7-7-2011

Modeling Techniques for the High-Resolution Interpretation of Cryo-Electron Microscopy Reconstructions

Mirabela Rusu

Dissertation

MODELING TECHNIQUES FOR THE HIGH-RESOLUTION INTERPRETATION OF CRYO-ELECTRON MICROSCOPY RECONSTRUCTIONS

by

Mirabela Rusu, MEng, MS

July 7, 2011

APPROVED:

_____

Jack W. Smith, MD, PhD

_____

M. Sriram Iyengar, PhD

_____

Todd R. Johnson, PhD

_____

Robert W. Vogler, DSN, MEd

MODELING TECHNIQUES FOR THE HIGH-RESOLUTION INTERPRETATION OF CRYO-ELECTRON MICROSCOPY RECONSTRUCTIONS

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfillment
of the Requirements

for the Degree of

Doctor of Philosophy

by

Mirabela Rusu, MEng, MS

Committee Members:

Jack W. Smith*, MD, PhD
M. Sriram Iyengar*,†, PhD
Todd R. Johnson‡, PhD
Robert W. Vogler*, DSN, MEd

*School of Biomedical Informatics, University of Texas Health Science Center at Houston, US
†NASA Johnson Space Center, US
‡Division of Biomedical Informatics, University of Kentucky, US

To my family and friends

Acknowledgments

This thesis is the outcome of a long journey that taught me about structural biology, science in general and life in science. This work was made possible only through the help of my mentors, colleagues, family and friends. My appreciation goes to all of you.

First and foremost, I would like to thank my parents, for trusting in me and in my decisions since an early age, for encouraging me to go as far as my thoughts and dreams take me (even when it meant leaving their side), for teaching me strong values for life and that I can do anything that I set my mind to.

I would like to acknowledge Dr. Willy Wriggers for his courage in bringing me to Houston, his mentorship and support over all these years, his scientific and career advice. My recognition also goes to Dr. Stefan Birmanns for his guidance and mentorship. Many thanks to my committee members, Dr. Jack Smith, Dr. M. Sriram Iyengar, Dr. Todd Johnson and Dr. Robert Vogler for their advice, endless support and patience along this journey. Also, I'm greatful to other scientists who guided me over the years: Dr. Teresa Ruiz, Dr. Michael Radermacher, Dr. Alexander Freiberg, Dr. Michael Sherman, Dr. Elmer Bernstam and Dr. Jorge Herskovic. My gratitude goes to my fellow scientists, Manuel Wahle, Dr. Zbigniew Starosolski and Dr. Jochen Heyd, as well as my fellow students at the School of Biomedical Informatics.

Last, but not least, I thank my friends for becoming my Houston family, for advising me and patiently supporting me at the toughest, as well as, joyful moments during this journey. Thank you Giovanni, Gustavo, Lucia, Peyman, Paulina, Tina, Robert and all the others.

Thank you!

Contents

Introduction

Essential biological processes are governed by organized, dynamic interactions between multiple biomolecular systems. Complexes are thus formed to enable the biological function and get dissembled as the process is completed. Examples of such processes include the translation of the messenger RNA into protein by the ribosome, the folding of proteins by chaperonins or the entry of viruses in host cells. Understanding these fundamental processes by characterizing the molecular mechanisms that enable then, would allow the (better) design of therapies and drugs. Such molecular mechanisms may be revealed trough the structural elucidation of the biomolecular assemblies at the core of these processes. Various experimental techniques may be applied to investigate the molecular architecture of biomolecular assemblies. High-resolution techniques, such as X-ray crystallography, may solve the atomic structure of the system, but are typically constrained to biomolecules of reduced flexibility and dimensions. In particular, X-ray crystallography requires the sample to form a three dimensional (3D) crystal lattice which is technically difficult, if not impossible, to obtain, especially for large, dynamic systems. Often these techniques solve the structure of the different constituent components within the assembly, but encounter difficulties when investigating the entire system. On the other hand, imaging techniques, such as cryo-electron microscopy (cryo-EM), are able to depict large systems in near-native environment, without requiring the formation of crystals. The structures solved by cryo-EM cover a wide range of resolutions, from very low level of detail where only the overall shape of the system is visible, to high-resolution that approach, but not yet reach, atomic level of detail.

In this dissertation, several modeling methods are introduced to either integrate cryo-EM datasets with structural data from X-ray crystallography, or to directly interpret the cryo-EM reconstruction. Such computational techniques were developed with the goal of creating an atomic model for the cryo-EM data. The low-resolution reconstructions lack the level of detail to permit a direct atomic interpretation, i.e. one can not reliably locate the atoms or amino-acid residues within the structure obtained by cryo-EM. Thereby one needs to consider additional information, for example, structural data from other sources such as X-ray crystallography, in order to enable such a high-resolution interpretation. Modeling techniques are thus developed to integrate the structural data from the different biophysical sources, examples including the work described in the manuscript I and II of this dissertation. At intermediate and high-resolution, cryo-EM reconstructions depict consistent 3D folds such as tubular features which in general correspond to alpha-helices. Such features can be annotated and later on used to build the atomic model of the system, see manuscript III as alternative.

Three manuscripts are presented as part of the PhD dissertation, each introducing a computational technique that facilitates the interpretation of cryo-EM reconstructions. The first manuscript is an application paper that describes a heuristics to generate the atomic model for the protein envelope of the Rift Valley fever virus. The second manuscript introduces the evolutionary tabu search strategies to enable the integration of multiple component atomic structures with the cryo-EM map of their assembly. Finally, the third manuscript develops further the latter technique and apply it to annotate consistent 3D patterns in intermediate-resolution cryo-EM reconstructions.

The first manuscript, titled "*An assembly model for Rift Valley fever virus*", was sub-

mitted for publication in the Journal of Molecular Biology. The cryo-EM structure of the Rift Valley fever virus was previously solved at 27Å-resolution by Dr. Freiberg and collaborators. Such reconstruction shows the overall shape of the virus envelope, yet the reduced level of detail prevents the direct atomic interpretation. High-resolution structures are not yet available for the entire virus nor for the two different component glycoproteins that form its envelope. However, homology models may be generated for these glycoproteins based on similar structures that are available at atomic resolutions. The manuscript presents the steps required to identify an atomic model of the entire virus envelope, based on the low-resolution cryo-EM map of the envelope and the homology models of the two glycoproteins. Starting with the results of the exhaustive search to place the two glycoproteins, the model is built iterative by running multiple multi-body refinements to hierarchically generate models for the different regions of the envelope. The generated atomic model is supported by prior knowledge regarding virus biology and contains valuable information about the molecular architecture of the system. It provides the basis for further investigations seeking to reveal different processes in which the virus is involved such as assembly or fusion.

The second manuscript was recently published in the of Journal of Structural Biology (doi:10.1016/j.jsb.2009.12.028) under the title "*Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions*". This manuscript introduces the evolutionary tabu search strategies applied to enable a multi-body registration. This technique is a hybrid approach that combines a genetic algorithm with a tabu search strategy to promote the proper exploration of the high-dimensional search space. Similar to the Rift Valley fever virus, it is common that the structure of a large multi-component assembly is available at low-resolution from cryo-EM, while high-resolution

structures are solved for the different components but lack for the entire system. Evolutionary tabu search strategies enable the building of an atomic model for the entire system by considering simultaneously the different components. Such registration indirectly introduces spatial constrains as all components need to be placed within the assembly, enabling the proper docked in the low-resolution map of the entire assembly. Along with the method description, the manuscript covers the validation, presenting the benefit of the technique in both synthetic and experimental test cases. Such approach successfully docked multiple components up to resolutions of 40Å.

The third manuscript is entitled "*Evolutionary Bidirectional Expansion for the Annotation of Alpha Helices in Electron Cryo-Microscopy Reconstructions*" and was submitted for publication in the Journal of Structural Biology. The modeling approach described in this manuscript applies the evolutionary tabu search strategies in combination with the bidirectional expansion to annotate secondary structure elements in intermediate resolution cryo-EM reconstructions. In particular, secondary structure elements such as alpha helices show consistent patterns in cryo-EM data, and are visible as rod-like patterns of high density. The evolutionary tabu search strategy is applied to identify the placement of the different alpha helices, while the bidirectional expansion characterizes their length and curvature. The manuscript presents the validation of the approach at resolutions ranging between 6 and 14Å, a level of detail where alpha helices are visible. Up to resolution of 12 Å, the method measures sensitivities between 70-100% as estimated in experimental test cases, i.e. 70-100% of the alpha-helices were correctly predicted in an automatic manner in the experimental data.

The three manuscripts presented in this PhD dissertation cover different computation

methods for the integration and interpretation of cryo-EM reconstructions. The methods were developed in the molecular modeling software Sculptor (http://sculptor.biomachina.org) and are available for the scientific community interested in the multi-resolution modeling of cryo-EM data. The work spans a wide range of resolution covering multi-body refinement and registration at low-resolution along with annotation of consistent patterns at high-resolution. Such methods are essential for the modeling of cryo-EM data, and may be applied in other fields where similar spatial problems are encountered, such as medical imaging.

Mirabela Rusu, M.Eng., M.S.

