peak of center-to-center vectors (figure 5.9 C), was clustered to extract a sample from the most predominant rotation peak. The sampled rotation peak consisted of 53 particle dimers (9.24% of the rotation distribution). The average of the resulting particle dimer arrangement is depicted in figure 5.9 D.1-D.2.

5.7.3 Side Neighbor Cluster

A radial shell ranging from 10.9 nm to 14.0 nm (figure 5.10 A) was used to extract a sample from the side cluster. The extracted peak sample of 1,689 vectors (from the 56,970 vectors captured within the source shell) was centered at a radial distance of 11.39 nm; the resulting average is depicted in figure 5.10 B. Subsequently, the relative rotations associated with the vector sample (figure 5.10 C) were clustered and the most predominant mode was sampled, the peak sample contained 103 rotations (6.09% of the rotation distribution). Finally, an average of the corresponding RuBisCO complex dimer was computed (figure 5.10 D).

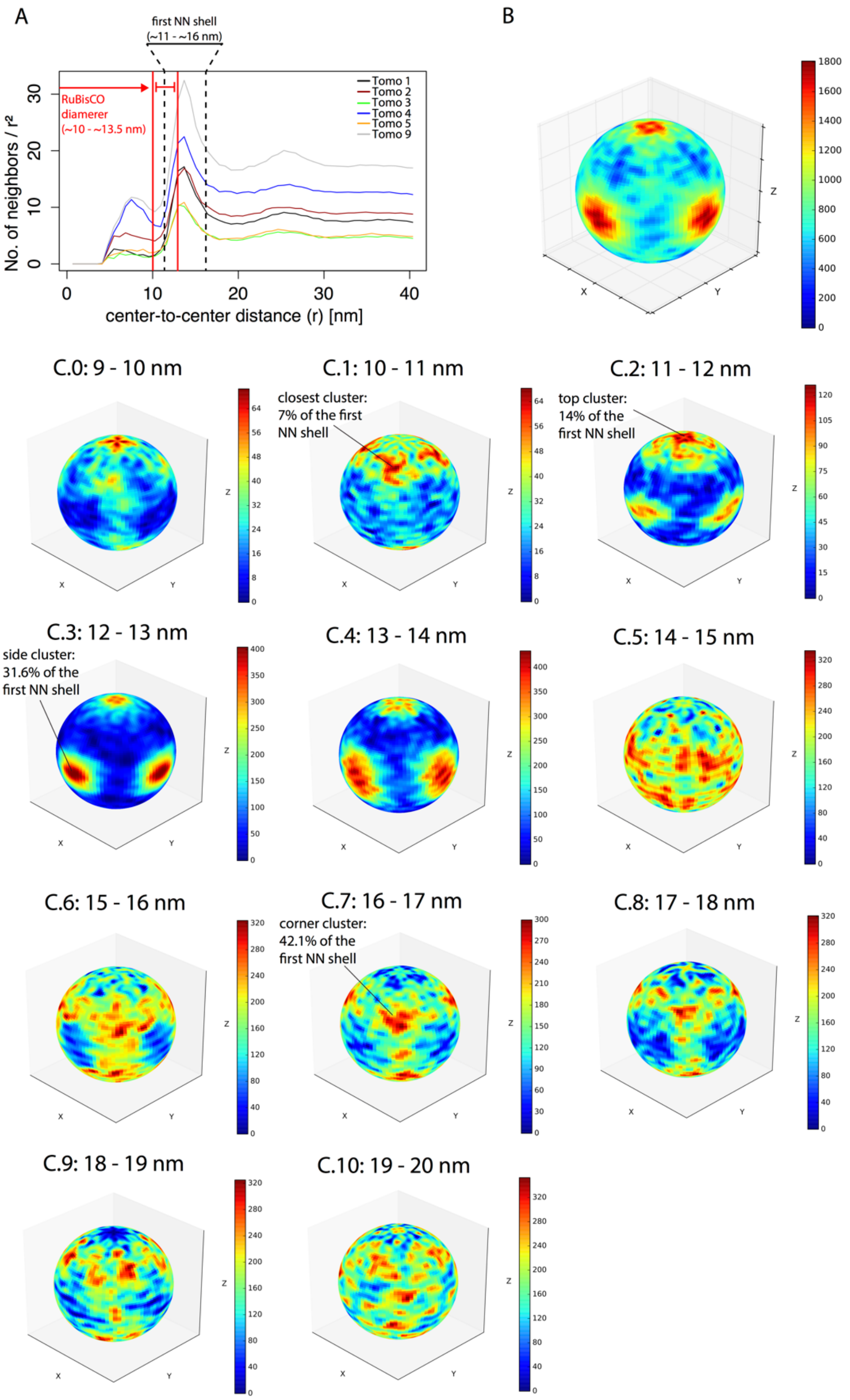

Figure 5.8: Position distribution of neighboring RuBisCO complexes. (A) Radial distribution of neighbor positions for a dataset of 6 tomograms, depicting the diameter range of the reference RuBisCO complex (red), and the radial shell attributed to the first layer of near neighbors (dashed lines). (B) distribution of center-to-center 3D vectors within the first NN shell (figure 5.7 A) mapped to the unit sphere. (C.0-C.10) Center-to-center vector distributions of 1 nm radial shells, ranging from 9 to 20 nm distance from the reference particle. All distributions displayed with D_4 symmetry applied.

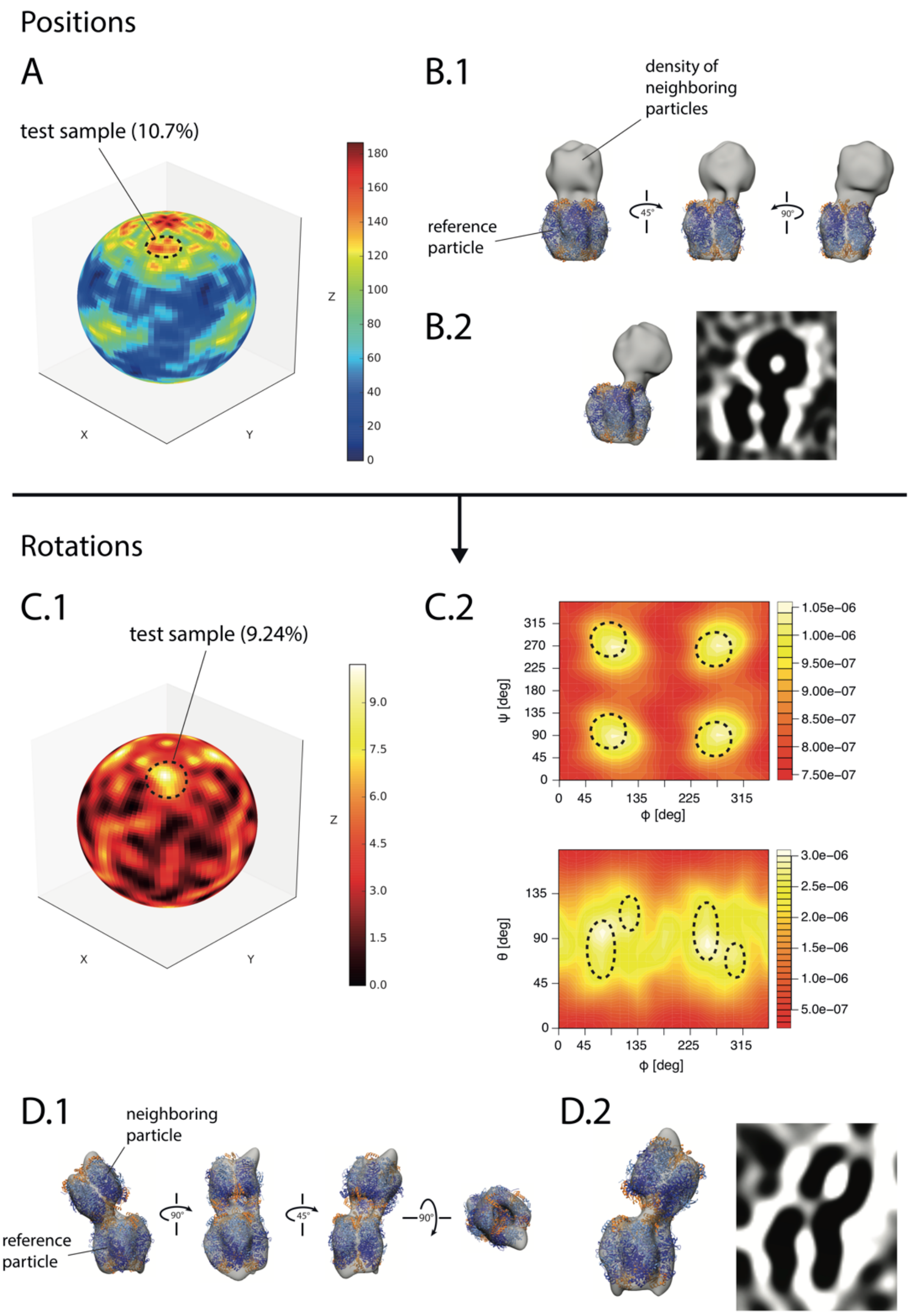

Figure 5.9: Geometric analysis of the closest cluster. (A) Full distribution of center-to-center 3D vectors from the 9.92 - 11.97 nm radial shell (figure 5.7A). The extracted peak sample from the vector cluster is depicted by a dashed circle, with the corresponding percentage from a tight radial shell containing the sample. (B.1) The resulting average (filtered to 3 nm resolution) displayed from different perspectives with a crystal structure fitted to the reference particle. (B.2) Middle cross-section of the average from the perspective depicted by the iso-surface representation. (C) Relative rotation distribution associated to the previously extracted 3D vector cluster, represented as a distribution of rotated 4-fold axes (figure 5.7 B) and in Euler angle space (C.1 and C.2 respectively). The extracted rotation sample is depicted with dashed circles. (D.1-D.2) Average (filtered to 4 nm resolution) of the resulting dimer arrangement displayed from different perspectives, crystal structures were fitted to both particles. (D.2) Middle cross-section of the average from the perspective depicted by the iso-surface representation. Fitted crystal structures [Taylor et al., 2001] with large subunits in blue and light blue, and small subunits in orange. All distributions displayed with D_4 symmetry applied.

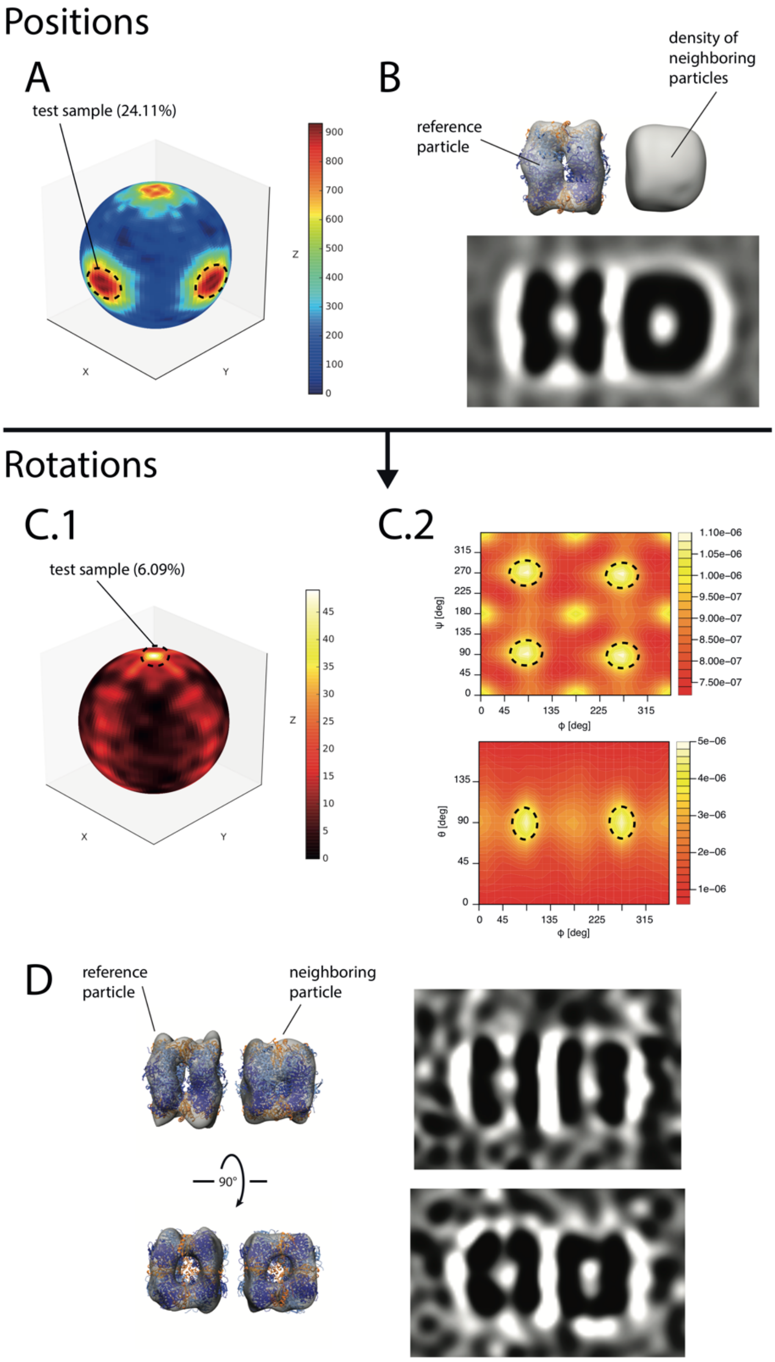

Figure 5.10: Geometric analysis of the side cluster. (A) Full distribution of center-to-center 3D vectors from the 10.94 – 14.02 nm radial shell (figure 5.7A). The extracted peak sample from the vector cluster is depicted by dashed circles, with the corresponding percentage from a tight radial shell containing the sample. (B) Resulting average with a crystal structure fitted to the reference particle, and a middle cross-section from the perspective depicted by the iso-surface representation. (C) Relative rotation distribution associated to the previously extracted 3D vector sample, represented as a distribution of rotated 4-fold axes (figure 5.7 B) and in Euler angle space (C.1 and C.2 respectively). The extracted rotation sample is depicted with dashed circles. (D) Average of the resulting particle dimer arrangement with crystal structures fitted to both complexes, and middle cross-sections of the average from the perspectives depicted by the iso-surface representations. Averages filtered to 3 nm resolution, fitted crystal structures [Taylor et al., 2001] with large subunits in blue and light blue, and small subunits in orange. All distributions displayed with D_4 symmetry applied.

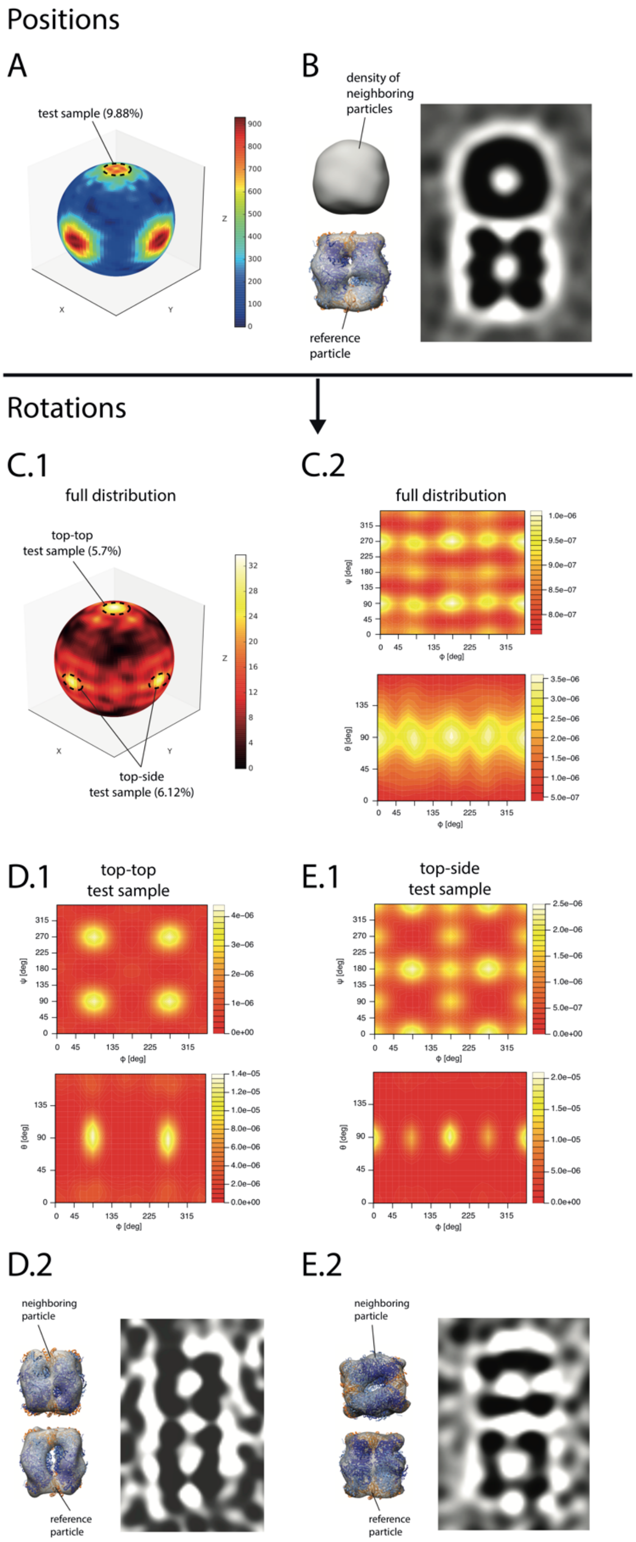

Figure 5.11: Geometric analysis of the top cluster. (A) Full distribution of center-to-center 3D vectors from the 10.6 – 14.02 nm radial shell (figure 5.7A). The extracted peak sample from the vector cluster is depicted by a dashed circle, with the corresponding percentage from a tight radial shell containing the sample. (B) Resulting average with a crystal structure fitted to the reference particle and a middle cross-section. (C) Relative rotation distribution associated to the previously sampled 3D vector cluster, represented as a distribution of

rotated 4-fold axes (figure 5.7 B) and in Euler angle space (C.1 and C.2 respectively). The extracted rotation samples are depicted with dashed circles. (D.1) Rotation distribution of the top-top sample, in spherical and Euler angle space. (D.2) Average of the resulting top-top dimer arrangement with crystal structures fitted to both particles, and middle cross-section of the average from the perspective depicted by the iso-surface representation. (E.1) Rotation distribution of the top-side sample, in spherical and Euler angle space. (E.2) Average of the resulting top-side dimer arrangement with crystal structures fitted to both particles, and middle cross-section of the average from the perspective depicted by the iso-surface representation. Averages filtered to 3 nm resolution. Fitted crystal structures [Taylor et al., 2001] with large subunits in blue and light blue, and small subunits in orange. All distributions displayed with D_4 symmetry applied.