July 2011

Manuscripts

An assembly model for Rift Valley fever virus *

*Manuscript submitted for publication in the Journal of Molecular Biology

Mirabela Rusu[†]

Richard Bonneau[‡]

Michael R. Holbrook[§,¶]

Stanley Watowich[‖]

Stefan Birmanns[†]

Willy Wriggers[**]

Alexander N. Freiberg[§,††]

[†]School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Tx 77030, USA

[‡]Computational Biology Program, Department of Biology, Center for Genomics and Systems Biology, Computer Science Department, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA

[§]Department of Pathology, Institute for Human Infection and Immunity, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555, USA

[¶]NIAID-Integrated Research Facility, 8200 Research Plaza, Ft. Detrick, Frederick, Maryland 21702, USA

[‖]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555, USA

[**]Department of Physiology and Biophysics and Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1300 York Ave., New York, NY 10065. Permanent address: D. E. Shaw Research, 120 West 45th Street, New York, NY 10036

[††]To whom correspondence should be addressed: anfreibe@utmb.edu

An assembly model for Rift Valley fever virus

Abstract

Rift Valley fever virus (RVFV) is a bunyavirus endemic to Africa and the Arabian Peninsula that infects humans and livestock. The virus encodes two glycoproteins, Gn and Gc, which represent the major structural antigens and are responsible for host cell receptor binding and fusion. Both glycoproteins are organized on the virus surface as cylindrical hollow spikes that cluster into distinct capsomers with the overall assembly exhibiting an icosahedral symmetry. Currently, no experimental three-dimensional structure for any entire bunyavirus glycoprotein is available. Using fold recognition, we generated molecular models for both RVFV glycoproteins and found significant structural matches between the RVFV Gn protein and the Influenza hemagglutinin protein and a separate match between Sindbis virus envelope protein E1 and the RVFV Gc protein. Using these models, the potential interaction and arrangement of both glycoproteins in the RVFV particle was analyzed, by modeling the placement within the cryo-electron microscopy density map of RVFV. Our assembly model for RVFV proposes that the ectodomain of Gn forms the majority of the protruding capsomer spikes and that Gc is mainly involved in formation of the capsomer base. Further, Gc could be identified as the glycoprotein forming the intercapsomer connection. We suggest that the highly organized arrangement of the glycoproteins play a major role for virion stability. The overall proposed arrangement of the two glycoproteins presents similarities to the one described for the alphavirus E1-E2 proteins. We believe that our model will provide guidance to better understand the assembly process of bunyaviruses at a molecular level.

Keywords: Bunyavirus, Protein structure prediction, cryo-electron microscopy, virus assembly model, hybrid modeling, multi-body refinement, multi-resolution registration

Introduction

Rift Valley fever virus (RVFV) is a member of the Bunyaviridae family (genus Phlebovirus), transmitted mainly by mosquitoes and is endemic throughout much of Africa with recent outbreaks in parts of the Arabian Peninsula. The virus causes disastrous outbreaks in a wide range of vertebrate hosts, with man and livestock being the most sensitive. Infection of livestock can result in economically disastrous abortion storms and high mortality among young animals. The virus causes a variety of pathologies in humans with approximately 1% of infections resulting in a fatal hemorrhagic fever or encephalitis [41]. However, during the outbreak in Kenya, November 2006 to January 2007, the case fatality rate in humans reached nearly 30% [30]. If RVFV were introduced into the Americas, the virus could become established as competent vector species are endemic and livestock populations are naïve. An outbreak of RVFV in the U.S. would severely disrupt the beef and dairy industry, and could potentially cause significant human illness and greater mortality than the West Nile virus. With these characteristics, RVFV is considered a high consequence emerging infectious disease and therefore, the intentional spread of RVFV is a concern of national biosecurity. RVFV is classified as Category A select agent by CDC and USDA. Currently, there are no commercially available vaccines or therapeutics.

As all bunyaviruses, RVFV is an enveloped virus and has a tri-segmented, negative-sense RNA genome that replicates in the cytosol. Bunyaviruses most likely enter their host cells

via receptor-mediated endocytosis and require an acid-activated membrane fusion step [29]. The two glycoproteins, Gn and Gc, are expressed as a precursor polypeptide, which is then co-translationally cleaved prior to maturation of the envelope glycoproteins [7, 49]. For the transport from the endoplasmic reticulum to the Golgi apparatus, both newly synthesized glycoproteins are required [12]. Within the virion, the surface glycoproteins represent two of the four major structural proteins associated with the virus particle. They are anchored in the envelope membrane as type-I integral membrane proteins and are responsible for receptor recognition and binding, and entry into the target cells through fusion between viral and cellular membranes. In contrast to most other negative-stranded RNA viruses, bunyaviruses lack the presence of a matrix protein and the cytoplasmic tails of Gn and Gc are likely to interact with the ribonucleoprotein complex inside the virus particle [32, 34]. Gn and Gc form oligomers and are organized on the virus surface as cylindrical hollow spikes that cluster into distinct capsomers. These capsomers cover the virus surface, arranged on an icosahedral lattice with a triangulation number of 12 [10, 42, 19]. Computational studies have predicted RVFV Gc to be a class II viral fusion protein [11]. Owing to their importance in the process of maturation, receptor binding and fusion with the host cell, both glycoproteins form attractive targets for the design of antiviral drugs blocking the receptor binding and/or fusion process.

At present, no crystallographic data are available for any entire bunyavirus glycoprotein. Molecular modeling has been regularly used at various levels of representation to gain insight into the 3D structure and biochemistry of biomolecules and help in elucidating mechanisms at the molecular level. Bioinformatics investigation of the bunyavirus Gc proteins from the five different genera revealed that they share a limited number of similar sequences

with each other and that they have sequence similarity with the alphavirus E1 protein, suggesting that bunyavirus Gc proteins are class II viral fusion proteins [11]. Further, experiments with other bunyaviruses supported the major role Gc plays during cell fusion [35, 36, 43]. Three-dimensional molecular model structures for Gc have been described for members from different genera, such as La Crosse virus (Orthobunyavirus), Sandfly fever virus (Phlebovirus), Andes and Tula viruses (Hantaviruses), and has been used successfully to study the functionality of fusion peptides or the interaction and oligomerization of glycoproteins [11, 46, 44, 15]. However, while most of these studies targeted the Gc protein, less information is available for the bunyavirus Gn protein. It has been suggested that Gn plays a role in receptor binding and that it might have structural similarity to the alphavirus E2 protein [11].

To better understand the assembly of bunyaviruses and the interaction between Gn and Gc glycoproteins, we sought to generate three-dimensional (3D) structure models for RVFV Gn and Gc monomers using bioinformatics approaches. Subsequently, we used these structures to evaluate possible positions within the existing cryo-electron microscopy (cryoEM) density map of RVFV to predict protein-protein interaction interfaces and to propose a model for the assembly of RVFV at a molecular level. We suggest that RVFV Gn and Gc follow a topological arrangement within the virus particle similar to the E1 and E2 proteins of alphaviruses. By structural analogy to these alphaviruses our model suggests that the Gn protein is most likely responsible for receptor binding and protects the fusion loop of Gc at neutral pH. Further, Gn and Gc form a highly organized arrangement which might contribute to the overall virion stability.

RVFV G_N/G_C [SwissProt #P21401]

Program	G_N [536 aa]	G_C[507 aa]
EXPASY	429 – 449 [21aa]	470 – 490 [21aa]
HMMTOP	429 – 451 [23aa] 464 – 486 [CTD] 517 – 535 [19aa][a]	470 – 494 [25aa]
SOSUI	433 – 455 [23aa] 464 – 479 [CTD] 515 – 536 [22aa][a]	469 – 491 [23aa]
TMHMM	432 – 454 [23aa]	469 – 491 [23aa]
Average	429 – 455 [27aa] 515 – 536 [22aa][a]	469 – 494 [26aa]
CTD	456 – 514 [59aa]	495 – 507 [13aa]

[a]2nd TMD in G_N corresponds to signal peptide

Table 1: Prediction of location of transmembrane domains (TMD)

Materials and Methods

Sequences

For sequence and structural analysis, the RVFV vaccine strain MP-12 glycoprotein encoding nucleotide sequence (Pubmed DQ380208) has been used. The secondary structure of RVFV Gn and Gc, respectively, were examined using Jpred3 [6]. To define the location of the glycoprotein transmembrane domains, as well as cytoplasmic tail domains, the EXPASY, HMMTOP, SOSUI, and TMHMM servers were used [47, 17, 25].

We used the NetN Glyc 1.0 Server [14] to predict the locations of N-glycosylation sites. Glycosylated Asn residues are indicated according to their location within the M-segment encoded polyprotein (Figure 1).

Protein structure prediction

Fold recognition was used to detect alignments between solved structures and the Gc and Gn proteins. Initial backbone models were generated using the fold recognition Meta Server [22], which used alignments from the FFAS_03 program to the two templates [21]. These models agreed with alignments found using other fold recognition methods, increasing our confidence in these fold predictions. Side chains were added and models were refined using Modeler [9]. The atomic model of the Gn glycoprotein was generated based on the 1918 influenza H1 hemagglutinin protein (PDB ID: 1RD8, [45]) for which a 14.15% sequence identity was observed. Similarly, the atomic model of the Gc glycoprotein was build based on the Sindbis virus structural E1 protein (PDB ID: 1LD4, [53]) from an observed sequence identity of 13.83%. In addition to these structures sub-optimal FFAS_03 alignments and derived models were also evaluated in the context of the cryoEM density including alignments of RVFV Gc to PDB structures of the Semliki Forest virus (PDB: 2XFC, [48]; PDB ID: 1RER, [13]), dengue virus E (PDB ID: 1P58, [52]), integrin binding fragment of human fibrillin-1 (PDB ID: 1UZJ, [26]) and alignment of Gn to the EAP45/ESCRT GLUE domain (PDB ID: 2HTH-chain A, [1]).

Due to the similar sequence identity between the different homologues, an alternative atomic model of the RVFV Gc protein was built based on the structure of the Chikungunya virus E1 protein fitted into the cryoEM reconstruction of Semliki Forest virus(PDB ID: 2XFC; chain A, [48]). The overall shape of Gc is conserved between the initial model based on the Sindbis virus E1 protein and the Semliki Forest virus-based model, however domain I and the connection between domain II and III are slightly better structured (data not

shown).

Fitting of glycoprotein structures into the cryoEM density

The atomic models of the RVFV Gn and Gc glycoproteins were fitted into the RVFV vaccine strain MP-12 cryoEM map, solved at 27Å-resolution [42]. The organization of the two glycoproteins within the RVFV envelope was identified following a hybrid approach that combined an interactive exploration of the exhaustive search outcome with a multi-body refinement procedure. The multi-body refinement is described in detail in [5] and a summary is provided here. The multi-step approach (Figure 2) is applied to generate an atomic model for the triangular face of the RVFV. First, an exhaustive search using the tool colores from the package *Situs* [51] is applied to explore possible placements for each of the Gc and Gn glycoproteins. The molecular modeling software *Sculptor* [5] was then utilized for the interactive exploration of the exhaustive search results to select placements that are in agreement with computed Gn/Gc ratio within each capsomer type [42] and that show reduced steric clashing. Several such docking locations were identified for both Gc and Gn, and multiple models were iteratively refined by searching for the architecture that best described the density of the asymmetric unit. A detailed description of the modeling steps is given in the next section.