5.7.4 Top Neighbor Cluster

The top cluster was sampled from a radial shell of 10.6 - 14.0 nm, containing a total of 58,546 vectors (figure 5.11 A). The extracted peak sample of 1,583 vectors had a center located 12.27 nm from the center of the reference particle; the sample's average is shown in figure 5.11 B. The relative rotation distribution corresponding to the extracted peak sample of 3D vectors had two predominant modes, designated as 'top-top' and 'top-side' (figure 5.11 C). Rotation clustering allowed the extraction of peak samples from both top-top and top-side modes, containing 91 (5.7% of the rotation distribution) and 97 (6.12% of the rotation distribution) rotations respectively. Figure 5.11 D shows the distribution of sampled rotations from the top-top mode, along with the resulting average, while figure 5.11 E shows analogous information for the top-side mode.

5.7.5 Corner Neighbor Cluster

The radial shell considered for sample extraction of the corner cluster ranged from 15 nm to 18 nm, containing a total of 76,504 vectors (figure 5.12 A). 3,251 vectors were sampled from the peak of the target cluster, the center of the sample was located at a radial distance of 16.1 nm. A sample from the predominant mode of the relative rotation distribution associated with this position sample (figure 5.12 C) was extracted, containing 185 rotations (5.69% of the rotation distribution). The average of the resulting RuBisCO dimer arrangement is depicted in figure 5.12 D.

5.7.6 Local Organization Model

Finally, the identified dimer arrangements of RuBisCO particles within the first NN shell (~10 - 18 nm) are were combined to build a geometric model of RuBisCO packing within the pyrenoid. Four dimer arrangements were used for the model: (1) top-top, (2) top-side, (3) side, and (4) corner. Individual arrangements are depicted in figure 5.13 A-C, and a complete model of the unit cell is shown in figure 5.13 D, with the percentage of first NN shell contribution per dimer arrangement. Different perspectives of the resulting 3D mesh of first NN positions are depicted in figures 5.13 E-F. Connected rings of first NN positions in figure 5.13 G, depict neighbor layers in an arrangement similar to close-packing of spheres: One hexagonal middle layer, and two surrounding layers above and below. However, in this model, both the top and bottom layer require 4 neighbor centers to fully saturate the space within the first NN shell, yielding a 14 neighbor configuration.

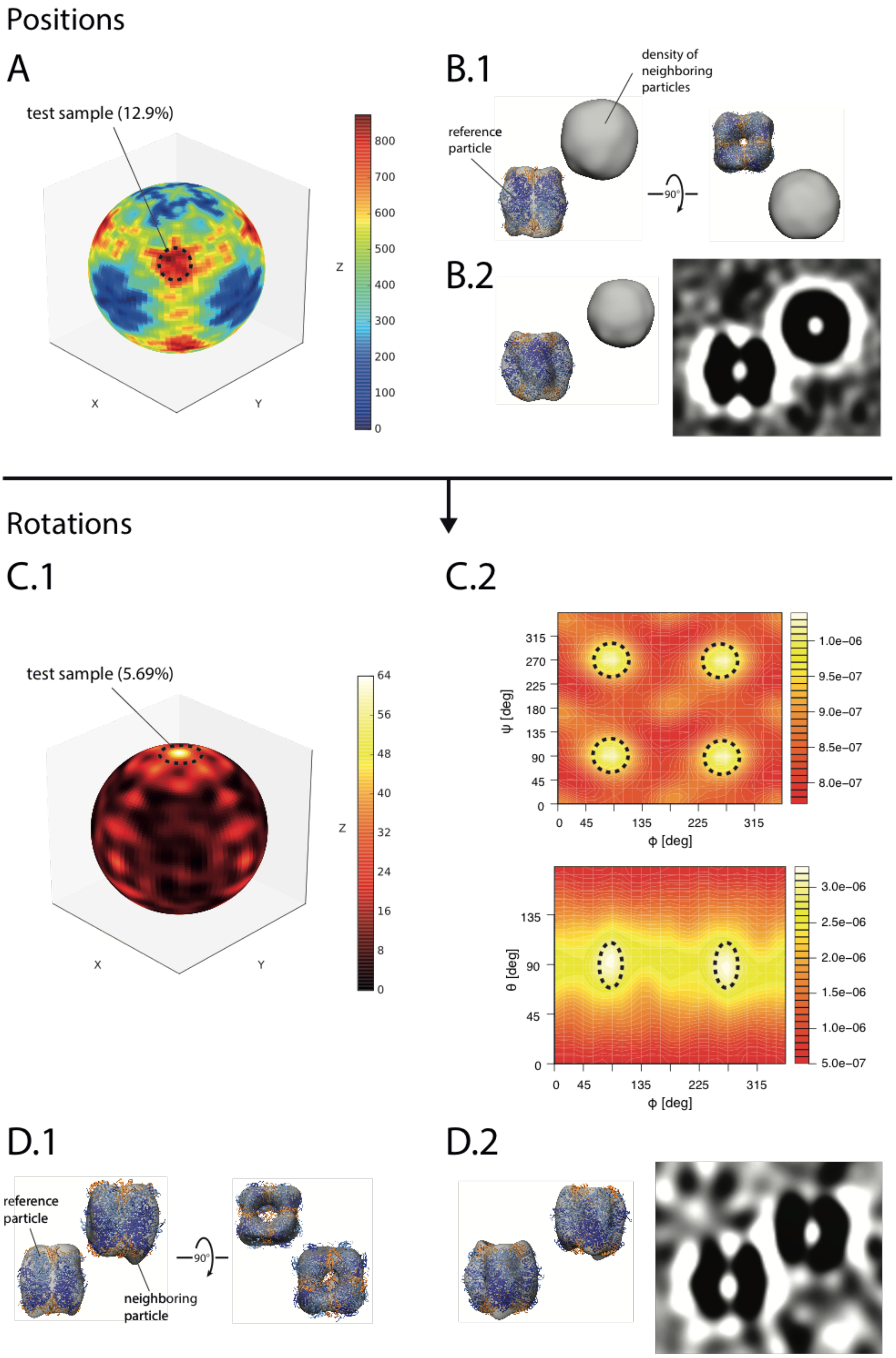

Figure 5.12: Geometric analysis of the corner cluster. (A) Full distribution of center-to-center 3D vectors from the 14.98 – 17.99 nm radial shell (figure 5.7A). The extracted peak sample from the vector cluster is depicted

by a dashed circle, with the corresponding percentage from a tight radial shell containing the sample. (B.1-B.2) Resulting average filtered to 3 nm resolution, with a crystal structure fitted to the reference particle. (B.2) Middle cross-section of the average from the perspective depicted by the iso-surface representation. (C) Relative rotation distribution associated to the previously extracted 3D vector sample, represented as a distribution of rotated 4-fold axes (figure 5.7 B) and in Euler angle space (C.1 and C.2 respectively). The extracted rotation sample is depicted with dashed circles. (D.1-D.2) Average (filtered to 4 nm resolution) of the resulting dimer arrangement displayed from different perspectives, crystal structures were fitted to both particles. (D.2) Middle cross-section of the average from the perspective depicted by the iso-surface representation. Fitted crystal structures [Taylor et al., 2001] with large subunits in blue and light blue, and small subunits in orange. All distributions displayed with D_4 symmetry applied.

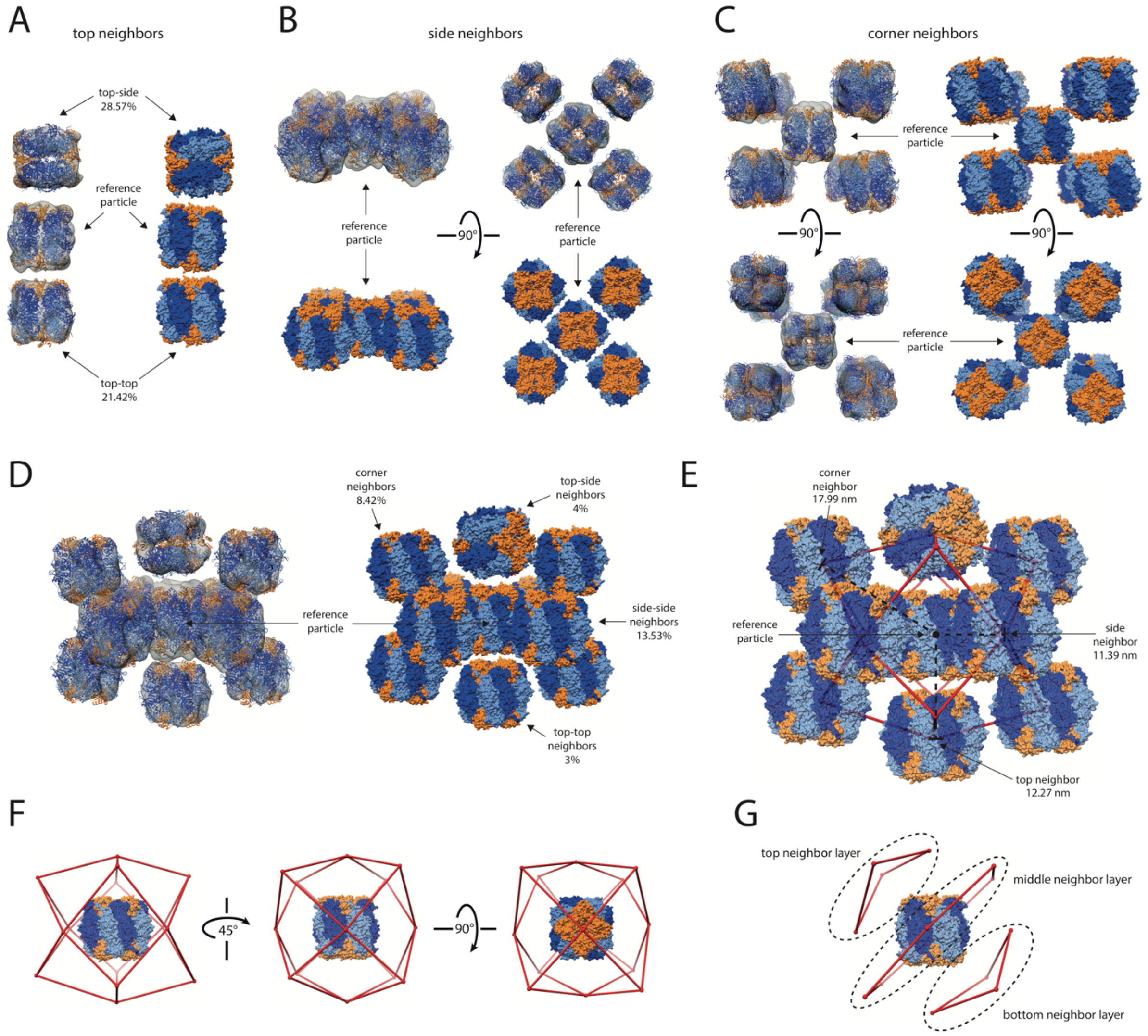

Figure 5.13: Proposed geometric model for the local organization of RuBisCO complexes in pyrenoids of *C. reinhardtii* cells. (A-C) Extracted averages of predominant dimer arrangements with fitted structures and corresponding surface models. (A) Percentage of the top distribution in top-top and top-side sub-clusters. (D) Integrative model of the RuBisCO complex unit cell, derived from 4 dimer arrangements (top-top, top-side, side, and corner clusters), displaying averages with fitted crystal structures, next to a corresponding surface model showing the percentages of the first NN shell per neighbor cluster. (E) Unit cell model of RuBisCO complexes showing a mesh over first NN positions (red) and radial distance to NN cluster centers. (F) Different perspectives of the first NN position mesh (red) with respect to the reference RuBisCO particle. (G) Rings of first NN positions (red) depicting neighbor layers resembling a configuration similar to close-packing

of spheres: a hexagonal middle layer and two surrounding layers with 4 points each. Fitted crystal structures and corresponding surface models [Taylor et al., 2001] depicting large subunits in blue and light blue, and small subunits in orange. Averages filtered to 3 nm resolution.

5.8 Discussion

Here, I present a study of the molecular organization of RuBisCO complexes in pyrenoids of *C. reinhardtii* cells. Subtomogram analysis of detected RuBisCO complexes in nine tomograms of cryo-FIB milled pyrenoid lamellas, was used to optimize the geometric information (i.e. position and orientation parameters of RuBisCO particles) needed to describe the *in situ* distribution of RuBisCO complexes in pyrenoid space. Initially, nine pyrenoid tomograms were subjected to template matching (using a reference structure of the RuBisCO complex [Taylor et al., 2001]), FRM alignment and AC3D classification. Approximately 50,000 subtomograms from each tomogram were used for classification, yielding a significant 'noise' class (17.8%), a class of 'RuBisCO-like' structures with missing densities in a variety of regions (41.6%), and a 'positive' class of RuBisCO complexes (40.5%). The positive class of RuBisCO particles was subjected to a final round of alignment, while D_4 symmetrized averages of individual tomograms yielded cross-resolution estimates of 19.15 to 23.94 Å, merging subtomogram sets from the top two performing tomograms produced a final *in situ* structure of nominal 16 Å resolution.

The proposed methodology for local geometric analysis (chapter 4) was applied to the set of particles labeled as positives by visual inspection of AC3D class averages, with the aim of dissecting the local arrangement of RuBisCO complexes. By inspection of the radial distribution of near neighbors, specifically within the first NN radial shell (~10 - 18 nm), five RuBisCO dimer arrangements were identified, corresponding to predominant modes in the associated distributions of relative positions and rotations: (1) The closest cluster (7.01% of the first NN shell) with position cluster center at a radial distance of 11.01 nm, displayed one predominant mode in rotation space. Interestingly, the extracted rotation sample generated an average showing a RuBisCO dimer configuration that brings neighboring small subunits in close proximity (figure 5.9 D.1, D.2). (2) The side cluster (31.57% of the first NN shell) was positioned 11.39 nm from the reference particle, and the associated distribution of relative rotations had a very distinctive mode, which yielded a 'side-side' dimer arrangement (figure 5.10 D). The rotation distribution of the top cluster (14.03% of the first NN shell, at a radial distance of 12.27 nm) was bimodal, the peak samples yielded the top-top (3) and top-side (4) dimer arrangements (figure 5.11 D.2, E.2 respectively). (5) The corner cluster (42.1% of the first NN shell), located at a 17.99 nm radial distance displayed one predominant mode in rotation space. The extracted rotation sample places the neighboring complex in a similar orientation as the reference particle (figure 5.12 D). It is noteworthy that retaining rigidity of RuBisCO dimer arrangements to produce clear averages, required extraction of samples within a small deviation from the modes. The four most predominant RuBisCO dimer arrangements of the first NN shell were integrated to propose a geometric model for the unit cell of RuBisCO complexes (figure 5.13 D-G). Figure 5.13 G shows layers of NN positions in a configuration resembling that of closely packed spheres, a hexagonal middle layer and 2 layers of 4 neighbor positions, creating a unit cell configuration that fully saturates the space within the first NN shell. Furthermore, the predominant RuBisCO dimer configurations suggest that the cube-like shape of the RuBisCO complex is a factor in the arrangement of near neighbors, as expected given the high density of RuBisCO complexes in the pyrenoid.

Neighboring complexes do not appear to be highly ordered. Even though predominant position clusters within the first NN shell account for a significant percentage of the shell (~88%), position clusters have large standard deviations (> 2 nm) when compared to the radius of the RuBisCO complex (~5 − 6.75 nm). While relative rotation distributions associated with position cluster do have predominant modes, most of the remaining distribution remains highly disorganized. Moreover, relative rotation clusters also exhibit angular variance above 4°. Since only a small percentage (~30%) of the first NN shell displays order under both relative position and rotation, is reasonable to conclude that there is a high degree of flexibility in particle dimer arrangements. These findings are consistent with a relatively amorphous pyrenoid matrix. Using a CCD-acquired tomographic dataset, a model based on close-packing of spheres was proposed to describe the organization of RuBisCO complexes in the pyrenoid [Engel et al., 2015]. Expansion of this preliminary analysis with the high-resolution data presented here has elucidated a highly dense and unorganized first NN shell. Moreover, the shape of the radial distribution functions of RuBisCO particles (figure 5.8 A) is characteristic of liquid substances, peaks gradually decrease as radial distance increases, completely disappearing within few factors of the particle diameter, i.e., a distribution with clear short-range order but lacking long-range order. In contrast, the radial distribution function of crystalline materials displays both decrease and increase of peak values, with peaks being clearly discernable at significantly higher radial distances [Zallen, 1998]. Furthermore, the lack of long-range order and high flexibility observed in the relative position and orientation distributions, supports the idea that the pyrenoid matrix is a dynamic and fluid-like environment, perhaps allowing flexible hexagonal and cubic close-packing configurations to appear transiently. Moreover, this could imply that first NN shells are not always, perhaps rarely, fully saturated, suggesting that unit cell configuration of 12 neighbors might be commonplace, as proposed by previous CET studies of the *C. reinhardtii* pyrenoid [Engel et al., 2015]. However, since a measure of sensitivity and specificity of the 'positive' particle class is difficult to obtain, this study is not suitable to reliably calculate the average number of neighbors in the first NN shell and quantify its variance. It is noteworthy that the local geometric analysis presented here, greatly benefited from efficient detection of particle neighborhoods (section 4.3). Since pyrenoid matrices have a high concentration of RuBisCO complexes (figure 2.9 B, figure 5.1), each pyrenoid tomogram provided a large number of RuBisCO particles (~20,000), yielding a highly densely cloud of particle position points (figure 5.5). Thus, detection of RuBisCO neighborhoods using near-neighbor queries with k-dimensional trees [Bentley, 1975] allowed computation of a large number geometric features (i.e. center-to-center vectors and relative rotations) in affordable time.