Results

Fold recognition of the RVFV glycoproteins

Both RVFV glycoproteins, Gn and Gc, are known to be type I integral transmembrane proteins. Before obtaining fold recognition and molecular model predictions of the two RVFV glycoproteins, the primary amino acid sequences of the entire Gn and Gc were analyzed for the predictions of transmembrane domains (TMD), ecto- and endodomains (cytoplasmic tail domains – CTD), glycosylation sites and consensus secondary structure. Gn displays a mixture of α-helical, β-strands and random coil secondary structural elements (Figure 1B). The N-terminus has a slightly higher content of β-strands, while the C-terminus is rich in α-helical elements located in the regions predicted for the transmembrane and cytoplasmic tail domains. RVFV Gc exhibits predominantly β-strands, a very low content of α-helices and a high content of random coil (Figure 1C). Most of the α-helical elements are found in the regions predicted for the transmembrane and short cytoplasmic tail domains, as already described for the Gn protein. Garry and Garry [11] suggested that the Gc glycoproteins of bunyaviruses are class II viral fusion proteins. Class II fusion proteins, such as the envelope glycoprotein E of tick-borne encephalitis virus or the E1 protein of Sindbis virus, are comprised mostly of antiparallel β-sheets, similar to the secondary structure prediction for RVFV Gc.

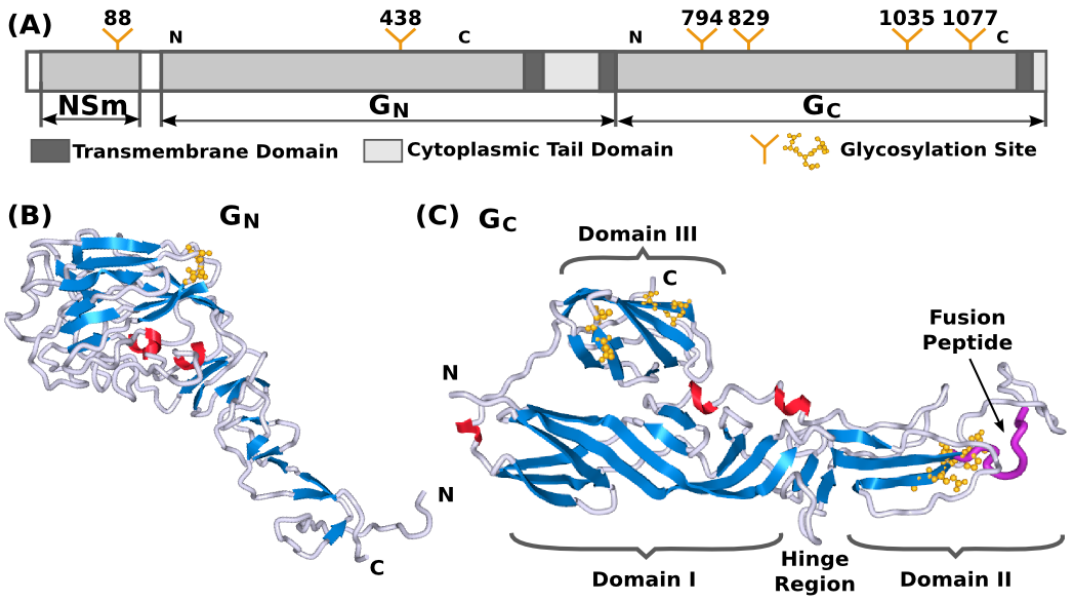

Figure 1: 3D structure models of RVFV Gn and Gc proteins. (A) Schematic representation of the RVFV M-segment polyprotein. Transmembrane and cytoplasmic tail domains are highlighted in dark grey or white bars, respectively. N-Glycosylation sites are indicated with the position of the respective Asn residue. The regions of the two glycoproteins used for molecular modeling are indicated with N and C. 3D molecular models for RVFV (B) Gn and (C) Gc are shown. Secondary structures are highlighted in blue for β-strands, red for α-helices, and grey for turns. The predicted location of the fusion peptide within Gc is represented in purple. The domain nomenclature in modeled Gc were used in adoption to the alphavirus E1 protein.The molecular graphics in this paper were generated with Sculptor [5] and Chimera [33]

Model building and structural description of the RVFV Gn and Gc glycoproteins

To determine the 3D structures of Gn and Gc, and to verify that RVFV Gc adopts a class II fusion protein fold, we applied molecular modeling protein structure prediction as described in the Material and Methods section. Our predictions initially focused on the nearly entire RVFV Gn (530 aa in length) and Gc (507 aa in length) protein sequences (Figure 1A). However, molecular models could only be generated for the two glycoprotein ectodomains, and the TMDs and CTDs have been removed from our further analysis. In the following

description, RVFV Gn and Gc are used to describe the ectodomain for each glycoprotein.

The molecular modeling results revealed that the best matching profile for RVFV Gn resulted in a hit which had similarity to the Influenza 1918 human H1 hemagglutinin (Figure 1B).. In contrast, the molecular model generated for RVFV Gc was obtained with the Sindbis virus and Chikungunya E1 protein, respectively. As shown in Figure 1B, the modeled structure for RVFV Gc resembles the overall fold of a class II fusion protein. All of the listed additional hits found, represent class II fusion proteins. This result was not surprising, since Garry and Garry [11] already predicted the bunyaviruses Gc protein to be class II fusion proteins.

The validity of the models was evaluated in terms of stereochemical and geometric parameters such as bond lengths, bond angles, torsion angles, and packing environment and was found to satisfy all stereochemical criteria (assessed by VADAR statistics software package [50]). For the 3D models, the (Φ, Ψ) values calculated for each amino acid residue of the individual model structures are within the allowed region of the Ramachandran plot [38] (data not shown). For RVFV Gn, 98% of the residues were assigned to the Ramachandran's plot most-favored and allowed regions, 1% to generously allowed regions and 1% to disallowed regions. For RVFV Gc the assignments to the three regions are 97%, 2% and 1%.

The Gc protein is highly extended and consists of three domains, with predominantly β-strand content, which is in accordance with the amino acid sequence analysis (data not shown). The nomenclatures of these three domains have been defined in accordance to the alphavirus E1 protein, and are domain I (central domain), domain II and domain III. Domain II contains two predicted glycosylation sites at positions N794 and N829 and also bears the predicted fusion loop of RVFV Gc that potentially inserts into the target host

membrane during the pH-dependent virus fusion step [39, 11]. The location of the fusion loop is highlighted in purple in Figure 1C. Domain III is separated from the first two domains by a short stretch, forms an Ig-like β-barrel structure and contains two glycosylation sites at position N1035 and N1077.

In contrast, the predicted 3D model for the ectodomain of RVFV Gn represents an elongated structure with a globular head-domain (Figure 1B). The membrane-distal domain consists of a globular head, which displays a mixture of β-strands, slightly lower α-helical and random coil content. A stem-like region connects the globular domain with the TMD, which are not displayed in the 3D structure. The head-domain also contains the glycosylation site at position N285.

The predicted N-glycosylation sites were in agreement with the findings from Kakach et al. [23]. The N-glycosylation sites are indicated as yellow spheres in the 3D models in Figures 1B and C. All glycosylation sites on Gn and Gc are fully surface accessible, which adds a validation to our model structures.

Glycoprotein modeling in the RVFV particle

Recently, we determined the 3D structure of the RVFV vaccine strain MP-12 by single-particle cryoEM at 27Å resolution [42]. The reconstruction shows the T=12 icosahedral envelope of the virions, depicting different types of capsomers (hollow cylinders resulting from a tight association of shell glycoproteins)[10, 42]. Having the two model structures available for the RVFV glycoproteins, we sought to identify their organization within the different capsomers identified in the cryoEM density map by means of cross-correlation and

to build a model for the whole glycoprotein layer of the phlebovirus particle.

The glycoprotein layer is composed of capsomers showing different symmetry order [10, 42]. Pentons are located around the 5-fold symmetry axis while hexons organize around the 3-fold, quasi 3-fold and 2-fold axis. Although an icosahedral symmetry is imposed when reconstructing the cryoEM map of the virus, the hexons show different symmetry orders and can be averaged to increase the level of detail of the volumetric data. Such practice is common in low resolution structures, where averaging is applied to increase the signal-to-noise ratio of the data. First, the three different types of hexons were extracted, aligned and then an averaged volume from the 11 copies was computed (rotations included). This averaged hexon displays a 6-fold symmetry and was then used to construct an average density for the asymmetric unit and the corresponding triangular face. The cryoEM density of the averaged face was utilized as target volume for the global docking of the Gc and Gn glycoproteins, respectively, inside the envelope. An exploration of all possible translations and rotations (9^{o} step size) was performed for each glycoprotein with the colores tool of the *Situs* package [51]. Such an exhaustive search estimates the optimal cross-correlation coefficient, providing the list of top scoring placements. Upon request, colores also provided the optimal score and corresponding rotation for each voxel in the cryoEM map. This 3D scoring landscape may be further investigated using interactive exploration techniques, as described below.

Due to the reduced resolution of the cryoEM map, the top scoring placements provided by the exhaustive search were identified in the high-density regions of the map. Such arrangement of glycoproteins generate an atomic model with major steric clashes and prevent the assembly of the capsomer according to the Gc/Gn ratios estimated by Sherman et al [42]. Thereby, we further investigated the results of the exhaustive search using the interactive

exploration techniques [16] provided by the molecular modeling software *Sculptor* [5]. This approach permits the selection of local optimal solutions that are in agreement with available pieces of information, such as the Gc/Gn ratio inside the capsomers. Multiple docking locations were thus selected for each type of glycoprotein resulting in several Gc/Gn pairs considered for the further modeling steps.