Considering that the RuBisCO linker protein EPYC1 is highly abundant in the pyrenoid and has been shown to directly bind RuBisCO [Mackinder et al., 2016], a density between neighboring RuBisCO complex is expected to be visualized in dimer averages. However, it is possible that dimer flexibility (only ~5 − 10% of the orientation distributions is ordered) precludes detection of this small 33 kDa protein by subtomogram averaging. Furthermore, test samples from the distributions of 3D position vectors and relative rotations (section 5.7) are far too small to provide sufficient signal that would allow visualization of such a small and flexible protein in subtomogram averages. Moreover, given the dynamic and disordered packing of RuBisCO in the pyrenoid, it is very likely that EPYC1 is bound sub-stoichiometrically to RuBisCO. Since 'top', 'side', and 'corner' dimer classes have symmetrical relationships between RuBisCO complexes, geometric classification alone is insufficient for detecting which EPYC1 binding sites between neighboring complexes are bound by the linker protein. On the other hand, additional image-based alignment of the area between neighboring RuBisCO particles, by using a small mask between complexes to focus the alignment procedure, can generate density between the complexes, thus, this approach might introduce significant bias. However, an interesting characteristic of the proposed unit cell model, is that small subunits of neighboring complexes appear to consistently align and come in close proximity, not only between the central and neighboring complexes, but also between adjacent neighbor complexes (e.g. side and corner neighbors in figure 5.13 D, E). Thus, the proposed unit cell model is consistent with biochemical studies, which predict that EPYC1 proteins bind to the two hydrophobic alpha-helices of the RuBisCO small subunit [Meyer & Griffiths, 2013; Meyer et al., 2012].

6. Polysome Detection

6.1 Introduction

While structure-function studies on isolated ribosomes have greatly elucidated its enzymatic function [Voorhees & Ramakrishnan, 2013], the organizational principles of polysomes and their functional consequences remain poorly understood.

Identification of polysomes enables the study of structural features from supramolecular arrangements of ribosomes during protein synthesis. However, polysome detection is an inherently challenging task due to the flexible nature of their preferred arrangements, in particular for densely populated macromolecular landscapes. Another significant challenge is the low SNR of CET data, which directly impacts specificity and sensitivity of template matching, hindering ribosome localization, and thus making analysis of polysome topologies cumbersome.

Once template matching has identified ribosomes in a tomogram, and their position and orientation have been refined through subtomogram alignment. Computational analysis of these parameters is required to analyze polysomes. Here a graph-based method for polysome detection is proposed. I model the probability distribution of two ribosomes translating the same mRNA molecule based on the local geometric analysis of adjacent ribosomes. This model is then used in a probabilistic framework where localized ribosomes are represented as a neighborhood graph [von Luxburg, 2007]. A MRF is embedded on the neighborhood graph and a message-passing algorithm infers a polysome-label for each ribosome, i.e., to cluster ribosomes into polysomes.

The performance of the method was evaluated on simulated and experimental datasets of bacterial lysates [F. Brandt et al., 2009], the method was subsequently applied to tomograms of rough microsomes derived from ER of mouse myeloma cells, with the objective of identifying cytosolic and ER-associated Ribosomes.

6.2 Polysome Detection Workflow

Here the polysome detection problem is formally defined. Given a set of n detected ribosome particles $P = \{p_1, \dots, p_n\}$, where each particle $p_i = (\mathbf{t_i}, R_i)$ is described by a 3D vector $\mathbf{t_i}$ and rotation R_i, denoting the position and orientation of a detected ribosome in the tomographic coordinate system. The aim of the method is to classify all elements of P into disjoint subsets, where each subset represents a polysome.

The following steps outline the developed statistical inference method: (1) model the probability of an mRNA molecule connecting two adjacent ribosomes, (2) topology graph construction, (3) classification using a MRF, and (4) polysome clustering. The model in step (1) is constructed using a training dataset of tomograms, while steps (2) to (4) operate directly on the data where polysome detection will be carried out. The overall workflow is depicted in figure 6.1.

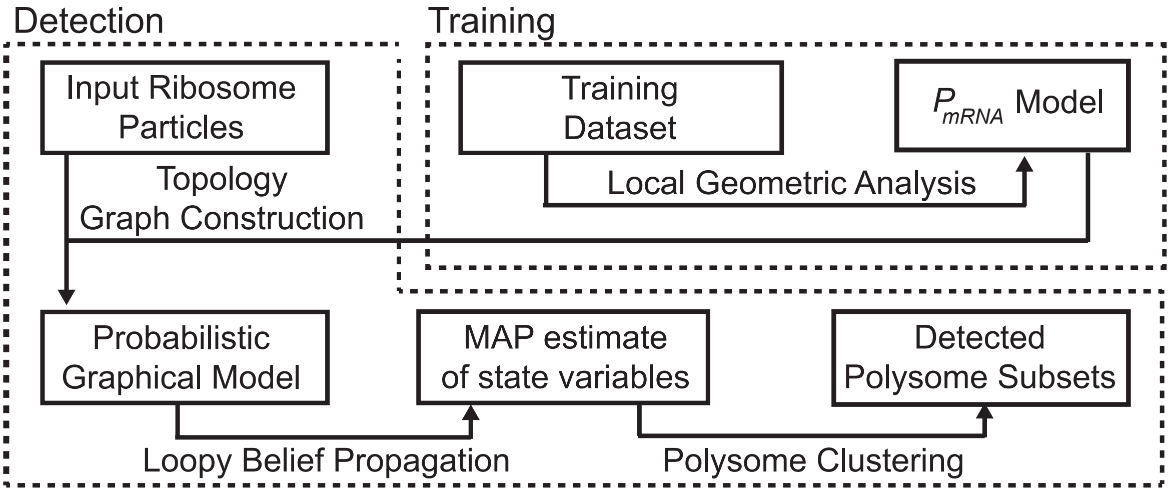

Figure 6.1: Polysome detection workflow. The dashed lines separate the training and detection phases. P_{mRNA} is generated in the training phase, and subsequently used in the detection phase for generating a topology graph from a set of input ribosome particles.

(1) Modeling the probability an mRNA connection between ribosomes. Here, a training dataset of tomograms is used to model the probability of two ribosome particles translating a single mRNA molecule, given their relative geometric arrangement. The model aims to capture the local arrangement of neighboring ribosomes within a polysome sequence. From the available data, a small number of tomograms are selected as a training dataset, the associated ribosome particles are subjected to a local geometric analysis and a probabilistic model P_{mRNA} is generated based on visual inspection.

(2) Topology graph construction. The input of the polysome detection method is a set of ribosome particles P from a single tomogram. The initial step of the detection phase is to generate a neighborhood graph using the set of 3D ribosome positions $\{\mathbf{t_1}, \dots, \mathbf{t_n}\}$ as vertices. Edge weights are derived from our previously constructed P_{mRNA} model.

(3) Classification using MRF. A MRF is defined on the topology graph by associating a state variable x_i to each vertex in the topology graph, and thus to each particle p_i. State variables denote polysome-labels, identifying a polysome subset for each particle. Our problem can now be stated as inferring the posteriori distribution of all states $X = (x_1, \dots, x_n)$ given observations P. More specifically, to obtain the MAP:

$$x^* = \arg\max{}_X P(X|P) \tag{6.1}$$

Where $x^* = (x_1^*, \dots, x_n^*)$ is the MAP for all n variables. Approximation of the MAP is performed using loopy belief propagation [Blake et al., 2011]. The underlying MRF is modeled to classify vertex sequences using the geometric information embedded in the topology graph, i.e. graph connectivity and edge weights.

(4) Polysome clustering. Ribosome particles are classified into polysome subsets according to the MAP estimate of polysome-labels. For each polysome subset, the sequence of ribosome particles from the 3' to 5' end can be extracted from the corresponding topology subgraph.

6.3 Probabilistic Model for the Local Geometric Arrangement of Polysomes

The aim of P_{mRNA} is to capture prior knowledge of the local organization of polysomes; it approximates the probability density function of an mRNA molecule being translated by two spatially adjacent ribosome particles. P_{mRNA}. The model $P_{mRNA}(\mathrm{i},\mathrm{j})$ is a function of the relative geometric arrangement of particles p_i and p_j. Given a training dataset of ribosome particles, the local distribution of particles is inspected to extract exit-to-entry vectors, i.e. 3D vectors between the mRNA exit site of the reference particle p_i, to the mRNA entry site of the neighboring particle p_j. The objective is to extract the cluster of exit-to-entry vectors corresponding to the polysomic neighbors on the 5' side of the reference particle.

Subsequently, the relative rotations corresponding to the 5' vector cluster are extracted. Once the 5' vector and rotation distributions have been identified, parametric density functions are fitted, labeled P_{vec} and P_{rot} respectively. By assuming statistical independence between P_{vec} and P_{rot}, a simplified model can be used:

$$P_{mRNA}(\mathrm{i},\mathrm{j}) = P_{vec}(p_i, p_j) \times P_{rot}(p_i, p_j) \tag{6.2}$$

6.4 Graphical Model for Probabilistic Polysome Detection

The objective of the polysome detection method can be stated as classification of input ribosome particles $P = \{p_1, \ldots, p_n\}$ into polysome subsets, i.e. inferring a polysome-label x_i for each ribosome particle p_i. The probability of states $X = (x_1, \ldots, x_n)$ given observations P can be expressed as $P(X|P) \propto P(P|X)P(X)$, thus the calculation of MAP-associated states takes the form:

$$x^* = \arg\max{}_X P(P|X)P(X) \tag{6.3}$$

Where $P(P|X)$ and $P(X)$ are the observation likelihood and label prior respectively.

The polysome-label is further defined as the index of the ribosome particle at the 3' end of the polysome sequence, thus, each variable x_i can take any value in our label space $L = \{1, \ldots, n\}$.

6.4.1 Topology Graph of Ribosomes

The topology graph describes the overall organization of input ribosome particles P in the coordinate system of tomogram. A neighborhood graph $G = (V, E)$ is derived from P, where the set of vertices $V = \{1, \ldots, n\}$ represents particles. An ordered pair $(i, j) \in E$ represents a directed edge connecting vertices $i, j \in V$. The neighborhood graph G is constructed by connecting vertices if the corresponding ribosome particles are spatially adjacent, i.e. an edge *(i,j)* is created if $\|\mathbf{t_i} - \mathbf{t_j}\| \le r_{max}$. The radial parameter r_{max} should be large enough to connect polysomic neighbors, but small enough to yield a ϵ-neighborhood graph. Range queries for extraction of near-neighboring particles within r_{max} are computed by querying a k-dimensional tree [Bentley, 1975] constructed from 3D positions $\{\mathbf{t_1}, \dots, \mathbf{t_n}\}$. In this context, directed edges can be understood as possible mRNA connections, in a 3' to 5' direction.

As $P_{mRNA}(i, j)$ models the probability of particle p_j being the polysomic neighbor of p_i on its 5' side, the likelihood function can be expressed as:

$$P(\mathrm{P}|\mathrm{X}) = \prod_{x_i \in \mathbf{x}} \xi(\mathrm{x_i})\tau(x_i, \mathrm{i}) \tag{6.4}$$

Where $\xi(i)$ is the probability of p_i being at the 3' end of a polysome, i.e. the probability of a polysome-label *i* . While $\tau(i, j)$ denotes the probability of an mRNA path from particle p_i (3' side) to particle p_j. Using graph G, $\xi(i)$ can be defined as:

$$\xi(\mathrm{i}) = \prod_{\mathrm{j} \in \mathrm{N(i)}} 1 - \mathrm{P_{mRNA}}(\mathrm{j}, \mathrm{i}) \tag{6.5}$$

Here $N(i)$ refers to the set of neighbors of *i* which contribute to its in-degree, i.e. the set of vertices *j* for which there is an edge $(j, i) \in E$.

Estimation of $\tau(i, j)$ considers the most probable mRNA path between vertices. The most probable path from vertex *i* to vertex *j*, denoted here as a vertex sequence $PATH_{i \to j}$, can be computed using Dijkstra's shortest path algorithm [Dijkstra, 1959] by setting edge weights as the negative log of P_{mRNA}, i.e., $w(i, j) = -\log\left(P_{mRNA}(i, j)\right)$. Assuming edges to be statistically independent, the probability of traversing $PATH_{i \to j}$ becomes:

$$\tau(\mathrm{i}, \mathrm{j}) = -\exp\left(\sum_{\mathrm{v_k} \in \mathrm{PATH_{i \to j}}} \mathrm{w}(\mathrm{v_k}, \mathrm{v_{k+1}})\right) \tag{6.6}$$

Where $\mathrm{v_k} \in \mathrm{PATH_{i \to j}}$ indicates the k^{th} vertex in the sequence and $\mathrm{v_{k+1}}$ the next vertex in the 5' direction. figure 6.1 illustrates a simple topology graph, exemplifying computation of ξ and τ.

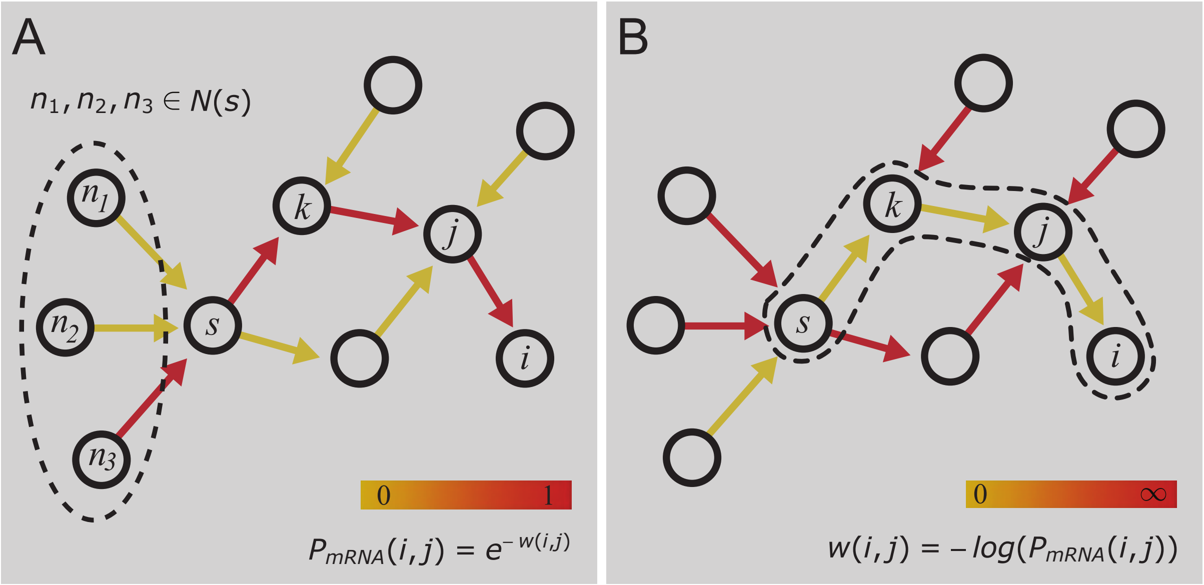

Figure 6.2: Illustration of a simple Topology graph. (A) Depiction of polysome-label probability ξ(s), explicitly showing *N(s)* (dashed line). (B) Using Dijkstra's shortest path algorithm to find the path that minimizes the sum of weights (dashed line), corresponding to the path with the largest mRNA probability$PATH_{s \to i}$.

6.4.2 Markov Random Field for Maximum-A-Posteriori Classification

A MRF on G acts as a prior model $P(X)$ for the hidden random variables X, under the set of observations P. We can express the posterior MRF in terms of a sum of energies (Gibbs energy), having a prior term Ψ and an observation likelihood term Φ [Blake et al., 2011]:

$$P(\mathrm{X}|P) = \frac{1}{Z(\mathrm{P})} \exp(-E(X, \mathrm{P})) \tag{6.7}$$

$$E(X, P) = \sum_{\mathrm{i} \in \mathrm{V}} \Phi_{\mathrm{i}}(\mathrm{x_i}, \mathrm{p_i}) + \sum_{(\mathrm{i,j}) \in \mathrm{E}} \Psi_{\mathrm{ij}}(\mathrm{x_i}, \mathrm{x_j}) \tag{6.8}$$

Both single and pairwise potential functions, Φ and Ψ respectively, can be defined as to model our polysome clustering problem. They are defined as follows:

$$\Phi_{\mathrm{i}}(\mathrm{x_j}, \mathrm{p_j}) = \begin{cases} -[\log(\xi(\mathrm{j})) + \log(\tau(\mathrm{j}, \mathrm{i}))] & \exists\, \mathrm{PATH}_{\mathrm{j} \to \mathrm{i}} \\ -\log(0) & \nexists\, \mathrm{PATH}_{\mathrm{j} \to \mathrm{i}} \end{cases} \tag{6.9}$$

$$\Psi_{\mathrm{ij}}(\mathrm{x_i}, \mathrm{x_j}) = \begin{cases} -\log(\mathrm{P_{mRNA}}(\mathrm{i}, \mathrm{j})) & \mathrm{x_i} = \mathrm{x_j} \text{ and } (\mathrm{i}, \mathrm{j}) \subset \mathrm{PATH}_{\mathrm{x_i} \to \mathrm{j}} \\ -\log(0) & \text{else} \end{cases} \tag{6.10}$$

Using this model, it is possible to efficiently approximate the MAP $x^* = \arg\max_x P(X|P)$. Using loopy belief propagation, specifically the max-product algorithm, the general energy function $E(X, P)$ can minimized. Since the energy function is the negative log of the posterior probability, we can compute the message from variable x_j to neighboring variable x_i in the following manner [Blake et al., 2011; Weiss & Freeman, 2001b]:

$$M_{j\rightarrow i}(x_i) = \min_{x_j}\left(\Phi_j(x_j, p_j) + \Psi_{ij}(x_i, x_j) + \sum_{k\in N(j)-\{i\}} M_{k\rightarrow j}(x_j)\right) \tag{6.11}$$

After a number of iterations (typically 3 to 5), the min-marginal belief b_i for every variable x_i is computed:

$$b_i(x_i) = \Phi_\mathrm{i}(x_i, p_i) + \sum_{k\in N(i)} M_{k\rightarrow i}(x_i) \tag{6.12}$$

A MAP estimate for variable x_i can later be obtained as $x_i^* = \arg\min_{x_i} b_i(x_i)$, where x_i^* indicates the inferred polysome-label for particle p_i.