Each such pair underwent the procedure described in Figure 2. First, the interactively selected placements are employed to create an initial model of the hexon located on the 3-fold axis by applying a 6-fold symmetry. This atomic model, composed of 6 Gc and 6 Gn units, was also placed in the neighboring capsomers. We proceeded with a multi-body Powell refinement of these fragments in the raw volume (as described in [5]) while applying boundary constraints. Such a local optimization simultaneously refines the translation and rotation of each glycoprotein in the capsomer by maximizing the cross-correlation coefficient. As multiple fragments are considered at the same time, the refinement prevents the glycoproteins from overlapping or from causing major steric clashes. The technique permits the introduction of boundary constraints in the form of atomic models describing the neighboring capsomers. Such constraints are not well defined in the first steps of the modeling and thereby the individual glycoproteins generating the neighboring capsomers are also considered in the multi-body refinement. As the different types of capsomers are identified, the neighboring capsomers become available and are utilized as constraints in the refinement. No symmetry is considered during the refinement, yet the units will eventually fall into the symmetry exhibited by the capsomer volume. For example, a 3-fold symmetry becomes apparent when refining the B capsomer which is organized around the 3-fold symmetry axis. The multi-body refinement was iterated several times until the placement of the

glycoproteins was stable. As an atomic model is generated for each type of capsomer, a final multi-body refinement is undertaken to create the asymmetric unit. Forty-six units, 23 Gc and 23 Gn glycoproteins, were simultaneously refined while constraining the 15 neighboring capsomers.

Intra- and intercapsomer placement of RVFV Gn and Gc

We applied the described procedure (Figure 2) to 11 Gc/Gn pairs obtained by combining the interactively selected Gc and Gn glycoproteins. Some of these pairs were discarded during the modeling as it become apparent that they will prevent the generation of models with good stereochemical quality or appropriate Gc/Gn ratios. At the end of the procedure, four models were produced with cross-correlation coefficients above 0.740. The top-scoring model had a correlation of 0.766 and is shown in Figures 3 and 4. This model had an estimated volume of approx. 1,300,000 Å3 for the hexon and approx. 1,100,000 Å3 for the pentons, in agreement with our previous calculations [42].

Although the resolution of the 3D map of RVFV was insufficient to visualize any boundaries between individual glycoprotein monomers, a plausible model was derived through docking of the molecular Gn and Gc models. Both glycoprotein models fit very well with the cryoEM density map (Figure 3A). This model clearly defines the location of Gn forming the protruding spikes and Gc is lying at an approximately 45° angle on the viral surface, forming an icosahedral scaffold. Gc can be ascribed to the density identified as the viral "skirt" around the base of each capsomer. The predicted arrangement of the two glycoproteins leads to both, homo- and heterodimeric contacts between Gn and Gc.

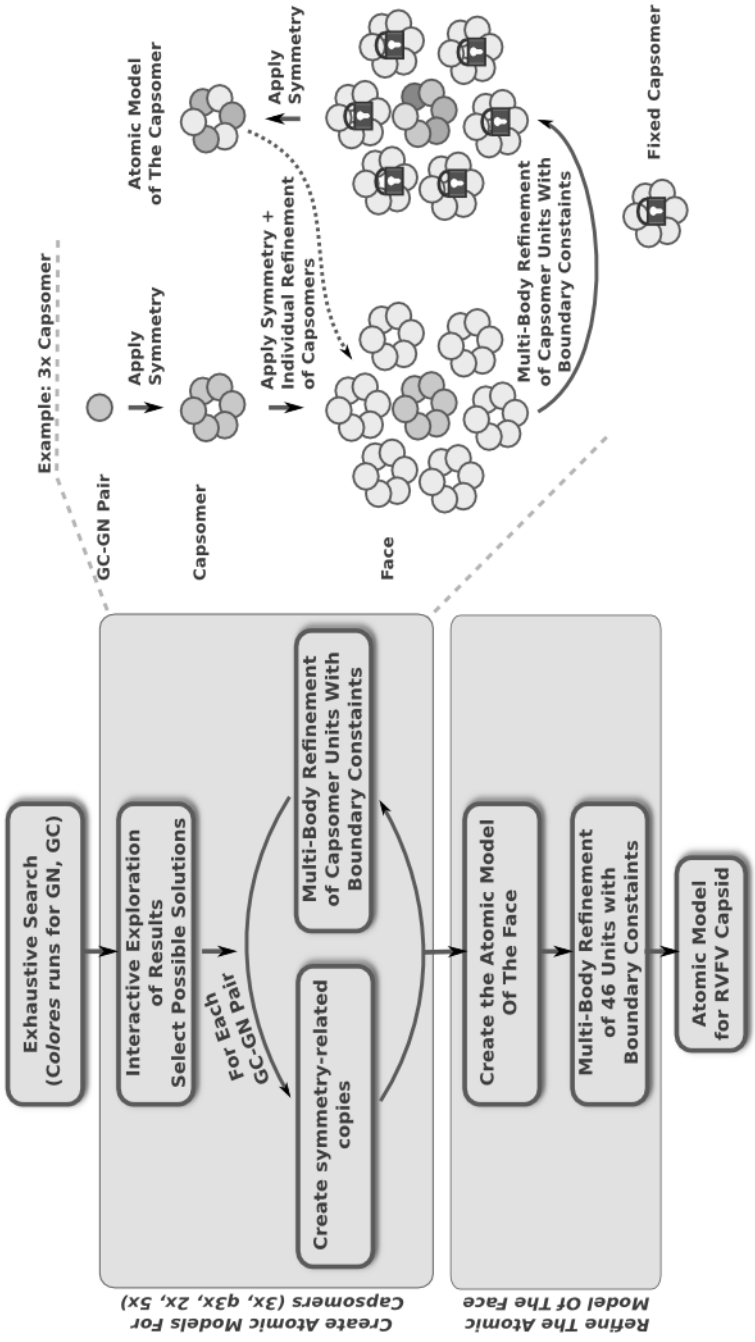

Figure 2: Schematic representation of the modeling steps undertaken to create the atomic model of the RVFV envelope. A detailed description of the individual steps is found in the text.

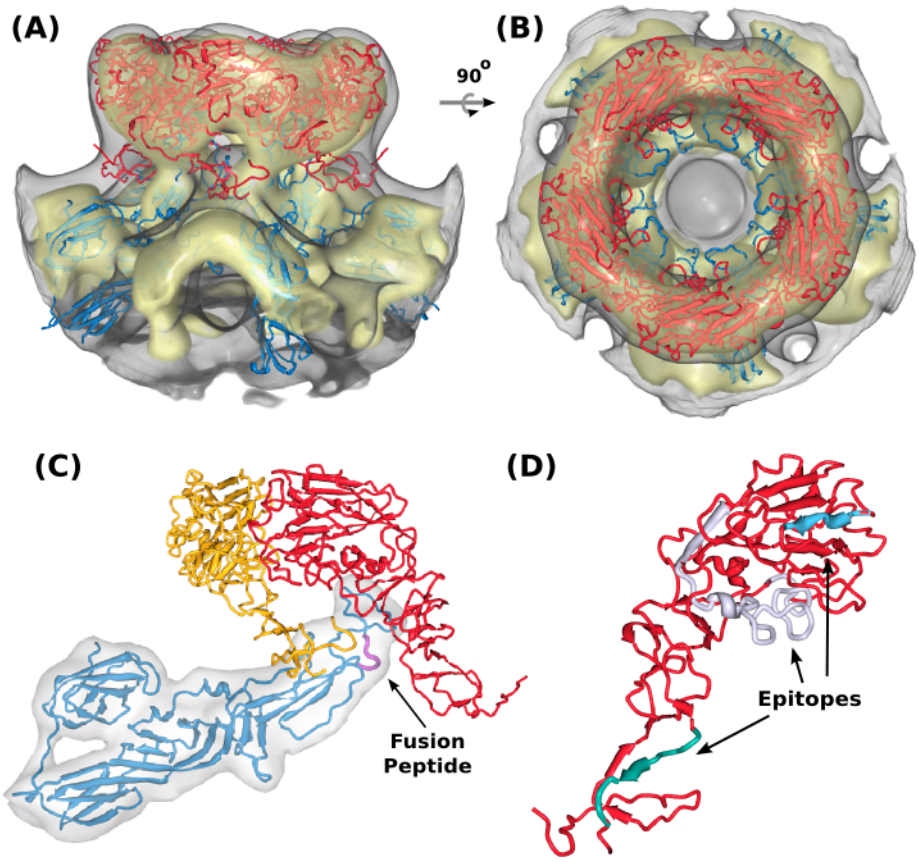

Figure 3: Positioning of the Gn and Gc molecular models into the RVFV cryoEM reconstruction. (A) and (B) show the glycoprotein arrangement within a penton extracted from the cryoEM density. The cryoEM density is represented as a grey transparent capsomer and the glycoprotein monomer models are indicated in red (Gn) and blue (Gc). Gn could only be positioned in the outer caldera of the capsomer and Gc in the skirt region of the capsomer. Two different viewing angles are shown (side-view, and top-view). (C) One structural unit (Gn-Gc heterodimer) and an adjacent Gn monomer have been extracted from the docking results shown in (A). Within the basic structural unit, the head domain of the Gn model (red and yellow) covers domain II of Gc. The predicted location of the fusion peptide shown in domain II of Gc is highlighted in magenta and indicated by the black arrow. (D) Epitopes for three monoclonal antibodies recognizing Gn [24] are highlighted. These epitopes are corresponding to the monoclonal antibodies 4-32-8D (grey), 4-D4 (blue) and 3C-10 (green).

Based on the limited similarities of the phlebovirus Gn and Gc proteins with the alphavirus E2 and E1 proteins, respectively, Gn has been depicted as the main receptor-binding protein and Gc as the main player during fusion with the target host membrane [11]. The location of the two glycoproteins suggested in our model is plausible. Gn fits into the outer density of the capsomers and the model is consistent with the available biological data on RVFV. Keegan & Collett [24] localized four distinct antigenic determinants along the Gn glycoprotein. We chose three of these mapped epitopes and highlighted them in our molecular model for Gn (Figure 3D). Two of these epitopes which are actually recognized by neutralizing monoclonal antibodies are surface exposed (highlighted in blue and grey in Figure 3D) and the epitope recognized by a non-neutralizing and non-protective antibody is located within the stem region of Gn and located in a region which interacts with Gc (highlighted in green in Figure 3D). In our model Gn interacts with domain II of Gc and also covers the fusion loop (highlighted in Figure 3C). The placement of Gc within the RVFV particle, reminds of the alphavirus E2 arrangement [40]. Domain II of E2 is the main interacting domain with E1, E2 has a position within the spike with a slight upward orientation on the virus surface and also forms the skirt of the spike [48, 28].

In the cryoEM reconstructions of RVFV, a strong density bridging neighboring capsomers has been described [10, 42, 19]. These ridges are located halfway between the rim of the capsomer and the lipid bilayer of the virion. Inside these ridges a channel of approximately 18 Å in diameter runs between adjacent capsomers and interconnects the inner cavities of the neighboring capsomers. According to our model of the glycoprotein arrangement, Gc could be placed into the density of these ridges (Figure 4A). Specifically, the domain III of two Gc molecules from adjacent capsomers filled the density (highlighted in red in Figure

4A and B). In the side view, one can clearly see, how the domain III form the tunnel-like structure (Figure 4B). Further, the position of the fusion peptide oriented to the capsomer center is displayed (arrow in Figure 4B).

A similar model for the envelope was also obtained when building the RVFV Gc glycoprotein based on the structure of the Chikungunya virus E1 protein (data not shown) [48]. Again, Gn forms the protrusion spikes of the capsomers, while Gc is the main component of the icosahedral scaffold. Similarly, the domain III of the Gc is the main component of the ridges between the capsomers, yet in this model the stem-like region of Gn is partially involved into the formation of the ridges as well (data not shown). Unlike the previous model, the fusion peptide located within Gc, points more outward from the capsomer but is still covered by the Gn glycoprotein.

Discussion

The family Bunyaviridae is organized into 5 distinct genera based on genetic and antigenic differences and represents the largest RNA virus family with more than 350 named isolates [8]. While many studies have focused on the characterization of the molecular biology aspects of transcription and replication, pathogenesis and vaccine development, little is known regarding the structural organization and physical interactions of bunyavirus glycoproteins within the virion. Recently, cryoEM structures have been solved for the phleboviruses RVFV [10, 42, 19] and Uukuniemi virus [31] and the hantaviruses Tula virus [18] and Hantaan [2]. These structures did not only increase our basic knowledge regarding the assembly of this important virus family but also revealed that the bunyavirus glycoproteins can result in

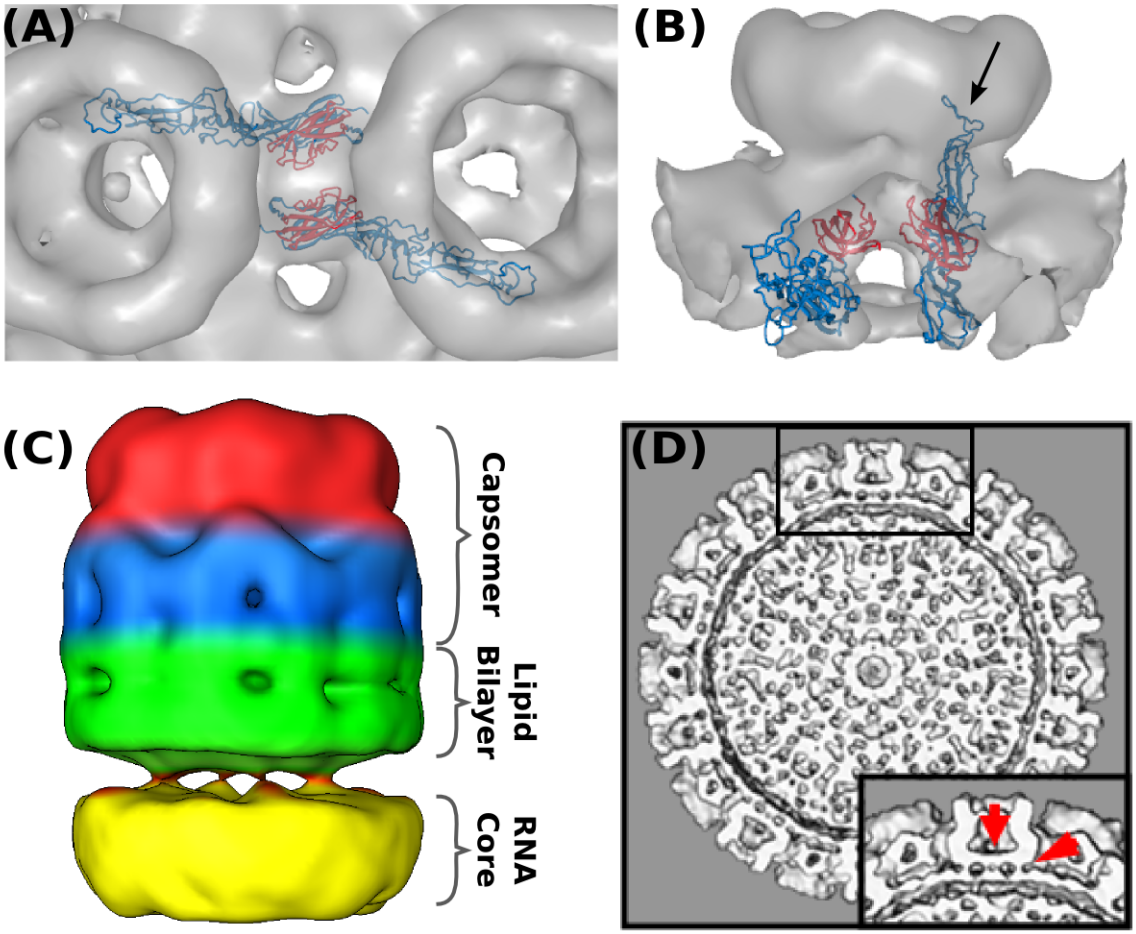

Figure 4: Inter-capsomer connections. (A) Top-view of two neighboring capsomers (grey cryoEM density) with two Gc monomers shown in blue. The domain III's (red) are very well positioned within the ridges connecting adjacent capsomers. The fusion peptide is directed to the capsomer center. (B) Side-view of one capsomer along the tunnel located beneath the connecting ridges. Two Gc's are shown and their proposed position within the cryoEM density. The black arrow indicates the location of the fusion peptide within domain II. The domain III's are highlighted in red to indicate their placement within the ridges. (C) CryoEM density of one extracted penton at a very low threshold (0.54). The outer region of the capsomer is indicated in red (representing mainly Gn molecules), the capsomer base in blue (representing mainly Gc molecules), the lipid envelope in green and the density corresponding to the RNP core is shown in yellow. Densities spanning the gap between the lipid bilayer and the RNP core are representing the glycoprotein cytoplasmic tails. (D) Surface-shaded representation of the central section of the RVFV cryoEM map viewed along the 5-fold orientation. The sections show glycoprotein protrusions on virus surface, lipid bilayer, and RNP core. In the lower right corner a blow-up of the boxed area is shown. Red arrows point to clearly defined densities spanning the lipid bilayer. These densities represent glycoprotein transmembrane domains and are located either on the outer edge of the capsomer or directly beneath the connecting channels.

multiple arrangements. While phlebovirus glycoproteins are arranged on the virion surface in a T=12 icosahedral symmetry, the hantavirus glycoproteins are arranged in a grid-like pattern. The bunyavirus Gc protein seems to be involved in the fusion process for each genus, but one remarkable difference between the two glycoproteins of the different genera is the protein size. It might be possible that while the general glycoprotein maturation process in the host cell follows a similar pathway for members of all the five genera, the size and domain architecture could be one of the factors contributing to a different arrangement of the glycoproteins on the virus surface. However, due to the lack of an experimental structure for any bunyavirus glycoprotein, we applied fold-recognition structure prediction to generate 3D structural models for the RVFV Gn and Gc monomers. The glycoprotein structures have been further analyzed in combination with the RVFV cryoEM structure previously solved by our group.

Hypothetical assembly model for Rift Valley fever virus

The two RVFV glycoproteins, Gn and Gc, are organized in 122 distinct capsomers on the virus surface, extending ~96 Å above the lipid envelope. Based on our docking results, Gn-Gc heterodimers form the basic structural unit. We hypothesize that hexons and pentons are comprised of six and five Gn-Gc heterodimers, respectively, with the Gn protein being more solvent exposed and forming the capsomer spike and the Gc protein lying partially underneath, closer to the lipid membrane forming the capsomer base. This arrangement is likely, since neutralizing monoclonal antibodies against Gn and Gc have been described [3]. In addition to interactions between Gn and Gc within each heterodimer, there are