Loop correction procedure:

Input: Initial polysome sets
Output: Merged polysome sets

```
for all labels $s$ do
    if $s \notin \Omega_s$ then
        for all labels $q$ do
            if $s \in \Omega_q$ then
                set $\Omega_q$ to $\Omega_q \cup \Omega_s$
            end if
        end for
    end if
end for
```

Algorithm 6.1: Procedure to merge polysome sets. Cycles in G with high P_{mRNA} values in every edge have an ill-defined polysome-label, since probability ξ is low for all vertices in the cycle. MAP classification tends to fragment these cycles into several polysome sets. This algorithm was used to merge polysome sets that present such behavior.

6.4.3 Polysome Clustering

Once the MAP estimate for X has been calculated, ribosome particles are grouped into polysome sets according to their polysome-label assignment. A polysome set with label s is denoted as Ω_s, representing a polysome where the ribosome particle s is positioned at the 3' end of the mRNA sequence.

Circular polysomes can generate loops in the topology graph (i.e., edges with high probabilities between 5' and 3' polysomic ends). In such cases the above method fragments true polysomes into separate polysomes. A consistent observation is that ribosome particles designated as 3'-labels of polysome fragments, are not elements of their own representative polysome set, i.e. $s \notin \Omega_s$, but elements of a neighboring polysome fragment. This observation can be used to overcome this issue, by merging polysome fragments in a post-processing step. Algorithm 6.1 is used to merge polysome sets when heavily weighted cycles in G lead to fragmentation.

6.5 Tomogram Datasets

The above-described detection was applied to three distinct datasets: (1) an expert-curated dataset of tomograms from bacterial lysates [F. Brandt et al., 2009] (section 3.2.2.1), and (2) a simulated dataset of bacterial lysate tomograms (section 3.2.2.2), analogous to the experimental dataset. This application was chosen since the topological features of characteristic polysome arrangements have been well investigated, moreover, the set of ribosome particles detected in the experimental dataset was subjected to visual inspection and manually classified into polysome and monosome classes [F. Brandt et al., 2009]. Using the polysome sets identified by [F. Brandt et al., 2009], a benchmark dataset was derived for quantitative evaluation of the polysome detection methodology. (3) a dataset of tomograms from microsomal preparations of rough ER, derived from mouse myeloma cells (section 3.2.2.3). Microsome tomograms contained both ER-associated ribosomes and cytosolic ribosomes, making the dataset an attractive candidate for polysome detection. For each dataset, a large amount of peaks was extracted to ensure high sensitivity, a requirement for recuperating true supramolecular arrangements. While specificity was sacrificed, for this particular analysis, false positive particles are preferred against false negatives.

6.6 Results

The first step in the proposed polysome detection methodology is to define a local arrangement model P_{mRNA} for each type of polysome. Here 3 models are defined, for cytosolic bacterial, ER-associated, and cytosolic mammalian polysomes. The cytosolic bacterial model was used for quantitative evaluation of the polysome detection method on simulated and experimental tomogram of bacterial lysates. The ER-associated and cytosolic mammalian models were used for polysome detection and analysis in tomograms of rough ER microsomes.

6.6.1 Local Model of Cytosolic Bacterial Polysomes

A training dataset of three tomograms of *E.coli* lysate (section 3.2.2.1) was used to derive a P_{mRNA} model for cytosolic bacterial polysomes. Previous analysis of this dataset revealed two characteristic local arrangements between neighboring 70S ribosomes, top-to-top (t-t) and top-to-bottom (t-b), giving rise to pseudo-helical and pseudo-planar organizations of polysomes [F. Brandt et al., 2009]. Local geometric analysis of the training dataset yielded results similar to what has previously reported in [F. Brandt et al., 2009], modes of the vector and relative rotation distributions were consistent with t-t and t-b configurations (figure 6.3).

From the training dataset, 360 ribosome particles were identified in pseudo-helical and pseudo-planar polysomes by visual inspection. Local geometric analysis revealed 2 distinct clusters of mRNA exit-to-entry vectors, corresponding to neighboring ribosomes on the 5' and 3' polysome directions from the reference ribosome (figure 6.3 A). A 3D Gaussian density function in Cartesian space (P_{vec}) was fitted to the 5' cluster of 183 vectors (figure 6.3 B). The distribution of relative rotations corresponding to the 5' vector cluster is bimodal, describing both t-t and t-b configurations. To model this orientation distribution, we define P_{rot} as mixture model of two Bingham distributions; one Bingham density function was fitted to the 31 quaternions of the t-b cluster, while the second was fitted to 152 quaternions of the t-t cluster.

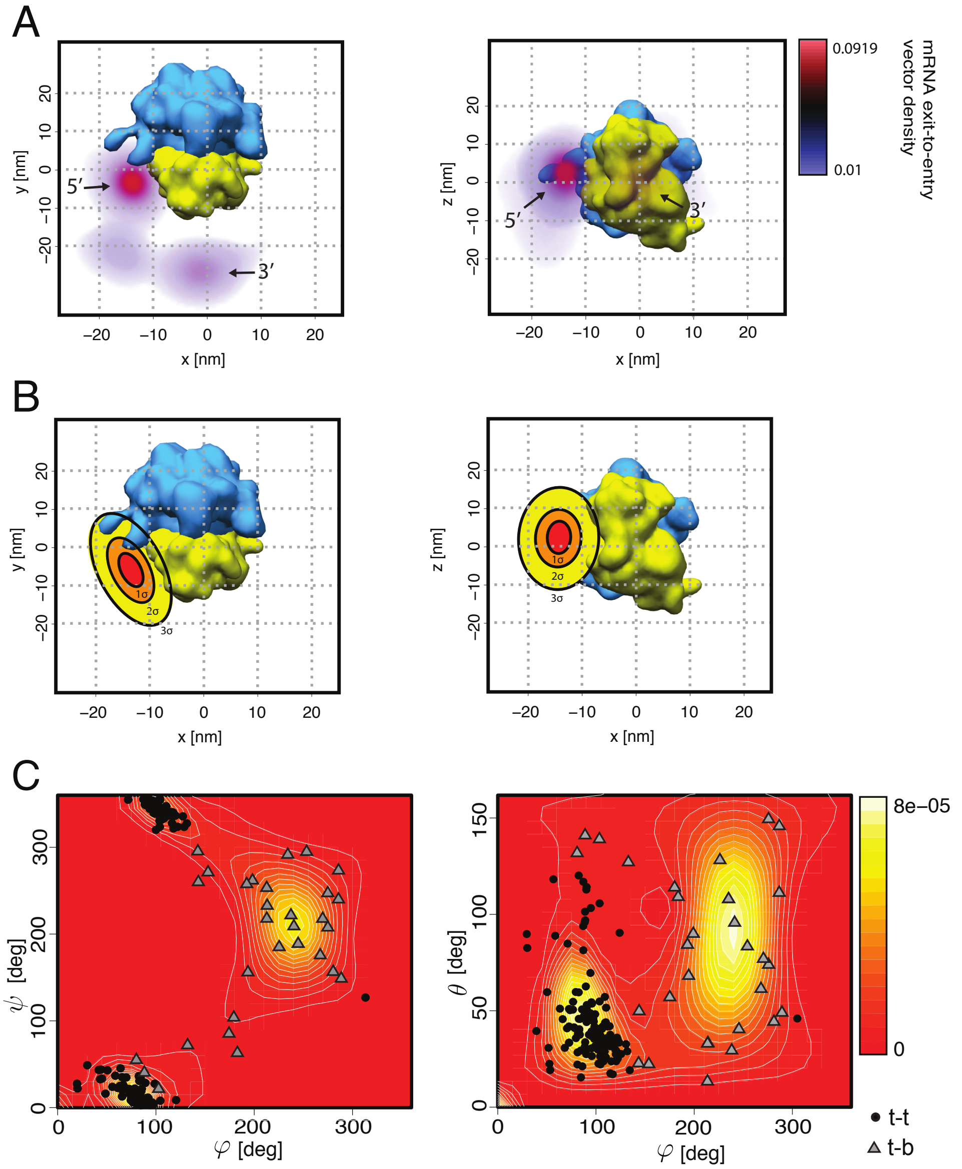

Figure 6.3: Local geometric analysis of bacterial Ribosomes. mRNA exit-to-entry vector and relative rotation distribution from a training dataset of 3 tomograms. (A) Density of mRNA exit-to-entry vectors. (B) 3D Gaussian fit for the 5' cluster with a goodness-of-fit $\chi^2_{red} = 1.28$. (C) Relative rotation distribution of the 5' cluster, showing the t-t and t-b clusters, Bingham distributions were fitted to the t-t and t-b clusters with a goodness-of-fit of $\chi^2_{red} = 1.28$ and $\chi^2_{red} = 7.41$ respectively. The Euler angle density shows the mixture model of the fitted Bingham distributions.

6.6.2 Quantitative Evaluation of Bacterial Polysome Detection

Once a P_{mRNA} model for cytosolic bacterial polysomes was defines, the experimental (the three remaining tomograms) and simulated datasets of cytosolic polysomes from bacterial lysate (sections 3.2.2.1, 3.2.2.2) were used for quantitative evaluation of the proposed polysome detection method. The experimental dataset was manually curated: a subset of ribosome particles was identified as polysomes in pseudo-helical and pseudo-planar topologies, their local arrangement was inspected to select only t-t and t-b local configurations [F. Brandt et al., 2009] and used as a positive class, whereas the remaining detected ribosomes were labeled as negative particles. The set of negative particles Ω_{neg} was defined as monosome particles, or false positive ribosomes. Furthermore, the amount of peaks extracted from the synthetic tomograms was oversampled to ensure high coverage of both positive and monosome classes.

A polysome set from the benchmarked positive class is denoted Ω'_s, as opposed to the inferred polysomes Ω_{s} from the detection method. Performance measures were calculated as follows:

$$\mathrm{TP} = \sum_{\mathrm{s}} |\Omega_{\mathrm{s}} \cap \ \Omega'_s| \tag{6.13}$$

$$\mathrm{FP} = \sum_{\mathrm{s}} |\Omega_{\mathrm{s}}| - |\Omega_{\mathrm{s}} \cap \ \Omega'_s| \tag{6.14}$$

$$\mathrm{TN} = \left| \ \Omega_{\mathrm{neg}} \ \right| - \left| \Omega_{\mathrm{neg}} \cap \bigcup_{\mathrm{s}} \Omega_{\mathrm{s}} \right| \tag{6.15}$$

$$\mathrm{FN} = \sum_{\mathrm{s}} | \ \Omega'_s| - |\Omega_{\mathrm{s}} \cap \ \Omega'_s| \tag{6.16}$$

Where TP: number of true positives, FP: number of false positives, TN: number of true negatives, and FN: number of false negatives. Performance measures are shown in table 6.1, examples of detected polysomes in the experimental dataset are depicted in figure 6.4.

Dataset	Accuracy	Precision	Recall
Synthetic	97.6%	81.3%	87.1%
Experimental	96.2%	82.6%	80.0%

Table 6.1: Performance measures of polysome detection on simulated and experimental datasets of bacterial lysate. Accuracy = (TP + TN)/(TP + TN + FP + FN), Precision = TP/(TP + FP), Recall = TP/(TP + FN).

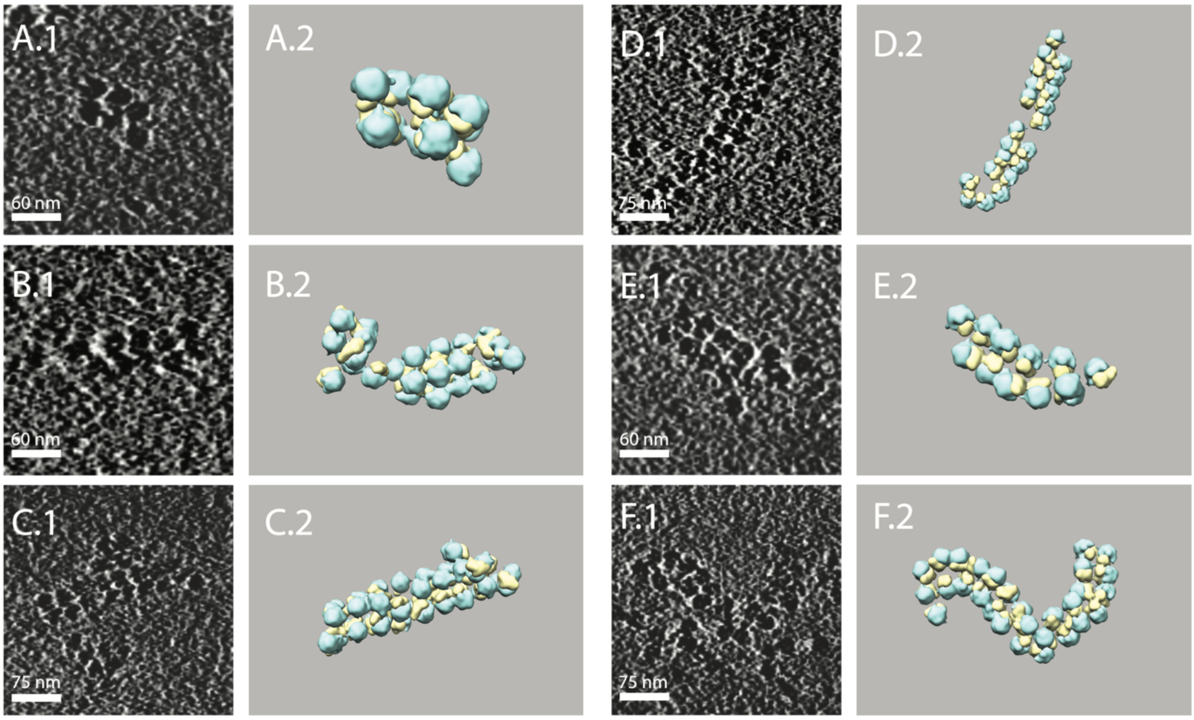

Figure 6.4: Detected cytosolic polysomes in tomograms of bacterial lysate. Examples of detected pseudo-helical (A, B, C) and pseudo-planar (D, E, F) topologies. Tomographic cross-sections (1) of the detected polysome are shown next to the rendered model of the corresponding Ribosome particles (2), small and large ribosomal subunit in yellow and blue respectively.

6.6.3 Local Model of ER-Associated Polysomes

From a training dataset of five tomograms of rough microsomes, exhaustive peak extraction during the template matching procedure yielded total of 9,063 particles. Subsequently, CPCA classification produced a subset of 7,081 positive ribosome particles. Visual inspection further reduced this set to 635 ribosome particles bound to the microsomal membrane.

Membrane bound ribosome particles were used to derive a P_{mRNA} model. Through local geometric analysis, 5' and 3' mRNA exit-to-entry vector clusters were identified (figure 6.5 A), and a 3D Gaussian function was fitted to 172 vectors in the 5' cluster (figure 6.5 B), designated as component P_{vec} in the P_{mRNA} model. Furthermore, the distribution of relative rotations associated to the 5' vector cluster was reduced to the in-plane rotation with respect to the membrane plane (figure 6.5 C) allowing the distribution to be described by one parameter only. P_{rot} was modeled as a 1D bimodal Gaussian mixture (figure 6.5 D).