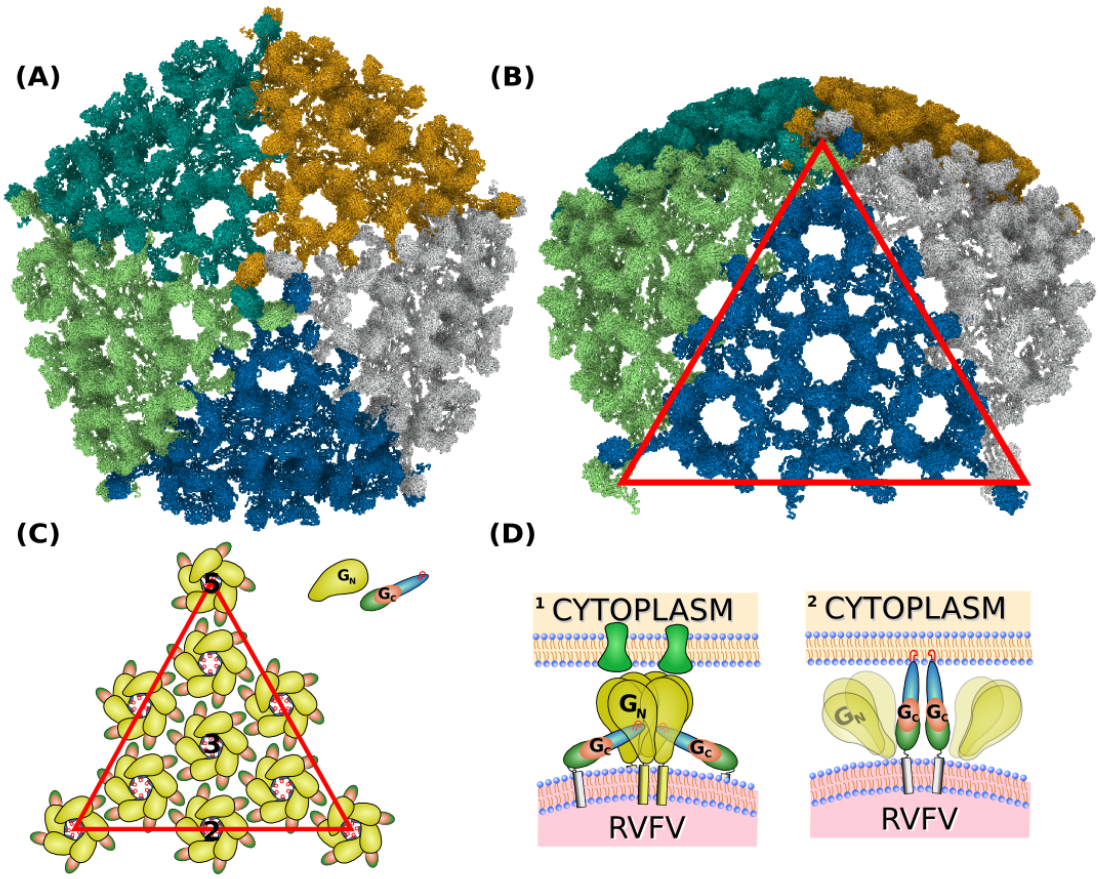

Figure 5: Overview of the RVFV glycoprotein shell. (A) The proposed T=12 icosahedral protein layer formed by Gn and Gc. Individual subunits are color coded. (B) Tilted representation as shown in (A). The red triangle represents one triangular face. (C) Schematic representation of the Gn and Gc contacts. Drawn is one of the 20 triangular faces of the icosahedrons enclosing the RVFV particle and the distribution of the Gn and Gc glycoproteins (corresponding to red triangle in (B)). Black numbers denote icosahedral 2-, 3-, and 5-fold symmetry axes. Gn monomers are represented as bulb-like structures in yellow, and Gc monomers as a tube-like structure. The individual domains are represented in red (domain I), blue (domain II) and green (domain III). The fusion peptides are indicated as red circles, and are pointing to the capsomer center. (D) Hypothetical model of the RVFV – host cell interaction. The RVFV glycoproteins Gn and Gc are represented according to our model and show similarities to the alphavirus E1 and E2 proteins. (1): Gn is depicted as the receptor-binding protein and binds to the host cell receptor (green). (2): After receptor binding the uptake of the RVFV particle is initiated and an acidification step of the endocytic vesicle triggers the dissociation of Gn and Gc. This results in the formation of potentially Gc trimers (in accordance with current models for class II fusion proteins) and insertion of the fusion peptides into the host cell membrane.

also interactions between neighboring structural units present. A Gc molecule from one heterodimer contacts the stalk region of an adjacent Gn molecule, which is part of another heterodimer. A recent study has shown, that hantavirus glycoproteins form complex intra- and intermolecular disulfide bonds between Gn and Gc and contribute to the assembly and stability of the virus particle [15]. The RVFV Gn and Gc ectodomains used for the molecular modeling have 23 and 20 cysteines, respectively, and it is possible that similar inter- and intramolecular disulfide bonds are present as well.

The spike complex of the alphavirus Semliki Forest virus consists of a trimer of heterodimers [(E1-E2)3] and is mediated by interactions between E2 and E1 TMDs [27, 37]. Even though we did not include the glycoprotein TMD and CTD in our fold predictions, it is possible that the Gn and Gc proteins potentially also interact with each other via their transmembrane regions and that the glycoproteins within one capsomer interact with the ribonucleoprotein complex via the Gn/Gc cytoplasmic tails. The interaction of the TMDs could represent an additional determinant of the heterodimer assembly. This hypothesis is strengthened by our description of protein densities spanning the space between the RNP core and the lipid bilayer within the RVFV particle (Figure 4C) [42]. A recently published study by Piper et al., [34], described the necessity of the RVFV Gn protein for genome packaging and showed that the Gn cytoplasmic tail is required for this process. In our RVFV cryoEM reconstruction we noticed the presence of densities spanning the virus envelope at the positions of capsomers [42]. These densities most likely represent the Gn and Gc TMDs and seem to be situated directly at the center of the ridges between neighboring capsomers and at the outer edges of the capsomers (red arrows in Figure 4D).

In contrast to many other lipid enveloped RNA viruses, bunyaviruses do not contain a

matrix protein which has the function of association and stabilization between nucleocapsid and viral envelope proteins. Based on our hypothetical assembly model, we suggest that a highly organized arrangement of the Gn and Gc glycoproteins is responsible for the overall virion stability and that the capsomer-capsomer interactions might play a central role in defining the icosahedral symmetry.

Multiple monoclonal antibodies against RVFV Gn and Gc have been described [3, 4] and for some of these antibodies the epitope on the ectodomain of Gn has been mapped [24]. In our model, the epitopes for the monoclonal antibodies 4-D4 and 4-32-8D, which have neutralizing and protective function, are localized and surface-exposed in the globular head domain of Gn (Figure 3D). This domain is capping the Gc domain II and fusion loop and it could be hypothesized that the neutralizing effect of these two antibodies might be explained by either preventing receptor binding or potential rearrangement of Gn post receptor attachment and, hence, inhibition of fusion, since the fusion loop will not be exposed to the host membrane. Another monoclonal antibody, 3C-10, has been described as non-neutralizing and non-protective in the mouse model and the epitope is localized in the stalk region of the Gn model (Figure 3D). In our model, this region can be found to be localized close to the ridges, connecting adjacent capsomers (Figure 3A). It might be possible that this epitope is not freely accessible in the native confirmation within the virion. The Gc domain III's forming the capsomer connections are representing a steric block preventing antibody binding.

In conclusion, structural models have been developed for the RVFV glycoproteins, Gn and Gc, and the structural aspects of the protein models allowed us to generate an assembly model which indicates how Gn and Gc could interact within and between capsomers. The

model for the icosahedral shell of RVFV is presented in Figure 5. Our hypothesized model has more similarity to the assembly of alphaviruses, than to flaviviruses. The main difference to flaviviruses derives from the fact, that in bunyaviruses the receptor binding and membrane fusion activities most likely reside in two different glycoproteins (similar to the alphaviruses), whereas in flaviviruses the E protein is responsible for both. Further, the fusion peptide of the RVFV Gc protein sticks up and packs against Gn, similar to the E1 and E2 proteins in SFV, while in the flaviviruses the fusion peptides are held down and pack against the interface of the E protein domain I and III.

The presented arrangement of Gn and Gc and description of their interactions may play an important role in glycoprotein folding and maturation, capsomer and virus assembly, virus fusion, and neutralization of infection. Future site-directed mutagenesis experiments using a reverse-genetics system [20] can be applied to evaluate the proposed glycoprotein interactions. The new information reported in this study, will not only impact the current understanding about the assembly of phleboviruses, but can also be exploited in understanding the antigenic and serological properties of this virus and help designing effective antivirals.

Acknowledgments

This work was supported in part by a grant from National Institutes of Health (R01GM62968).

References

[1] S. L. Alam, C. Langelier, F. G. Whitby, S. Koirala, H. Robinson, C. P. Hill, and W. I. Sundquist. Structural basis for ubiquitin recognition by the human ESCRT-II EAP45 GLUE domain. *Nature Structural & Molecular Biology*, 13(11):1029–1030, 2006.

[2] A. J. Battisti, Y. K. Chu, P. R. Chipman, B. Kaufmann, C. B. Jonsson, and M. G. Rossmann. Structural studies of hantaan virus. *J. Virol.*, 85(2):835–841, 2011.

[3] T. G. Besselaar and N. K. Blackburn. Topological mapping of antigenic sites on the Rift Valley fever virus envelope glycoproteins using monoclonal antibodies. *Archives of Virology*, 121(1-4):111–124, 1991.

[4] T. G. Besselaar and N. K. Blackburn. The effect of neutralizing monoclonal antibodies on early events in Rift Valley fever virus infectivity. *Research in Virology*, 145(1):13–19, 1994.

[5] S. Birmanns, M. Rusu, and W. Wriggers. Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. *J. Struct. Biol.*, 173(3):428–35, 2011.

[6] C. Cole, J.D. Barber, and G.J. Barton. The Jpred 3 secondary structure prediction server. *Nucl. Acids Res.*, 36(suppl 2):W197–W201, 2008.

[7] M. S. Collett, A. F. Purchio, K. Keegan, S. Frazier, W. Hays, D. K. Anderson, M. D. Parker, C. Schmaljohn, J. Schmidt, and J. M. Dalrymple. Complete nucleotide se-

quence of the m RNA segment of rift valley fever virus. *Virology*, 144(1):228–245, 1985.

[8] R. M. Elliott. Bunyaviruses and climate change. *Clinical Microbiology and Infection*, 15(6):510–517, 2009.

[9] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*, Chapter 5:Unit 5.6, Oct 2006.

[10] A. N. Freiberg, M. B. Sherman, M. C. Morais, M. R. Holbrook, and S. J. Watowich. Three-dimensional organization of Rift Valley fever virus revealed by cryoelectron tomography. *J. Virol.*, 82(21):10341–8, 2008.

[11] C. E. Garry and R. F. Garry. Proteomics computational analyses suggest that the carboxyl terminal glycoproteins of Bunyaviruses are class II viral fusion protein(beta-penetrenes). *Theoretical Biology and Medical Modelling*, 1(1):10, 2004.