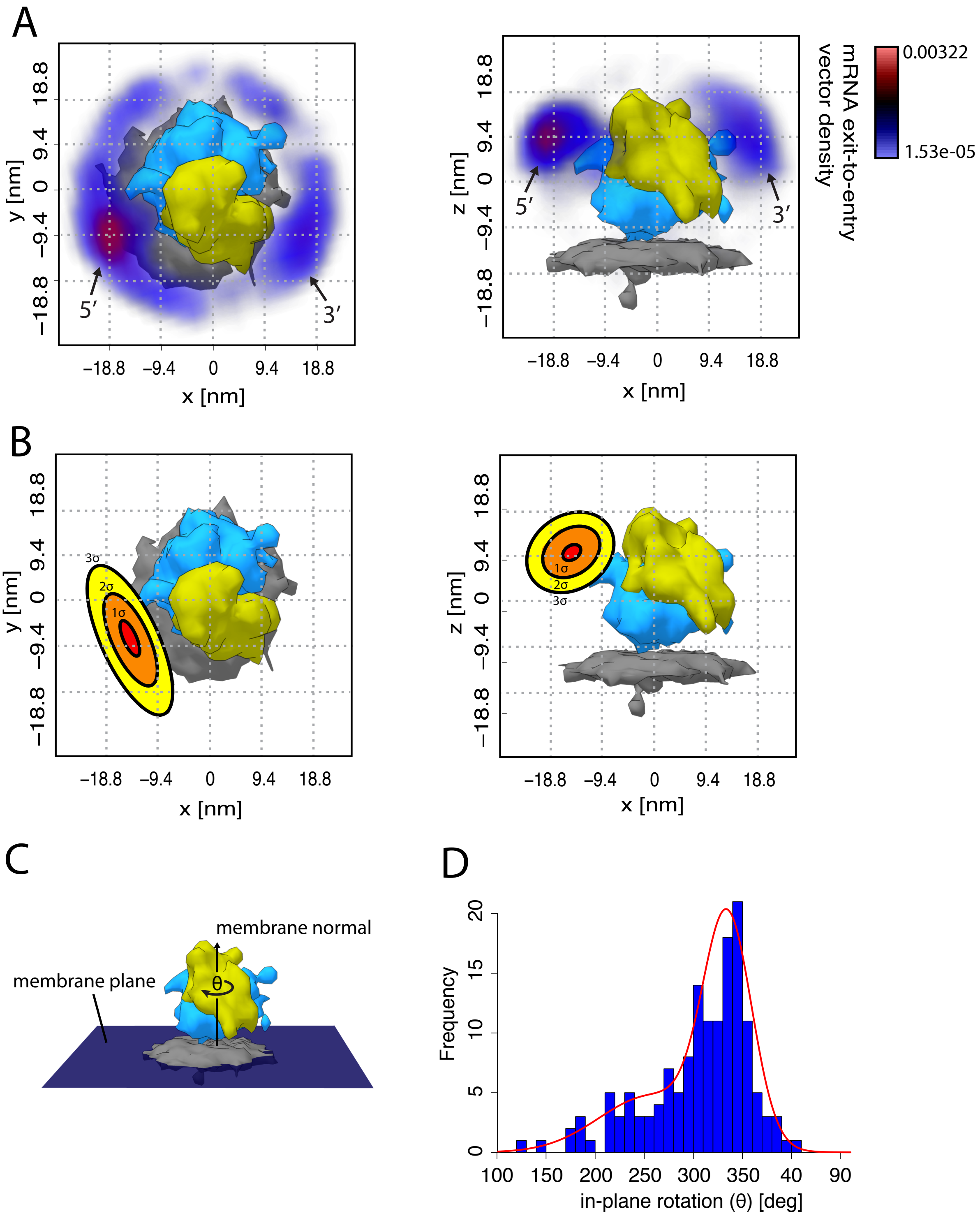

Figure 6.5: Local geometric analysis of neighboring ER-associated ribosomes. mRNA exit-to-entry vector and relative rotation distribution from a training dataset of 5 tomograms. (A) Density of mRNA exit-to-entry vectors. (B) 3D Gaussian fit for the 5' cluster with a goodness-of-fit $\chi^2_{red} = 1.16$. The distribution of (C) relative in-plane rotations with respect to the ER membrane, (D) associated with the 5' cluster of mRNA exit-to-entry vectors, show a peak at 335°. The rotation distribution was modeled by a bimodal Gaussian mixture (red curve) with a goodness-of-fit $\chi^2_{red} = 0.98$ ($\mu_1 = 333.8°, \sigma_1 = \ 24.8°, \mu_2 = 256.9°, \sigma_2 = 53.9°$).

6.6.4 Local Model of Cytosolic Mammalian Polysomes

Using the local model for ER-associated polysomes (section 6.6.3), the polysome detection method was applied to the training dataset of five rough microsome tomograms, with the aim of segregating ER-associated polysomes from cytosolic ribosomes, thus identifying potential candidates for cytosolic polysomes. From the set of 7,081 positive ribosome particles, the polysome detection method produced a subset of 5,914 monosome particles, i.e. ribosome particles uncorrelated with the P_{mRNA} model for ER-associate polysomes. This monosome class was subsequently used to generate the P_{mRNA} model for cytosolic mammalian polysomes.

Local geometric analysis of candidate particles for cytosolic polysomes elucidates two well defined clusters of mRNA exit-to-entry vectors, in agreement with 3' and 5' polysomic neighbors (figure 6.6 A). Furthermore, a 3D Gaussian function was fitted to the 5' cluster of 1,917 coordinates and defined as component P_{vec} (figure 6.6 B). Similarly to the bacterial model, the relative rotation distribution of the 5' cluster appears to be bimodal (figure 6.6 C). The rotation distribution was segmented into two clusters of 1,100 (figure 6.6 F) and 817 quaternions (figure 6.6 H). Two Bingham functions were fitted (figure 6.6 G, I), and P_{rot} was defined as the corresponding Bingham mixture model. Bingham modes show local arrangements consistent with bacterial t-t and t-b configurations (figure 6.6 D, E).

6.6.5 Analysis of ER-Associated and Cytosolic Mammalian Polysomes

Once P_{mRNA} models for ER-associated and cytosolic mammalian polysomes were derived from the training dataset, polysome detection was applied to a dataset of 13 tomograms of rough ER microsomes. Visual inspection of tomograms revealed two distinct populations of ribosomes, membrane bound and cytosolic (figure 6.11 A). Initially, a total of 25,186 ribosome particles from 13 tomograms, both true positives and false positives, were used as input to detect membrane bound polysomes. The method was applied using the P_{mRNA} model from ER-associated ribosomes (section 6.6.3). A second round of polysome detection was applied using the cytosolic mammalian P_{mRNA} model (section 6.6.4). In this manner, membrane bound (1,756 particles, 6.97% of the total amount of particles) and cytosolic polysome sets (7,533 particles, 29.91% of the total amount of particles) were extracted, and the remaining monosome particles from both detection rounds were merged into a monosome set (15,897 particles, 63.11% of the total amount of particles).

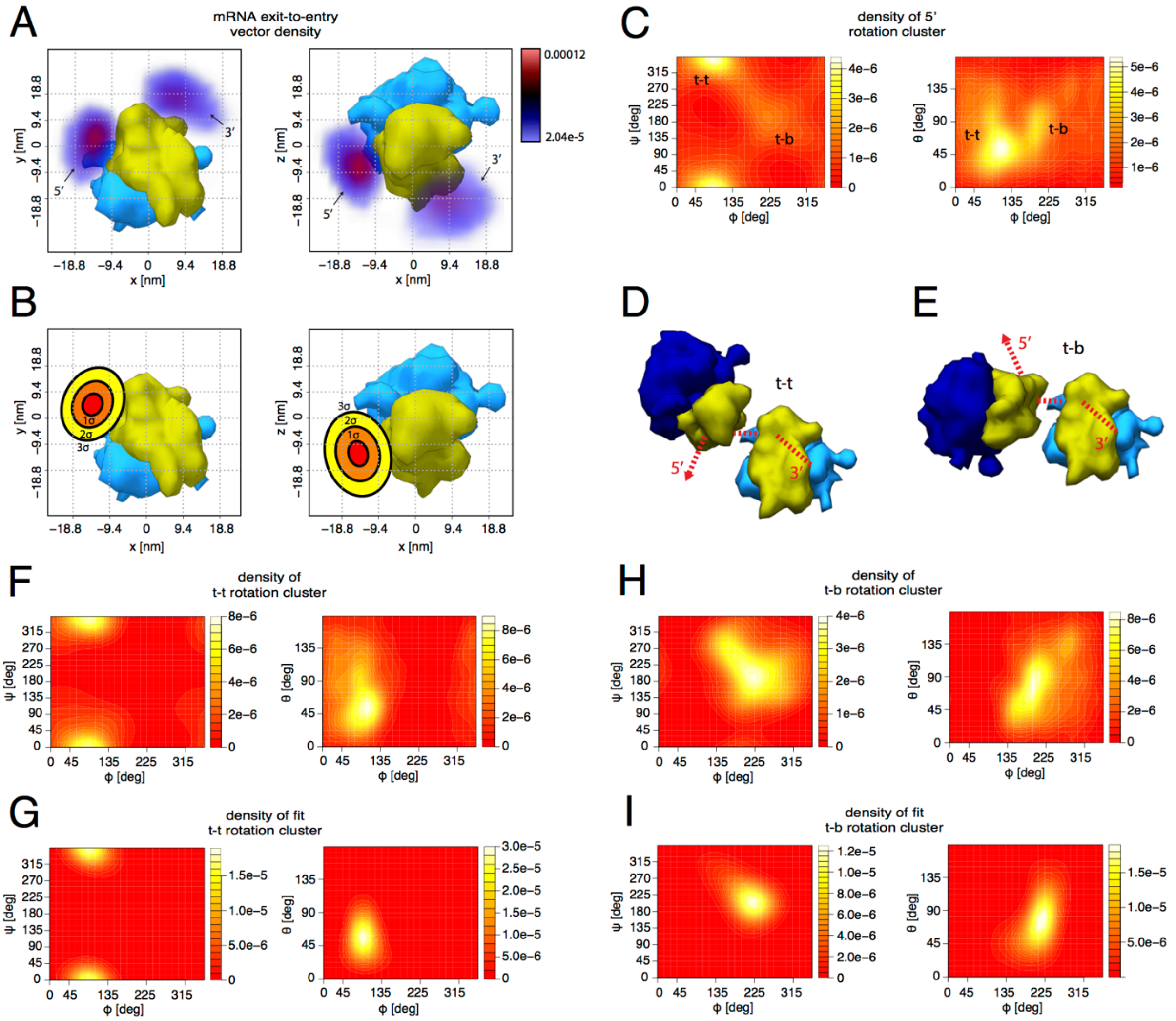

Figure 6.6: Local geometric analysis of cytosolic 80S Ribosomes. mRNA exit-to-entry vector and relative rotation distribution from a training dataset of 5 tomograms. (A) Density of mRNA exit-to-entry vectors. (B) 3D Gaussian fit for the 5' cluster with a goodness-of-fit $\chi^2_{red} = 1.02$. (C) Relative rotation distribution of the 5' clusters in Euler angle space showing 2 clear clusters, corresponding to the top-to-top (t-t) and top-to-bottom (t-b) arrangements. Rendered models of cluster modes, depicting t-t (D) and t-b (E) arrangements (large subunit of reference Ribosome in light blue, large subunit of neighboring Ribosome in dark blue, small subunits in yellow). Bingham distributions were fitted to t-t (F) and t-b (H) clusters, the goodness-of-fit for t-t cluster was $\chi^2_{red} = 1.17$ (G), while the fit for the t-b cluster (I) provided a $\chi^2_{red} = 1.15$.

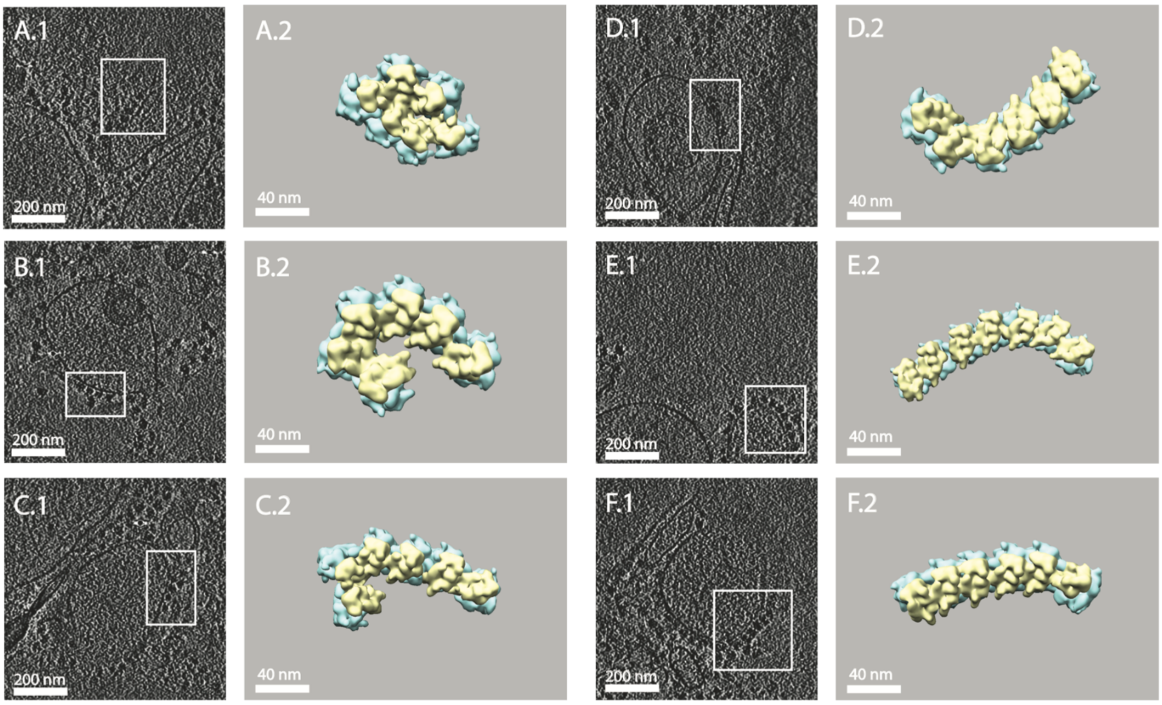

Figure 6.7: Detected ER-associated polysomes in tomograms of microsome preparations. Examples from a range of detected topologies, from slightly curved (E, F) to spiral-like (B, C) and circular (A). Tomographic cross-sections (1) of detected polysomes (framed region) are shown next to a rendered model (2), small and large ribosomal subunit in yellow and blue respectively.

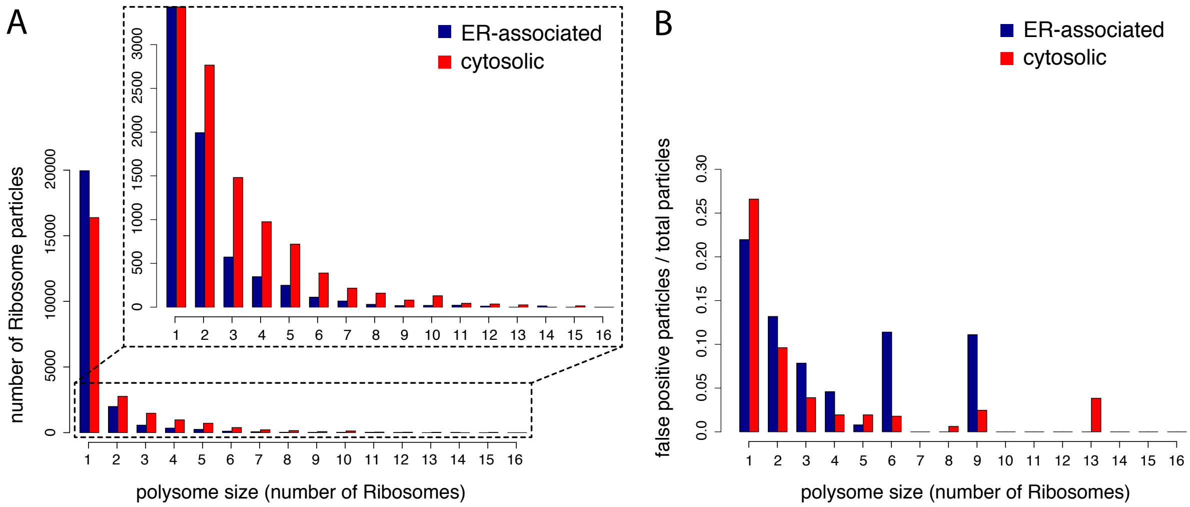

Figure 6.8: Histogram of polysome size, and relative amount of false positive particles by polysome size. Particle distribution by polysome size, for ER-associated and cytosolic polysomes (A). Relative concentration of false positive ribosome particles over detected polysome size (B).

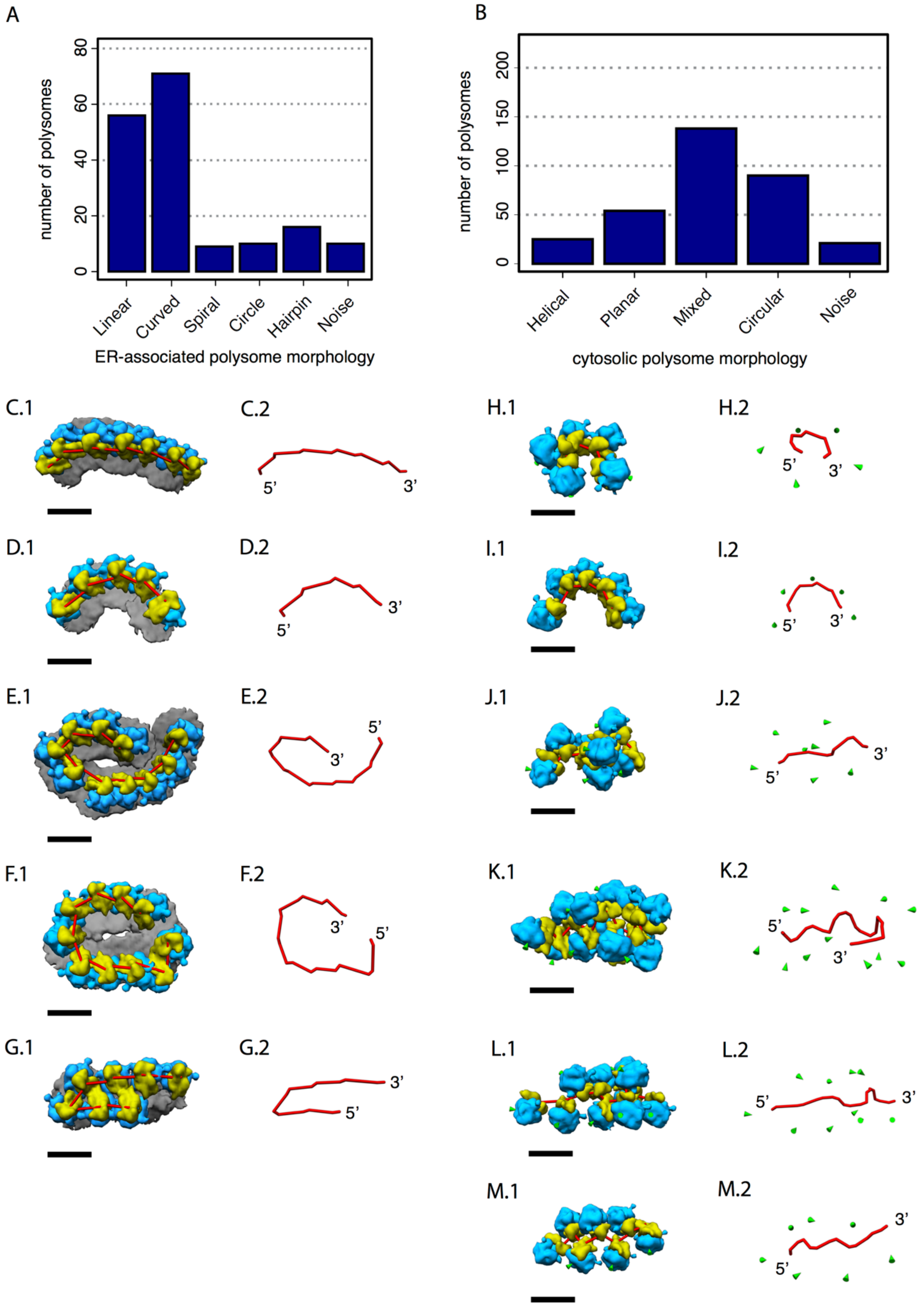

Figure 6.9: Morphology of detected ER-associated and cytosolic polysomes. Histogram of detected ER polysome (A) and cytosolic (B) topologies. Rendered models of predominant topology types, show top views of linear (C), curved (D), spiral-like (E), circular (F) and hairpin (G) ER-associated polysomes, and examples of cytosolic polysomes in circular (H, I), helical (J), mixed (K), and planar (L, M) conformations. Full models (1)

are shown next to their corresponding mRNA paths (2). Large and small ribosomal subunits in blue and yellow respectively, mRNA path in red, ER membrane depicted in gray, green cones marking peptide exit. Scale bar: 50 nm.

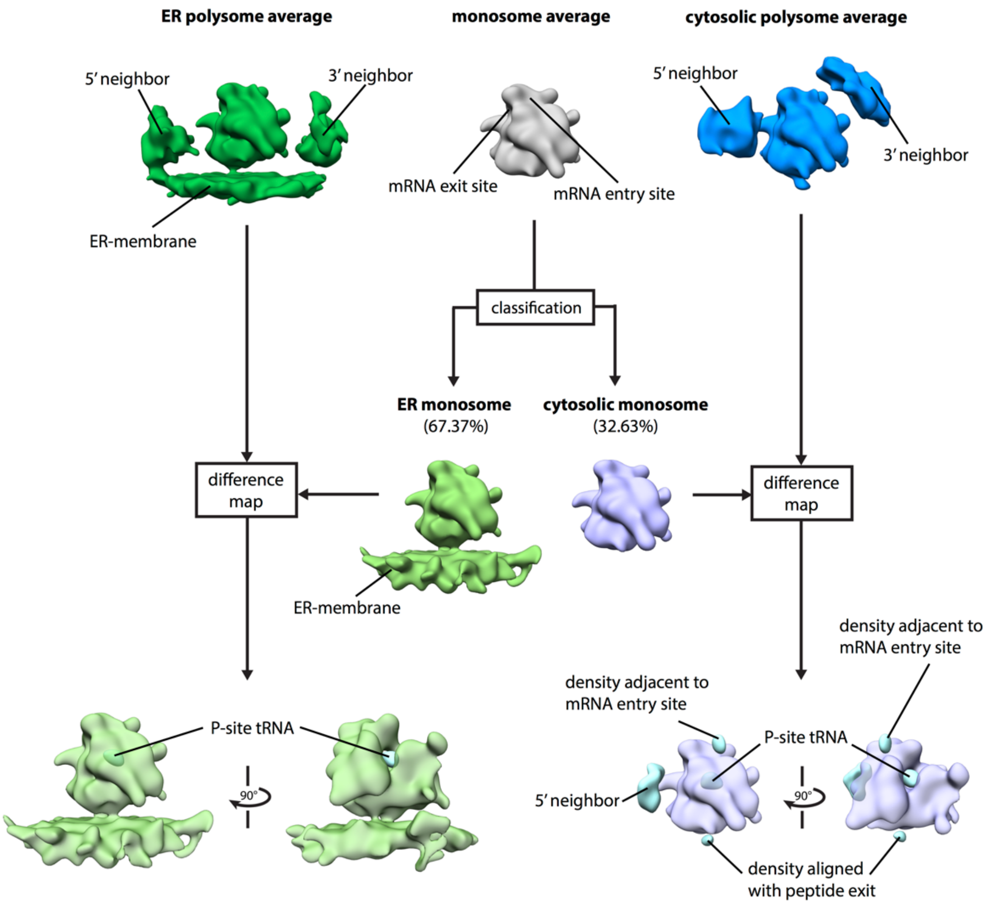

Figure 6.10: Subtomogram classification based on polysome detection of ER and mammalian cytosolic polysomes. Three subtomogram classes of 80S Ribosomes were derived from polysome detection: ER polysome, cytosolic polysome, and a monosome class. Averages of these 3 classes where computed, the monosome class was further subjected to CPCA classification to separate the ER-associated and cytosolic monosome classes. Difference maps of monosome and polysome classes (light blue iso-surface, 4σ) reveal significant density differences at the tRNA P-site. Since this analysis was mainly focused on large structural features, all averages were filtered to 5 nm resolution.

figure 6.7 shows examples of ER-associated polysomes, as detected by the proposed methodology. The distribution of ribosomes and concentration of false positive particles over polysome size is depicted on figure 6.8 A and 6.8 B respectively. figure 6.9 shows the amount of characteristic topology types found for both membrane-bound and cytosolic polysomes.

Furthermore, CPCA classification was applied to the subset of true positive ribosome particles in the monosome class (10,644 particles). Classification was focused on the ER membrane, i.e., a classification mask covering only the membrane region was used, yielding 2 subclasses, one with clear membrane signal (7,171 particles, 67.37% of the true positive particles in the monosome class), and one without (3,473 particles, 32.63% of the true positive particles in the monosome class). Averages of the ER and cytosolic monosome subclasses were compared against averages of true positive ER (1,380 particles) and cytosolic (6,675 particles) polysome sets, respectively (figure 6.10).

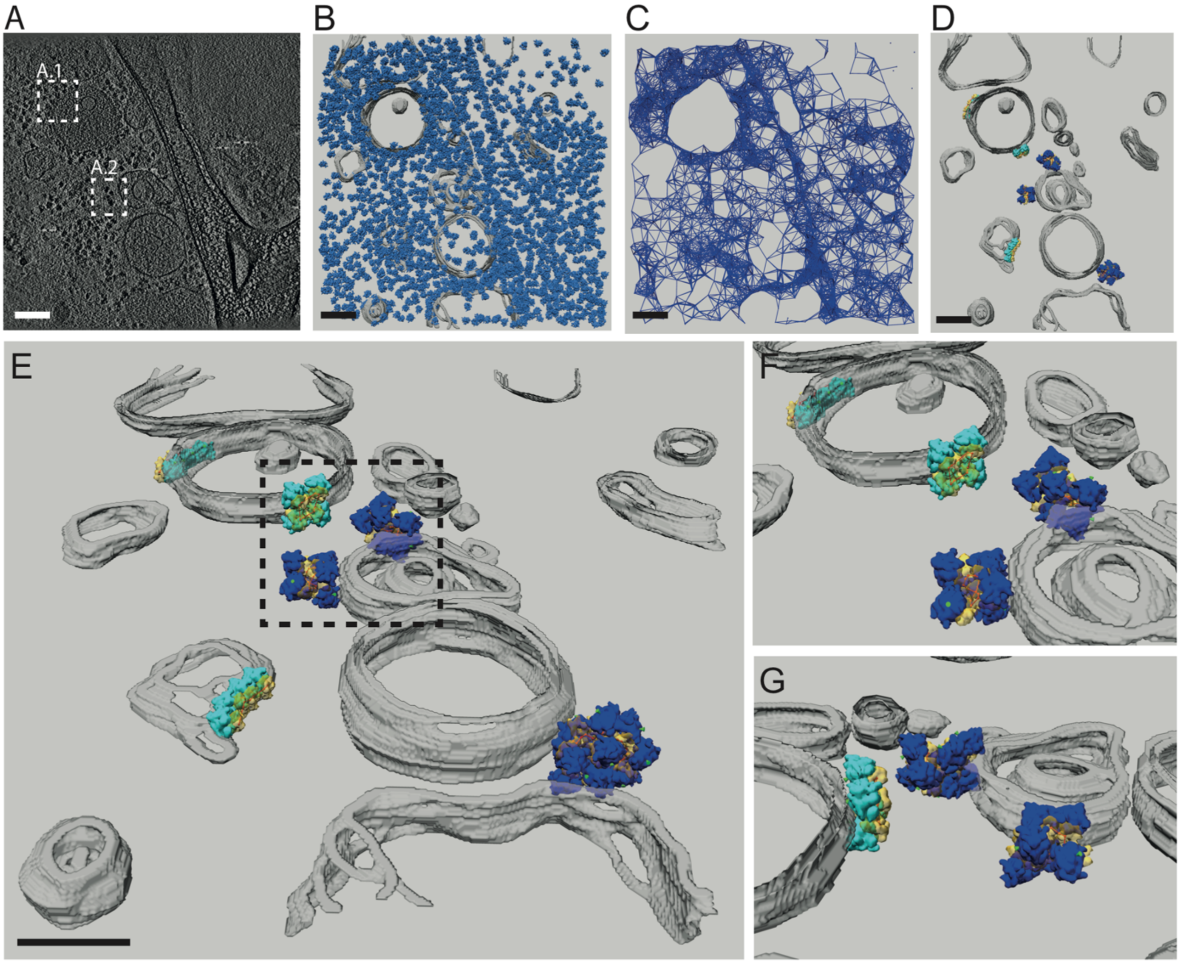

Figure 6.11: Experimental overview of the polysome detection methodology on tomograms of rough ER microsomes. (A) Cross-section of source tomogram, marking examples of membrane bound (A.1) and cytosolic (A.2) polysomes. (B) Rendered tomogram of detected 80S ribosome particles (blue) and microsomal membrane (gray). (C) Topology graph of ribosome particles. (D, E) Rendered examples of detected ER-associated (large subunit in light blue) and cytosolic (large subunit in dark blue) polysomes, small subunits in yellow, mRNA path in red, ER membrane in gray, green cones marking peptide exit. (F, G) Different perspectives of highlighted polysomes in (E). Scale bars: 250 nm.

6.7 Discussion

Here a method is presented for detection of polysomes in cryo-electron tomograms. Putative ribosome particles are represented as a 3D neighborhood graph, a probabilistic graphical model is embedded on the graph to classify ribosomes into polysomes, based on local geometric templates of predominant polysome topologies, as observed in training datasets. A neighborhood graph of the 3D distribution of ribosome positions in a tomogram can be efficiently computed using a k-dimensional tree for 3D range queries, allowing large position sets to be processed in affordable time. Edge weights in the graph are derived from the likelihood of mRNA connection, based on previously constructed models of the local geometric signature of polysome topologies. Graph theory was used to define polysome detection as a classification problem on a MRF. Finally, loopy belief propagation is used to cluster ribosome particles into polysomes.

We assessed the performance of the method on simulated and experimental tomograms of *E. coli* lysates. Quantitative evaluation shows encouraging results, in particular in terms of prediction accuracy (>96%). Moreover, the method was able to retrieve pseudo-helical and pseudo-planar topologies from ribosome-rich environments of bacterial lysate [F. Brandt et al., 2009]: 30S subunits are buried within the supramolecular structure, bringing mRNA exit and entry sites from adjacent ribosomes in close proximity. The 50S subunits are arranged outwards towards the cytosol, exposing the tRNA entry site and peptide exit site. It has been suggested that these characteristic topologies provide protection from mRNA decay by sequestering the mRNA molecule inside the polysome structure, shielding it from RNases [Arnold et al., 1998], while the positions of the 50S subunits provide space for the co-translational folding machinery, protecting nascent polypeptides from potentially toxic aggregation [Deuerling et al., 1999; Teter et al., 1999] in the highly crowded cellular environment. Furthermore, detected ER-associated polysomes are also in good agreement with previously observed topologies of circular, spiral, and curved polysome arrangements [S. Y. Lee et al., 1971; Palade, 1964]. Neighboring 40S subunits tend to minimize the distance between their mRNA exit and entry sites, bringing the mRNA molecule into a smoothly curved path, while the 60S subunits point their peptide exit sites towards the ER-membrane, exposing their nascent polypeptides to the translocation machinery [Pfeffer et al., 2012]. Interestingly, detected topologies of cytosolic polysomes in tomograms of bacterial lysate and mammalian microsomal preparations were remarkably similar, and in agreement with those observed in CET studies of human cells [F. Brandt et al., 2010], indicating that this methodology may also enable structural analysis of polysomes under physiological conditions. Moreover, this method was successfully applied in a cryo-FIB milled tomogram from the nuclear periphery of a HeLa cell, with the aim of detecting polysomes bound to the nuclear envelope and the ER [Mahamid et al., 2016].

Theoretically speaking, since the polysome detection method assumes statistical independence of graph edges, the probability of an mRNA path tends to collapse as the path length increases. This methodological property could preclude detection of long polysomes, while the presence of this systemic bias remains to be evaluated, it could explain why the majority of 80S ribosome particles were classified as monosomes or assigned to small polysomes, with only few inferred polysomes larger than 6 ribosomes long. However, subtomogram analysis of the predicted polysome and monosome classes does not suggest a significant bias, averages of ribosome particles from cytosolic and ER-associated polysomes classes show clear densities co-localizing with polysomic neighbors, as opposed the monosome average. Subsequent comparison of polysome averages with their corresponding monosome class average, show a significant density difference at the tRNA P-site, consistent with passive and actively translating ribosomes. It is noteworthy that the relative concentration of false positive 80S particles was significantly higher in the monosome class and in small polysomes (2 – 3 ribosome long), as expected since their local geometric configurations are unlikely to correlate with P_{mRNA}. Furthermore, the proposed method was applied to detect yeast cytosolic polysomes in tomograms of monosome and polysome containing fractions of cell lysate. The analysis yielded a distribution of inferred polysome lengths consistent with polysome profiling by sucrose gradient centrifugation of cell lysates [Pospísek & Valásek, 2013], thus providing biochemical validation of the above-described polysome detection method [Burbaum, 2015].

The method still holds considerable potential for improvement. The P_{mRNA} model aims to approximate the likelihood of mRNA connection between two adjacent ribosomes, given their relative geometric arrangement. Here, the mRNA exit-to-entry vector and relative rotation of a ribosome dimer are considered statistically independent. This is certainly a simplified model, since it is expected that both components are functions of the same physiological elements driving the pressure for the preferred supramolecular arrangements of polysomes e.g., shape of the ribosome, spatial restrictions in the cellular environment, and optimization of the folding process of nascent peptides. However, this convoluted intracellular process is challenging to model. It is noteworthy that tomograms of different biological systems and sample preparations with varying amounts of ribosome density can provide significantly different parameters for P_{mRNA}, overfitting the local geometric model towards a specific dataset.

Furthermore, polysome detection is defined in terms of potential functions within a Markov random field (equations 6.9 and 6.10), which theoretically speaking, do not have specific probabilistic interpretations. While direct probabilistic modeling of the polysome detection problem could be possible using a different graphical models (e.g. a Bayesian network or a factor graph), their performance for this particular application is still to be evaluated. Another caveat of the proposed methodology is the lack of theoretical assurances for the convergence of loopy belief propagation into a stable solution. Nevertheless, quantitative evaluation (table 6.1) and visual inspection of detected polysomes confirms that loopy belief propagation in conjunction with the simplified P_{mRNA} model, is capable of approximating a MAP classification solution with enough accuracy to detect characteristic polysome topologies.

Moreover, the method hinges on the accuracy of ribosome detection. False positives and negatives from template matching can potentially affect the performance of the polysome detection method. False positives are not likely to generate large errors at polysome detection; the method regards them as monosome particles since the 3D arrangement of their neighboring particles is highly unlikely to correlate with the P_{mRNA} model. In contrast, false negatives may significantly reduce chance of providing the required ribosome particles necessary for detecting complete polysome sequences, in such cases, the detection method will identify sequence fragments as separate polysomes. For the purpose of polysome detection, it is preferable to oversample the amount of template matching peaks.

7. Conclusion

This thesis presents a generalized statistical method for quantifying the 3D organization of macromolecular complexes within cryo-electron tomograms. Chapter 4 describes the procedure for local geometric analysis of adjacent macromolecules (e.g. center-to-center vectors and relative rotations), with the aim of identifying particle pairs that have specific geometric relationships. Previous visual proteomics studies have been restricted to the generation of macromolecular atlases in cellular volumes, providing only the 3D position and orientation of particles [Beck et al., 2009; Ortiz et al., 2006], but not progressing further. While some CET studies have presented near-neighbor analyses, these applications have been focused on ribosomes [F. Brandt et al., 2009, 2010; Pfeffer et al., 2012; Pfeffer, Woellhaf, et al., 2015]. This thesis proposes an analysis that can be applied to any macromolecular complex, incorporating macromolecule symmetry and can generate RDFs, allowing statistical description of macromolecular distributions. Implementation of the toolbox for local geometric analysis was based on the PyTom software [Hrabe et al., 2012], in the hope that future users might find it useful.