[12] S. R. Gerrard and S. T. Nichol. Characterization of the Golgi retention motif of Rift Valley fever virus G(N) glycoprotein. *J. Virol.*, 76(23):12200–12210, 2002.

[13] D. L. Gibbons, M.-C. Vaney, A. Roussel, A. Vigouroux, B. Reilly, J. Lepault, M. Kielian, and F. A. Rey. Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus. *Nature*, 427(6972):320–325, 2004.

[14] R. Gupta, E. Jung, and S. Brunak. Prediction of N-glycosylation sites in human proteins, In preparation, 2004.

[15] J. Hepojoki, T. Strandin, A. Vaheri, and H. Lankinen. Interactions and oligomerization of hantavirus glycoproteins. *J. Virol.*, 84(1):227–242, 2010.

[16] J. Heyd and S. Birmanns. Immersive structural biology: a new approach to hybrid modeling of macromolecular assemblies. *Virtual Reality*, 13:245–255, 2009.

[17] T. Hirokawa, S. Boon-Chien, and S. Mitaku. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, 1998.

[18] J. T. Huiskonen, J. Hepojoki, P. Laurinmäki, A. Vaheri, H. Lankinen, S. J. Butcher, and K. Grunewald. Electron Cryotomography of Tula Hantavirus Suggests a Unique Assembly Paradigm for Enveloped Viruses. *J. Virol.*, 84(10):4889–4897, 2010.

[19] J. T. Huiskonen, A. K. Överby, F. Weber, and K. Grunewald. Electron cryo-microscopy and single-particle averaging of Rift Valley fever virus: evidence for GN-GC glycoprotein heterodimers. *J. Virol.*, 83(8):3762, 2009.

[20] T. Ikegami, S. Won, C. J. Peters, and S. Makino. Rift Valley fever virus NSs mRNA is transcribed from an incoming anti-viral-sense S RNA segment. *J. Virol.*, 79(18):12106–12111, 2005.

[21] L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, and A. Godzik. FFAS03: a server for profile–profile sequence alignments. *Nucl. Acids Res.*, 33(Web Server issue):W284–288, July 2005. PMID: 15980471.

[22] L. Kaján and L. Rychlewski. Evaluation of 3D-Jury on CASP7 models. *BMC Bioinformatics*, 8:304, 2007.

[23] L. T. Kakach, J. A. A. Suzich, and M. S. Collett. Rift Valley fever virus M segment: phlebovirus expression strategy and protein glycosylation. *Virology*, 170(2):505–510, 1989.

[24] K. Keegan and M. S. Collett. Use of bacterial expression cloning to define the amino acid sequences of antigenic determinants on the G2 glycoprotein of Rift Valley fever virus. *J. Virol.*, 58(2):263–270, 1986.

[25] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, 305(3):567–580, 2001.

[26] S. S. J. Lee, V. Knott, J. Jovanović, K. Harlos, J. M. Grimes, L. Choulier, H. J. Mardon, D. I. Stuart, and P. A. Handford. Structure of the integrin binding fragment from fibrillin-1 gives new insights into microfibril organization. *Structure*, 12(4):717–729, 2004.

[27] J. Lescar, A. Roussel, M. W. Wien, J. Navaza, S. D. Fuller, G. Wengler, G. Wengler, and F. A. Rey. The Fusion glycoprotein shell of Semliki Forest virus: an icosahedral assembly primed for fusogenic activation at endosomal pH. *Cell*, 105(1):137–148, 2001.

[28] L. Li, J. Jose, Y. Xiang, R. J. Kuhn, and M. G. Rossmann. Structural changes of envelope proteins during alphavirus fusion. *Nature*, 468(7324):705–708, 2010.

[29] P. Y. Lozach, R. Mancini, D. Bitto, R. Meier, L. Oestereich, A. K. Överby, R. F. Pettersson, and A. Helenius. Entry of bunyaviruses into mammalian cells. *Cell Host & Microbe*, 7(6):488–499, 2010.

[30] MMWR. Centers for Disease Control and Prevention. *Rift Valley fever outbreak Kenya*, 56(4):73–76, Feb, 2 2007.

[31] A. K. Överby, R. F. Pettersson, K. Grünewald, and J. T. Huiskonen. Insights into bunyavirus architecture from electron cryotomography of Uukuniemi virus. *PNAS*, 105(7):2375–2379, 2008.

[32] A. K. Överby, V. L. Popov, R. F. Pettersson, and E. P. A. Neve. The cytoplasmic tails of Uukuniemi Virus (Bunyaviridae) G(N) and G(C) glycoproteins are important for intracellular targeting and the budding of virus-like particles. *J. Virol.*, 81(20):11381–11391, 2007.

[33] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera – A visualization system for exploratory research and analysis. *J. Comp. Chem.*, 25(13):1605–1612, 2004.

[34] M. E. Piper, D. R. Sorenson, and S. R. Gerrard. Efficient cellular release of Rift Valley fever virus requires genomic RNA. *PloS One*, 6(3):e18070, 2011.

[35] M. L. Plassmeyer, S. S. Soldan, K. M. Stachelek, J. Martín-García, and F. González-Scarano. California serogroup Gc (G1) glycoprotein is the principal determinant of pH-dependent cell fusion and entry. *Virology*, 338(1):121–132, 2005.

[36] M. L. Plassmeyer, S. S. Soldan, K. M. Stachelek, S. M. Roth, J. Martín-García, and Francisco González-Scarano. Mutagenesis of the La Crosse Virus glycoprotein supports a role for Gc (1066-1087) as the fusion peptide. *Virology*, 358(2):273–282, 2007.

[37] S. V. Pletnev, W. Zhang, S. Mukhopadhyay, B. R. Fisher, R. Hernandez, D. T. Brown, T. S. Baker, M. G. Rossmann, and R. J. Kuhn. Locations of carbohydrate sites on alphavirus glycoproteins show that E1 forms an icosahedral scaffold. *Cell*, 105(1):127–136, 2001.

[38] G. N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23:283–438, 1968.

[39] J. T. Roehrig, A. J. Johnson, A. R. Hunt, R. A. Bolin, and M. C. Chu. Antibodies to dengue 2 virus E-glycoprotein synthetic peptides identify antigenic conformation. *Virology*, 177(2):668–675, 1990.

[40] A. Roussel, J. Lescar, M. C. Vaney, G. Wengler, G. Wengler, and F. A. Rey. Structure and Interactions at the Viral Surface of the Envelope Protein E1 of Semliki Forest Virus. *Structure*, 14(1):75–86, 2006.

[41] C. S. Schmaljohn and S. T. Nichol. *Bunyaviridae*, volume 2 of *Fields Virology, D. M. Knipe and P. M. Howley (eds.)*. Wolters Kluwer, Philadelphia, PA., fifth edition, 2007.

[42] M. B. Sherman, A. N. Freiberg, M. R. Holbrook, and S. J. Watowich. Single-particle cryo-electron microscopy of Rift Valley fever virus. *Virology*, 387(1):11–15, 2009.

[43] X. Shi, J. Goli, G. Clark, K. Brauburger, and R. M. Elliott. Functional analysis of the Bunyamwera orthobunyavirus Gc glycoprotein. *The Journal of General Virology*, 90(Pt 10):2483–2492, October 2009.

[44] S. S. Soldan, B. S. Hollidge, V. Wagner, F. Weber, and F. González-Scarano. La Crosse virus (LACV) Gc fusion peptide mutants have impaired growth and fusion phenotypes, but remain neurotoxic. *Virology*, 404(2):139–147, 2010.

[45] J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson. Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science*, 303(5665):1866–1870, March 2004.

[46] N. D. Tischler, Aa Gonzalez, Ta Perez-Acle, M. Rosemblatt, and P. D. T. Valenzuela. Hantavirus Gc glycoprotein: evidence for a class II fusion protein. *The Journal of General Virology*, 86(Pt 11):2937–2947, 2005.

[47] G. E. Tusnády and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, 283(2):489–506, 1998.

[48] J. E. Voss, M. C. Vaney, S. Duquerroy, C. Vonrhein, C. Girard-Blanc, E. Crublet, A. Thompson, G. Bricogne, and F. A. Rey. Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature*, 468(7324):709–712, 2010.

[49] T. L. Wasmoen, L. T. Kakach, and M. S. Collett. Rift Valley fever virus M segment: cellular localization of M segment-encoded proteins. *Virology*, 166(1):275–280, 1988.

[50] L. Willard, A. Ranjan, H. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, and D. S. Wishart. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucl. Acids Res.*, 31(13):3316 –3319, 2003.

[51] W. Wriggers. Using Situs for the integration of multi-resolution structures. *Biophysical Reviews*, 2:21–27, 2010.

[52] W. Zhang, P. R. Chipman, J. Corver, P. R. Johnson, Y. Zhang, S. Mukhopadhyay, T. S. Baker, J. H. Strauss, M. G. Rossmann, and R. J. Kuhn. Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. *Nature Struct. Biol.*, 10(11):907–912, 2003.

[53] W. Zhang, S. Mukhopadhyay, S. V. Pletnev, T. S. Baker, R. J. Kuhn, and M. G. Rossmann. Placement of the structural proteins in Sindbis virus. *J. Virol.*, 76(22):11645–11658, 2002.

Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions*

*Manuscript Published in Journal of Structural Biology - doi:10.1016/j.jsb.2009.12.028

Mirabela Rusu[†,‡]

Stefan Birmanns[†,‡,§]

[†]School of Biomedical Informatics, University of Texas Health Science Center at Houston
[‡]These authors contributed equally to this work.
[§]Correspondening author. Fax: +1 713 500 3907

Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions

Abstract

A structural characterization of multi-component cellular assemblies is essential to explain the mechanisms governing biological function. Macromolecular architectures may be revealed by integrating information collected from various biophysical sources - for instance, by intepreting low-resolution electron cryomicroscopy reconstructions in relation to the crystal structures of the constituent fragments. A simultaneous registration of multiple components is beneficial when building atomic models as it introduces additional spatial constraints to facilitate the native placement inside the map. The high-dimensional nature of such a search problem prevents the exhaustive exploration of all possible solutions. Here we introduce a novel method based on genetic algorithms, for the efficient exploration of the multi-body registration search space. The classic scheme of a genetic algorithm was enhanced with new genetic operations, tabu search and parallel computing strategies and validated on a benchmark of synthetic and experimental cryo-EM datasets. Even at a low level of detail, for example 35-40Å, the technique successfully registered multiple component biomolecules, measuring accuracies within one order of magnitude of the nominal resolutions of the maps. The algorithm was implemented using the Sculptor molecular modeling framework, which also provides a user-friendly graphical interface and enables an instantaneous, visual exploration of intermediate solutions.