In chapter 5, the method described in chapter 4 was applied to study the local organization of RuBisCO complexes within the pyrenoid of *C. reinhardtii* cells. Subtomogram analysis yielded a 16 Å *in situ* structure of the RuBisCO complex. Local geometric analysis of RuBisCO particles suggests a fluid-like pyrenoid matrix. The RDF (figure 5.8 A) indicates short-range order but no long-range order, with the first NN radial shell located within a center-to-center distance of ~10-18 nm. Furthermore, the distribution of center-to-center vectors and relative rotations suggests high flexibility in particle pair configurations. Finally, predominant 3D arrangements of RuBisCO complex pairs were identified and unified into a geometric model of the unit cell of RuBisCO complexes, with a 3D configuration similar to that of closely packed spheres.

Chapter 6 progresses from local neighborhood analysis to the identification of supramolecular structures, presenting a graph-based method for probabilistic polysome detection. This method is the first of its kind, there has not been any method for CET that allows detection of higher-order structures. Moreover, since this method utilizes only geometric information (i.e. positions and orientations of ribosomes), it is intrinsically compatible with visual proteomics. Briefly, the approach outlined in chapter 4 was applied to a training dataset of tomograms to derive statistical models for the 3D configuration of neighboring ribosomes within polysomes. These P_{mRNA} models were then used as local geometric templates within a Markov random field to detect flexible extrapolations of ribosome pair arrangements. Quantitative evaluation of the method indicated a 96% prediction accuracy. Furthermore, visual inspection of the predicted polysomes confirmed the identification of characteristic polysome structures.

The proposed methodologies in this thesis can still be improved and extended. For the quaternion analysis presented in chapter 4, it would be useful to implement a maximum-likelihood method for Bingham mixture model fitting. An expectation–maximization method could use the quaternion clustering procedure described in chapter 4 as an initialization strategy to increase the rate of convergence. Furthermore, the P_{mRNA} model proposed in chapter 6 still has room for improvement. A simplified P_{mRNA} model was used, which considers the distributions of mRNA exit-to-entry vectors and relative rotations from adjacent ribosomes to be statistically independent. However, it is clear that these variables are correlated. Thus, a P_{mRNA} model that incorporates such a relationship might yield better results.

Visual proteomics aims to quantify the 3D organization of a variety of complexes within the cellular landscapes of *in situ* cryo-electron tomograms. To accomplish this goal, it will be important to develop geometric analysis methods that handle multiple particle classes (i.e. different types macromolecular complexes). An input dataset could be generated by template matching a tomogram with a library of structural templates, followed by refinement with subtomogram averaging and classification. In a multi-class geometric analysis scheme, one class of macromolecules could be set as 'reference' particles (e.g. RuBisCO complexes), while a second class could be defined as 'neighbor' particles (e.g. RuBisCO activases). This would enable the detection of particles in the neighbor class that are adjacent to reference particles. Subsequent geometric analysis with the methodology outlined in chapter 4 would allow the identification of interacting partners and their specific 3D arrangements. Fitting atomic models into subtomogram averages of these interacting partners would help elucidate the molecular surfaces that mediate their interaction. An additional research direction worth exploring is a graph-based method for supramolecular structure detection. This approach can handle multiple particle classes and can identify flexible structures that have higher topological complexity than the linear sequences of macromolecules found in polysomes.

Visual proteomics will require methods for the parallel statistical analysis of geometric information from numerous classes of molecular complexes. While it is true that geometry-based methods hinge on the accuracy of macromolecule identification in cryo-electron tomograms, recent advances in TEM hardware, including direct electron detectors [Cheng, 2015] and the Volta phase plate [Fukuda et al., 2015; Khoshouei et al., 2016], have increased the quality of micrographs by an order of magnitude and greatly improved the accuracy of tomographic reconstructions. Thus, higher specificity and sensitivity of macromolecule identification is to be expected, empowering geometric analysis to dissect the complex molecular organization of native cellular volumes.

Abbreviations

2D	two-dimensional
3D	three-dimensional
3PG	3-phosphoglycerate
aa-tRNA	aminoacyl tRNA
AC3D	auto-focused classification
ADP	adenosine diphosphate
ATP	adenosine triphosphate
C. reinhardtii	Chlamydomonas reinhardtii
CCD	charged coupled device
CCM	carbon concentration mechanism
CET	cryo-electron tomography
CPCA	constrained principal component analysis
CPU	central processing unit
cryo-FIB	cryo-focused ion beam
CTF	contrast transfer function
E. coli	Escherichia coli
EF-G	elongation factor G
EF-Tu	elongation factor Tu
EM grid	electron microscopy grid
EPYC1	essential pyrenoid component 1
ER	endoplasmic reticulum
FEG	field emission gun
FN	false negative
FP	false positive
FRM	fast rotation matching
FSC	Fourier shell correlation
G3P	glyceraldehyde-3-phosphate
GB	Gigabyte
GMM	Gaussian mixture model
MAP	maximum a posteriori
MRF	Markov random field
mRNA	messenger ribonucleic acid
MTF	modulation transfer function
NADPH	nicotinamide adenine dinucleotide phosphate
NE	nuclear envelope
NMR	nuclear magnetic resonance
NN	near-neighbor
NPC	nuclear pore complex
PDB	protein data bank
Pi	inorganic phosphate
RAM	Random-access memory
rbcL	RuBisCO large subunit
RBCS	RuBisCO small subunit
RCA1	RuBisCO activase

RDF	radial distribution function
RS	real space
RuBisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
RuBP	ribulose-1,5-bisphosphate
SNR	signal-to-noise ratio
SPA	single particle analysis
t-b	top-to-bottom
t-t	top-to-top
TEM	transmission electron microscope
TN	true negative
TP	true positive
tRNA	transfer ribonucleic acid
WBP	weighted backprojection

Bibliography

Amat, F., Castaño-Diez, D., Lawrence, A., Moussavi, F., Winkler, H., & Horowitz, M. (2010). Alignment of Cryo-Electron Tomography Datasets. In *Methods in Enzymology* (Vol. 482, pp. 343–367). http://doi.org/10.1016/S0076-6879(10)82014-2

Andersson, I., & Backlund, A. (2008). Structure and function of Rubisco. *Plant Physiology and Biochemistry*, *46*(3), 275–291. http://doi.org/10.1016/j.plaphy.2008.01.001

Arnold, T. E., Yu, J., & Belasco, J. G. (1998). mRNA stabilization by the ompA 5' untranslated region: two protective elements hinder distinct pathways for mRNA degradation. *RNA (New York, N.Y.)*, *4*(3), 319–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9510333

Asano, S., Engel, B. D., & Baumeister, W. (2016). In Situ Cryo-Electron Tomography : A Post-Reductionist Approach to Structural Biology. *Journal of Molecular Biology*, *428*(2), 332–343. http://doi.org/10.1016/j.jmb.2015.09.030

Asano, S., Fukuda, Y., Beck, F., Aufderheide, A., Förster, F., Danev, R., & Baumeister, W. (2015). Proteasomes. A molecular census of 26S proteasomes in intact neurons. *Science (New York, N.Y.)*, *347*(6220), 439–42. http://doi.org/10.1126/science.1261197

Beck, M., Malmström, J. a, Lange, V., Schmidt, A., Deutsch, E. W., & Aebersold, R. (2009). Visual proteomics of the human pathogen Leptospira interrogans. *Nature Methods*, *6*(11), 817–23. http://doi.org/10.1038/nmeth.1390

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, *18*(9), 509–517. http://doi.org/10.1145/361002.361007

Berner, A., Bokeloh, M., Wand, M., Schilling, A., & Seidel, H.-P. (2008). A Graph-Based Approach to Symmetry Detection. In *Proceedings of the Fifth Eurographics / IEEE VGTC Conference on Point-Based Graphics* (pp. 1–8). inproceedings, Aire-la-Ville, Switzerland, Switzerland: Eurographics Association. http://doi.org/10.2312/VG/VG-PBG08/001-008

Berrou, C., Glavieux, A., & Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Proceedings of ICC '93 - IEEE International Conference on Communications* (Vol. 2, pp. 1064–1070). IEEE. http://doi.org/10.1109/ICC.1993.397441

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. *Springer* (Vol. 16). http://doi.org/10.1117/1.2819119

Blake, A., Kohli, P., & Rother, C. (2011). *Markov Random Fields for Vision and Image Processing*. The MIT Press.

Bollobás, B. (1979). *Graph Theory* (Vol. 63). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4612-9967-7

Brandt, F., Carlson, L.-A., Hartl, F. U., Baumeister, W., & Grünewald, K. (2010). The three-dimensional organization of polyribosomes in intact human cells. *Molecular Cell*, *39*(4), 560–9. http://doi.org/10.1016/j.molcel.2010.08.003

Brandt, F., Etchells, S. a, Ortiz, J. O., Elcock, A. H., Hartl, F. U., & Baumeister, W. (2009). The Native 3D Organization of Bacterial Polysomes. *Cell*, *136*(2), 261–271. http://doi.org/10.1016/j.cell.2008.11.016

Brandt, S., Heikkonen, J., & Engelhardt, P. (2001a). Automatic Alignment of Transmission Electron Microscope Tilt Series without Fiducial Markers. *Journal of Structural Biology*, *136*(3), 201–213. http://doi.org/10.1006/jsbi.2001.4443

Brandt, S., Heikkonen, J., & Engelhardt, P. (2001b). Multiphase Method for Automatic Alignment of Transmission Electron Microscope Images Using Markers. *Journal of Structural Biology*, *133*(1), 10–22. http://doi.org/10.1006/jsbi.2001.4343

Buchan, J. R., & Stansfield, I. (2007). Halting a cellular production line: responses to ribosomal pausing during translation. *Biology of the Cell*, *99*(9), 475–487. http://doi.org/10.1042/BC20070037

Burbaum, L. (2015). *Strukturanalyse von Hefe-Polysomen mittels Kryo-Elektronentomographie*. Westfälische Wilhelms-Universität Münster.

Cargill III, G. S. (1975). Structure of Metallic Alloy Glasses. In *Solid State Physics* (Vol. Volume 30, pp. 227–320). http://doi.org/http://dx.doi.org/10.1016/S0081-1947(08)60337-9

Chandramouli, P., Topf, M., Ménétret, J.-F., Eswar, N., Cannone, J. J., Gutell, R. R., … Akey, C. W. (2008). Structure of the Mammalian 80S Ribosome at 8.7 Å Resolution. *Structure*, *16*(4), 535–548. http://doi.org/10.1016/j.str.2008.01.007

Chen, Y., & Förster, F. (2014). Iterative reconstruction of cryo-electron tomograms using nonuniform fast Fourier transforms. *Journal of Structural Biology*, *185*(3), 309–316. http://doi.org/10.1016/j.jsb.2013.12.001

Chen, Y., Hrabe, T., Pfeffer, S., Pauly, O., Mateus, D., Navab, N., & Forster, F. (2012). Detection and identification of macromolecular complexes in cryo-electron tomograms using support vector machines. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1373–1376). IEEE. http://doi.org/10.1109/ISBI.2012.6235823

Chen, Y., Pfeffer, S., Fernández, J. J., Sorzano, C. O. S., & Förster, F. (2014). Autofocused 3D Classification of Cryoelectron Subtomograms. *Structure*, *22*(10), 1528–1537. http://doi.org/10.1016/j.str.2014.08.007

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M., & Förster, F. (2013). Fast and accurate reference-free alignment of subtomograms. *Journal of Structural Biology*, *182*(3), 235–245. http://doi.org/10.1016/j.jsb.2013.03.002

Cheng, Y. (2015). Single-Particle Cryo-EM at Crystallographic Resolution. *Cell*, *161*(3), 450–

457. http://doi.org/10.1016/j.cell.2015.03.049

Clifford, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems. A Volume in Hon- our of John M. Hammersley* (pp. 19–32). Oxford University Press.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, *42*(2–3), 393–405. http://doi.org/10.1016/0004-3702(90)90060-D

Crowther, R. A., DeRosier, D. J., & Klug, A. (1970). The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *317*(1530), 319–340. http://doi.org/10.1098/rspa.1970.0119

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. http://doi.org/10.1.1.133.4884

Deuerling, E., Schulze-Specking, a, Tomoyasu, T., Mogk, a, & Bukau, B. (1999). Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature*, *400*(6745), 693–696. http://doi.org/10.1038/23301

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. http://doi.org/10.1007/BF01386390

Dubochet, J., Adrian, M., Chang, J. J., Homo, J. C., Lepault, J., McDowall, a W., & Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, *21*(2), 129–228. http://doi.org/10.1017/S0033583500004297

Ellis, R. J. (1979). The most abundant protein in the world. *Trends in Biochemical Sciences*, *4*(11), 241–244. http://doi.org/10.1016/0968-0004(79)90212-3

Ellis, R. J. (2010). Biochemistry: Tackling unintelligent design. *Nature*, *463*(7278), 164–165. http://doi.org/10.1038/463164a

Engel, B. D., Schaffer, M., Kuhn Cuellar, L., Villa, E., Plitzko, J. M., & Baumeister, W. (2015). Native architecture of the Chlamydomonas chloroplast revealed by in situ cryo-electron tomography. *eLife*, *4*(4), 1–29. http://doi.org/10.7554/eLife.04889

Faruqi, A. R., & Mcmullan, G. (2011). Electronic detectors for electron microscopy. *Quarterly Reviews of Biophysics*, *3*(44), 357–390.

Fernández, J. J., Li, S., & Crowther, R. A. (2006). CTF determination and correction in electron cryotomography. *Ultramicroscopy*, *106*(7), 587–596. http://doi.org/10.1016/j.ultramic.2006.02.004

Förster, F., Han, B.-G., & Beck, M. (2010). Visual Proteomics. In *Methods in Enzymology* (1st ed., Vol. 483, pp. 215–243). Elsevier Inc. http://doi.org/10.1016/S0076-6879(10)83011-3

Förster, F., & Hegerl, R. (2007). Structure Determination In Situ by Averaging of Tomograms. In *Methods in Cell Biology* (Vol. 2007, pp. 741–767). http://doi.org/10.1016/S0091-679X(06)79029-X

Förster, F., Pruggnaller, S., Seybert, A., & Frangakis, A. S. (2008). Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of Structural Biology*, *161*(3), 276–286. http://doi.org/10.1016/j.jsb.2007.07.006

Frangakis, A. S., Bohm, J., Forster, F., Nickell, S., Nicastro, D., Typke, D., … Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proceedings of the National Academy of Sciences*, *99*(22), 14153–14158. http://doi.org/10.1073/pnas.172520299

Frank, G., Bartesaghi, A., Kuybeda, O., Borgnia, M. J., White, T. A., Sapiro, G., & Subramaniam, S. (2012). Computational separation of conformational heterogeneity using cryo-electron tomography and 3D sub-volume averaging. *Journal of Structural Biology*, *178*(2), 165–76. http://doi.org/10.1016/j.jsb.2012.01.004

Frank, J. (2006). *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. book, Oxford: Oxford University Press.

Freeman, W. T., Pasztor, E. C., Carmichael, O. T., Hall, S., & Ave, F. (2000). Learning Low-Level Vision, *40*(1), 25–47.

Freidman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226. http://doi.org/10.1145/355744.355745

Frey, B. J., Koetter, R., & Petrovic, N. (2001). Very loopy belief propagation for unwrapping phase images. In *Advances in Neural Information Processing Systems* (Vol. 14, pp. 1–7). http://doi.org/10.1.1.11.7737

Frey, B., & MacKay, D. (1998). A Revolution: Belief Propagation in Graphs With Cycles. *Advances in Neural Information Processing Systems*, *10*, 479. http://doi.org/10.1.1.15.6659

Fukuda, Y., Laugks, U., Lučić, V., Baumeister, W., & Danev, R. (2015). Electron cryotomography of vitrified cells with a Volta phase plate. *Journal of Structural Biology*, *190*(2), 143–154. http://doi.org/10.1016/j.jsb.2015.03.004

Glaeser, R. M. (1971). Limitations to significant information in biological electron microscopy as a result of radiation damage. *Journal of Ultrasructure Research*, *36*(3–4), 466–482. http://doi.org/10.1016/S0022-5320(71)80118-1

Glover, J., & Kaelbling, L. P. (2013). Tracking 3-D rotations with the quaternion Bingham filter. *MIT CSAIL*. Retrieved from http://hdl.handle.net/1721.1/78248

Glover, J., Rusu, R., & Bradski, G. (2011). Monte Carlo Pose Estimation with Quaternion Kernels and the Bingham Distribution. In H. F. Durrant-Whyte, N. Roy, & P. Abbeel (Eds.), *Robotics: Science and Systems VII* (p. 97). inproceedings, Robotics: Science and Systems Foundation. http://doi.org/10.15607/RSS.2011.VII.015

Han, H. M., Zuber, B., & Dubochet, J. (2008). Compression and crevasses in vitreous sections under different cutting conditions. *Journal of Microscopy*, *230*(2), 167–171. http://doi.org/10.1111/j.1365-2818.2008.01972.x

Hartl, F. U., & Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. *Nature Structural & Molecular Biology*, *16*(6), 574–581. http://doi.org/10.1038/nsmb.1591

Hartman, F. C., & Harpel, M. R. (1994). Structure, Function, Regulation, and Assembly of D-Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase. *Annual Review of Biochemistry*, *63*(1), 197–232. http://doi.org/10.1146/annurev.bi.63.070194.001213

Hauser, T., Popilka, L., Hartl, F. U., & Hayer-Hartl, M. (2015). Role of auxiliary proteins in Rubisco biogenesis and function. *Nature Plants*, *1*(6), 15065. http://doi.org/10.1038/nplants.2015.65

Hrabe, T., Chen, Y., Pfeffer, S., Kuhn Cuellar, L., Mangold, A.-V., & Förster, F. (2012). PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *Journal of Structural Biology*, *178*(2), 177–188. http://doi.org/10.1016/j.jsb.2011.12.003

Hrabe, T., & Förster, F. (2011). Structure Determination by Single Particle Cryo-electron Tomography. In *Encyclopedia of Life Sciences* (pp. 1–11). Chichester, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9780470015902.a0023175

Indyk, P., O'Rourke, J., & Goodman, J. (2004). Nearest Neighbors in High-Dimensional Spaces. In *Handbook of Discrete and Computational Geometry* (pp. 877–892). http://doi.org/10.1.1.10.3826

Khoshouei, M., Pfeffer, S., Baumeister, W., Förster, F., & Danev, R. (2016). Subtomogram analysis using the Volta phase plate. *Journal of Structural Biology*, (May). http://doi.org/10.1016/j.jsb.2016.05.009

Kremer, J. R., Mastronarde, D. N., & McIntosh, J. R. (1996). Computer Visualization of Three-Dimensional Image Data Using IMOD. *Journal of Structural Biology*, *116*(1), 71–76. http://doi.org/10.1006/jsbi.1996.0013

Krivanek, O. L., Friedman, S. L., Gubbens, A. J., & Kraus, B. (1995). An imaging filter for biological applications. *Ultramicroscopy*, *59*(1–4), 267–282. http://doi.org/10.1016/0304-3991(95)00034-X

Kschischang, F. R., & Frey, B. J. (1998). Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Journal on Selected Areas in Communications*, *16*(2), 219–230. http://doi.org/10.1109/49.661110

Kuffner, J. J. (2004). Effective sampling and distance metrics for 3D rigid body path planning. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004* (Vol. 4, p. 3993–3998 Vol.4). IEEE. http://doi.org/10.1109/ROBOT.2004.1308895

Lee, D. T., & Wong, C. K. (1977). Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees. *Acta Informatica*, *9*(1), 23–29. http://doi.org/10.1007/BF00263763

Lee, S. Y., Krsmanovic, V., & Brawerman, G. (1971). Attachment of ribosomes to membranes during polysome formation in mouse sarcoma 180 cells. *The Journal of Cell Biology*, *49*(3), 683–91. http://doi.org/10.1083/jcb.49.3.683

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. http://doi.org/10.1109/TIT.1982.1056489

Louca, D., & Takeshi, E. (1999). Local lattice distortions in <math display="inline"> <mrow> <msub> <mrow> <mi mathvariant="normal">La</mi> </mrow> <mrow> <mn>1</mn> <mi>−</mi> <mi>x</mi> </mrow> </msub> </mrow> <mrow> <msub> <mrow> <mi mathvariant="normal">Sr</mi> </mrow> <mrow> <mi>x</. *Physical Review B*, *59*(9), 6193–6204. http://doi.org/10.1103/PhysRevB.59.6193

Lucić, V., Förster, F., & Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. *Annual Review of Biochemistry*, *74*, 833–65. http://doi.org/10.1146/annurev.biochem.73.011303.074112

Lucić, V., Rigort, A., & Baumeister, W. (2013). Cryo-electron tomography: The challenge of doing structural biology in situ. *Journal of Cell Biology*, *202*(3), 407–419. http://doi.org/10.1083/jcb.201304193

Mackinder, L. C. M., Meyer, M. T., Mettler-Altmann, T., Chen, V. K., Mitchell, M. C., Caspari, O., … Jonikas, M. C. (2016). A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. *Proceedings of the National Academy of Sciences*, *113*(21), 5958–5963. http://doi.org/10.1073/pnas.1522866113

Mahamid, J., Pfeffer, S., Schaffer, M., Villa, E., Danev, R., Cuellar, L. K., … Baumeister, W. (2016). Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science (New York, N.Y.)*, *351*(6276), 969–72. http://doi.org/10.1126/science.aad8857

Marabini, R., Herman, G. T., & Carazo, J. M. (1998). 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy*, *72*(1–2), 53–65. http://doi.org/10.1016/S0304-3991(97)00127-7

Mardia, K. V. (1975). Characterization of Directional Distributions. In *Statistical Distributions in Scientific Work* (pp. 365–385). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-010-1848-7_34

Mastronarde, D. N. (1997). Dual-Axis Tomography: An Approach with Alignment Methods That Preserve Resolution. *Journal of Structural Biology*, *120*(3), 343–352. http://doi.org/10.1006/jsbi.1997.3919

McEliece, R. J., MacKay, D. J. C., & Jung-Fu Cheng. (1998). Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications*, *16*(2), 140–152. http://doi.org/10.1109/49.661103

Meyer, M. T., Genkov, T., Skepper, J. N., Jouhet, J., Mitchell, M. C., Spreitzer, R. J., & Griffiths, H. (2012). Rubisco small-subunit -helices control pyrenoid formation in Chlamydomonas. *Proceedings of the National Academy of Sciences*, *109*(47), 19474–19479. http://doi.org/10.1073/pnas.1210993109

Meyer, M. T., & Griffiths, H. (2013). Origins and diversity of eukaryotic CO2-concentrating mechanisms: lessons for the future. *Journal of Experimental Botany*, *64*(3), 769–86. http://doi.org/10.1093/jxb/ers390

Meyer, M. T., McCormick, A. J., & Griffiths, H. (2016). Will an algal CO2-concentrating mechanism work in higher plants? *Current Opinion in Plant Biology*, *31*(Ccm), 181–188. http://doi.org/10.1016/j.pbi.2016.04.009

Mitra, N. J., Pauly, M., Wand, M., & Ceylan, D. (2013). Symmetry in 3D geometry: Extraction and applications. *Computer Graphics Forum*, *32*(6), 1–23. http://doi.org/10.1111/cgf.12010

Miziorko, H. M., & Lorimer, G. H. (1983). Ribulose-1,5-Bisphosphate Carboxylase-Oxygenase. *Annual Review of Biochemistry*, *52*(1), 507–535. http://doi.org/10.1146/annurev.bi.52.070183.002451

Murphy, K., Murphy, K., Weiss, Y., Weiss, Y., Jordan, M., & Jordan, M. (1999). Loopy-belief Propagation for Approximate Inference: An Empirical Study. *15*, 467–475.

Nickell, S., Förster, F., Linaroudis, A., Net, W. Del, Beck, F., Hegerl, R., … Plitzko, J. M. (2005). TOM software toolbox: acquisition and analysis for electron tomography. *Journal of Structural Biology*, *149*(3), 227–234. http://doi.org/10.1016/j.jsb.2004.10.006

Nickell, S., Kofler, C., Leis, A. P., & Baumeister, W. (2006). A visual approach to proteomics. *Nature Reviews Molecular Cell Biology*, *7*(3), 225–230. http://doi.org/10.1038/nrm1861

Ortiz, J. O., Förster, F., Kürner, J., Linaroudis, A. A., & Baumeister, W. (2006). Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *Journal of Structural Biology*, *156*(2), 334–341. http://doi.org/10.1016/j.jsb.2006.04.014

Palade, G. E. (1964). the Organization of Living Matter. *Proceedings of the National Academy of Sciences*, *52*(2), 613–634. http://doi.org/10.1073/pnas.52.2.613

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*.

San Francisco, CA: Morgan Kaufmann Publishers.

Penczek, P. (2010). Resolution Measures in Molecular Electron Microscopy. In *Methods in Enzymology* (1st ed., Vol. 482, pp. 73–100). Elsevier Inc. http://doi.org/10.1016/S0076-6879(10)82003-8

Penczek, P. a, Renka, R., & Schomberg, H. (2004). Gridding-based direct Fourier inversion of the three-dimensional ray transform. *Journal of the Optical Society of America A*, *21*(4), 499. http://doi.org/10.1364/JOSAA.21.000499

Penczek, P., Radermacher, M., & Frank, J. (1992). Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy*, *40*(1), 33–53. http://doi.org/10.1016/0304-3991(92)90233-A

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. http://doi.org/10.1002/jcc.20084

Pfeffer, S., Brandt, F., Hrabe, T., Lang, S., Eibauer, M., Zimmermann, R., & Förster, F. (2012). Structure and 3D Arrangement of Endoplasmic Reticulum Membrane-Associated Ribosomes. *Structure*, *20*(9), 1508–1518. http://doi.org/10.1016/j.str.2012.06.010

Pfeffer, S., Burbaum, L., Unverdorben, P., Pech, M., Chen, Y., Zimmermann, R., … Förster, F. (2015). Structure of the native Sec61 protein-conducting channel. *Nature Communications*, *6*, 8403. http://doi.org/10.1038/ncomms9403

Pfeffer, S., Dudek, J., Gogala, M., Schorr, S., Linxweiler, J., Lang, S., … Förster, F. (2014). Structure of the mammalian oligosaccharyl-transferase complex in the native ER protein translocon. *Nature Communications*, *5*, 3072. http://doi.org/10.1038/ncomms4072

Pfeffer, S., Dudek, J., Zimmermann, R., & Förster, F. (2016). Organization of the native ribosome–translocon complex at the mammalian endoplasmic reticulum membrane. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1860*(10), 2122–2129. http://doi.org/10.1016/j.bbagen.2016.06.024

Pfeffer, S., Woellhaf, M. W., Herrmann, J. M., & Förster, F. (2015). Organization of the mitochondrial translation machinery studied in situ by cryoelectron tomography. *Nature Communications*, *6*, 6019. http://doi.org/10.1038/ncomms7019

Pospísek, M., & Valásek, L. (2013). Polysome Profile Analysis – Yeast. In *Methods in enzymology* (1st ed., Vol. 530, pp. 173–181). Elsevier Inc. http://doi.org/10.1016/B978-0-12-420037-1.00009-9

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Akad. Wiss.*, *69*, 262–277. article.

Reimer, L. (1984). *Transmission Electron Microscopy: Physics of Image Formation and Microanalysis*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Reimer, L., & Kohl, H. (2008). *Transmission Electron Microscopy Physics of Image Formation*. *Springer series in optical sciences* (Vol. 51). New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-34758-5

Rich, A., Warner, J. R., & Goodman, H. M. (1963). The Structure and Function of Polyribosomes. *Cold Spring Harbor Symposia on Quantitative Biology*, *28*, 269–285. http://doi.org/10.1101/SQB.1963.028.01.043

Rigort, A., Villa, E., Bäuerlein, F. J. B., Engel, B. D., & Plitzko, J. M. (2012). Integrative Approaches for Cellular Cryo-electron Tomography. In *Methods in Cell Biology* (Vol. 111, pp. 259–281). Elsevier. http://doi.org/10.1016/B978-0-12-416026-2.00014-5

Roseman, A. M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy*, *94*(3–4), 225–236. http://doi.org/10.1016/S0304-3991(02)00333-9

Sandberg, K., Mastronarde, D. N., & Beylkin, G. (2003). A fast reconstruction algorithm for electron microscope tomography. *Journal of Structural Biology*, *144*(1–2), 61–72. http://doi.org/10.1016/j.jsb.2003.09.013

Saxton, W. O., & Baumeister, W. (1982). The correlation averaging of a regularly arranged bacterial cell envelope protein. *Journal of Microscopy*, *127*(2), 127–138. http://doi.org/10.1111/j.1365-2818.1982.tb00405.x

Schaffer, M., Engel, B. D., Laugks, T., Mahamid, J., Plitzko, J. M., & Baumeister, W. (2015). Cryo-focused Ion Beam Sample Preparation for Imaging Vitreous Cells by Cryo-electron Tomography. *Bio-Protocol*, *5*(17), 1–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27294174

Scheres, S. H. W., Melero, R., Valle, M., & Carazo, J.-M. (2009). Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization. *Structure*, *17*(12), 1563–1572. http://doi.org/10.1016/j.str.2009.10.009

Schmeing, T. M., & Ramakrishnan, V. (2009). What recent ribosome structures have revealed about the mechanism of translation. *Nature*, *461*(7268), 1234–1242. http://doi.org/10.1038/nature08403

Schuwirth, B. S. (2005). Structures of the Bacterial Ribosome at 3.5 A Resolution. *Science*, *310*(5749), 827–834. http://doi.org/10.1126/science.1117230

Schweikert, G. (2004). *Quantitativer vergleich der strahlschädigung biologischer proben im transmissions-elektronenmikroskop bei stickstoff und helium temperatur*. Technische Universität München.

Shimony, S. E. (1994). Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, *68*(2), 399–410. http://doi.org/10.1016/0004-3702(94)90072-8

Sivan, G., Kedersha, N., & Elroy-Stein, O. (2007). Ribosomal Slowdown Mediates Translational Arrest during Cellular Division. *Molecular and Cellular Biology*, *27*(19), 6639–6646. http://doi.org/10.1128/MCB.00798-07

Stölken, M., Beck, F., Haller, T., Hegerl, R., Gutsche, I., Carazo, J.-M., … Nickell, S. (2011). Maximum likelihood based classification of electron tomographic data. *Journal of Structural Biology*, *173*(1), 77–85. http://doi.org/10.1016/j.jsb.2010.08.005

Studer, D., Humbel, B. M., & Chiquet, M. (2008). Electron microscopy of high pressure frozen samples: Bridging the gap between cellular ultrastructure and atomic resolution. *Histochemistry and Cell Biology*, *130*(5), 877–889. http://doi.org/10.1007/s00418-008-0500-1

Takeshi, E., & Billinge, S. J. L. (2012). *Chapter 3 - The Method of Total Scattering and Atomic Pair Distribution Function Analysis*. *Underneath the Bragg Peaks Structural Analysis of Complex Materials* (Vol. 16). http://doi.org/http://dx.doi.org/10.1016/B978-0-08-097133-9.00003-4

Taylor, T. C., Backlund, A., Bjorhall, K., Spreitzer, R. J., & Andersson, I. (2001). First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. *The Journal of Biological Chemistry*, *276*(51), 48159–64. http://doi.org/10.1074/jbc.M107765200

Teter, S. A., Houry, W. A., Ang, D., Tradler, T., Rockabrand, D., Fischer, G., … Hartl, F. U. (1999). Polypeptide Flux through Bacterial Hsp70. *Cell*, *97*(6), 755–765. http://doi.org/10.1016/S0092-8674(00)80787-4

Tevs, A., Bokeloh, M., Wand, M., Schilling, A., & Seidel, H.-P. (2009). Isometric registration of ambiguous and partial data. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1185–1192). IEEE. http://doi.org/10.1109/CVPRW.2009.5206775

Tissières, A. (1974). Ribosome Research: Historical Background. In *Ribosomes* (pp. 3–12). Cold Spring Harbor, NY: Cold Spring Harbor Lab. http://doi.org/10.1101/087969110.4.3

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. http://doi.org/10.1007/s11222-007-9033-z

Voorhees, R. M., & Ramakrishnan, V. (2013). Structural Basis of the Translational Elongation Cycle*. *Annual Review of Biochemistry*, *82*(1), 203–236. http://doi.org/10.1146/annurev-biochem-113009-092313

Wan, X., Zhang, F., Chu, Q., Zhang, K., Sun, F., Yuan, B., & Liu, Z. (2011). Three-dimensional reconstruction using an adaptive simultaneous algebraic reconstruction technique in electron tomography. *Journal of Structural Biology*, *175*(3), 277–287. http://doi.org/10.1016/j.jsb.2011.06.002

Weiss, Y., & Freeman, W. T. (2001a). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, *47*(2), 736–744. http://doi.org/10.1109/18.910585

Weiss, Y., & Freeman, W. T. (2001b). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, *47*(2), 736–744. http://doi.org/10.1109/18.910585

Williams, D. B., & Carter, C. B. (2009). *Transmission Electron Microscopy: A textbook for materials science*. *Transmission Electron Microscopy*. Boston, MA: Springer US. http://doi.org/10.1007/978-0-387-76501-3

Zallen, R. (1998). *The Physics of Amorphous Solids*. (R. Zallen, Ed.)*The Physics of Amorphous Solids*. Weinheim, Germany: Wiley-VCH Verlag GmbH. http://doi.org/10.1002/9783527617968

Zanetti, G., Riches, J. D., Fuller, S. D., & Briggs, J. A. G. (2009). Contrast transfer function correction applied to cryo-electron tomography and sub-tomogram averaging. *Journal of Structural Biology*, *168*(2), 305–312. http://doi.org/10.1016/j.jsb.2009.08.002

Zhang, W., Kimmel, M., Spahn, C. M. T., & Penczek, P. A. (2008). Heterogeneity of Large Macromolecular Complexes Revealed by 3D Cryo-EM Variance Analysis. *Structure*, *16*(12), 1770–1776. http://doi.org/10.1016/j.str.2008.10.011

Acknowledgements

I would like to thank my supervisor Dr. Friedrich Förster, for introducing me to the exciting field of cryo-electron tomography and for supporting and guiding me during my PhD studies.

I would also like to thank my colleagues for a friendly working atmosphere and fruitful discussions, in particular to Dr. Yuxiang Chen, Dr. Stefan Pfeffer, Dr. Pia Unverdorben, Dr. Jan Schuller, Antje Aufderheide, Laura Burbaum and Robert Englmeier.

I would like to express my gratitude to Dr. Ben Engel and Dr. Miroslava Schaffer for supporting me and providing the data for the pyrenoid study. Special thanks to Dr. Julio Ortiz for providing the processed tomograms of *E. coli* lysate and to Dr. Stefan Pfeffer for providing and processing the tomograms of rough microsomes. I would also like to thank Dr. Friedrich Förster, Dr. Ben Engel and Dr. Stefan Pfeffer for reading and correcting my thesis. I would like to acknowledge Robert Englmeier for helping me translate the abstract of this thesis into german.

Many thanks to Dr. Thomas Hrabe, Dr. Yuxiang Chen and my supervisor Dr. Friedrich Förster for insightful discussions and for developing the PyTom software package.

Special thanks to Prof. Dr. Wolfgang Baumeister for allowing me to do my PhD studies in his department.