Keywords

simultaneous registration, multi-body registration, multicomponent, macromolecular assembly, cryo-electron microscopy, cryo-EM, multi-resolution modeling, genetic algorithms, tabu search

Introduction

Fundamental biological processes such as DNA transcription, protein translation or cellular transport are efficiently carried out by macromolecular assemblies through the coordinated interaction of their constituent biomolecules [2]. Thousands of different macromolecules coexist at a given time inside a cell, but only few have a well-characterized molecular mechanism [37]. The structural description of such assemblies is crucial to explain their functional behaviors. X-ray crystallography, a main source of high-resolution information, solved structures of cellular assemblies such as the ribosome [4, 40] or RNA polymerase II [19] . However, multicomponent complexes are refractory to structural determination by crystallography due to their large size and intrinsic flexibility. Therefore, crystal structures are often available only for individual fragments.

Alternatively, electron cryomicroscopy (cryo-EM) is an imaging technique suitable for the structural characterization of large systems in near-native environments. Two-dimensional projections are collected from the sample in solution and used for the reconstruction of a 3D volumetric map [14]. Although the number of cryo-EM maps determined at high-resolutions (3-5Å) has considerably increased over the last decade, low-resolution maps are still commonly obtained for asymmetric or/and dynamic assemblies. Such cryo-EM reconstructions provide information about the overall shape of macromolecules, but their reduced level of

detail prevents a direct atomic characterization. Yet, such low-resolution cryo-EM maps may be interpreted in relation to the crystal structure of component fragments through the application of multi-resolution modeling techniques.

Hybrid approaches are employed to integrate information from various biophysical sources, including, but not restricted to, X-ray crystallography and cryo-EM [5]. Atomic models of low-resolution cryo-EM maps may be generated by docking the atomic structure of the constituent biomolecules. Such models are often obtained by independently placing each fragment either using interactive molecular graphics software [31, 7] or by employing automatic techniques to optimize a goodness-of-fit measure. The optimization may be constrained to rigid-body transformations - translations and rotations [42, 39, 33, 34, 9, 24, 17] but can also include flexible deformations [41, 35].

Simultaneous registration of multiple subunits is beneficial to identify their native spatial organization inside the assembly. The additional information thus introduced provides spatial constraints that facilitate proper docking and prevent steric clashing. At low resolutions, independently fitted fragments may measure maximal correlations at the interior of the maps, where densities are high, but far from their correct docking position. Such spurious solutions are caused by the reduced interior detail of the reconstruction and/or due to the resolution heterogeneities [43]. By simultaneously registering all constituents, major steric clashes are limited as the correlation scores would be reduced for such cases.

Although valuable, such a simultaneous registration has a prohibitive computational cost. Identifying the optimal docking of one probe involves the exploration of six degrees of freedom. As the number of fragments increases, the dimensionality of the search space grows exponentially with a complexity of $O(n^{6N})$, where N is the number of registered pieces.

Albeit an exhaustive exploration of all possible rotations and translations can be achieved for one component [42], such investigation is unfeasible as additional constituents are taken into account.

A possible approach to solve the multi-body registration problem, while overcoming the computational complexity of an exhaustive exploration, involves limiting the search to a portion of the space. Computational techniques were proposed following this strategy. Some iteratively refine one component at the time while either masking the others [39] or removing the occupied volumes of already docked fragments [34]. Other methods are inspired from crystallographic refinement, and assume that an overall correct placement is already known before performing a local simultaneous refinement in real [11, 16] or reciprocal [22] space. Recently, *Lasker et. al.* proposed a simultaneous global docking technique that discretizes the search space around centroid points [28].

Here we introduce a novel optimization technique for the simultaneous registration of multiple atomic structures into cryo-EM envelopes. Based on a genetic algorithm, MOSAEC (Multi-Object Simultaneous Alignment by Evolutionary Computing) makes no assumption about the scoring landscape and enables the multi-body global registration without restricting the search to a particular region. Genetic algorithms (GA) are heuristics inspired by evolutionary biology, commonly employed to solve high-dimensional optimization problems [21, 20, 13]. Darwin's concepts of natural selection and survival of the fittest [12] are introduced in an iterative scheme to enable the optimization of a scoring function. An abstract representation of the solution is generated by converting the variable to be optimized - here the rotation and translation of the constituents - into a linear form known as a chromosome. A population of such individuals adapts towards an optimal score following a process that

mimics biological evolution. In MOSAEC, we adapted the classic scheme of a genetic algorithm to enhance the exploration of the search space. New genetic operators were introduced to preserve the genetic diversity of the population and were used in combination with parallel evolution of subpopulations. Moreover, the exploration of the complex search space was improved by including tabu regions - areas of the search space which are marked as local optima and thereby should not be further sampled.

In the following section, we will describe MOSAEC by first giving an overview of the method followed by the details of the implementation. Then, in the 'Results' section, we present the testing and validation of the algorithm on a series of synthetic and experimental datasets. We conclude with a discussion of the results.

Material and methods

MOSAEC is an optimization technique derived from genetic algorithms (GAs) that explores and identifies optima in the highly dimensional search space of the multi-body registration problem. An overview of the procedure is given next (also summarized in figure 1) followed by a more detailed description of MOSAEC's implementation.

Genetic algorithms

GAs are computational methods that mimic biological evolution to optimize a scoring function. These algorithms integrate the concept of natural selection and survival of the fittest, in an iterative scheme that progressively improves the solution while exploring the parameter search space [21, 20, 13]. Evolutionary algorithms such as GAs can be distinguished from

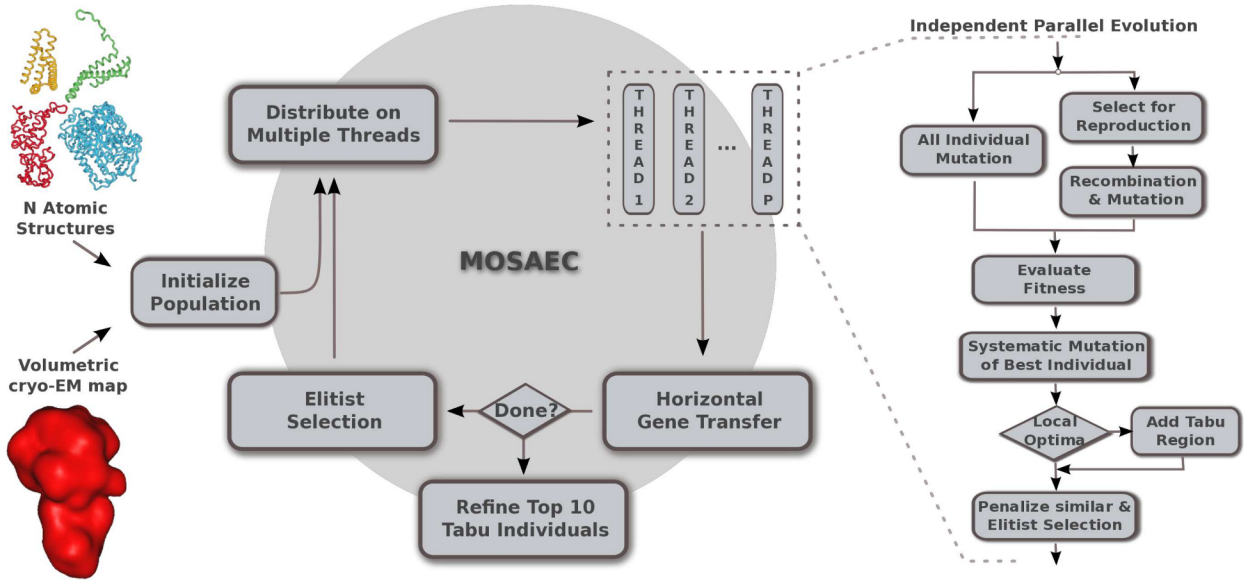

Figure 1: Schematic rendering of MOSAEC. (Left) The atomic structure of the constituent fragments and the volumetric map of the entire assembly are used as input for the algorithm. (Center) Parallel computing strategies are implemented to exploit both the multi-core architecture of current computers and the ability of GAs to explore different paths in the search space. (Right) In each independent thread, MOSAEC follows the classic GA scheme which was enhanced with new genetic operators and tabu-search strategies

the other optimization techniques as they consider a population of solutions instead of just a single one at a given point in time. The individuals in this population are a linear representation of the parameters to be optimized (see section 'Encoding of a candidate solution' on how the multi-body registration problem is represented). Each such individidual has a fitness value that indicates the optimality of the solution, i.e. the scoring evaluated for the encoded parameters. The algorithm starts with a set of individuals initialized through a random sampling of the search space. This population iteratively evolves under the influence of genetic operators while maximizing the fitness function, here the cross-correlation coefficient of the encoded atomic model and the target map. At each generation, a reproduction pool is selected with probabilities proportional to the fitness of the individuals. Recombination and mutation are applied to these solutions (sections 'Recombination' and '*Mutation*'), as well as

novel genetic operators (see description in the section '*Other genetic operators*'). Following mating, an improved population is selected based on an elitist reinsertion scheme detailed in section '*Reinsertion*', which ensures that better scoring individuals have higher chances to reproduce. Following this scheme that mimics mating, the scoring function is optimized progressively as better individuals are selected for the future generations. Also, tabu regions are introduced during reinsertion to prevent unnecessary explorations of regions marked as local optima (see section '*Tabu search*'). Moreover, MOSAEC exploits the stochastic nature of evolutionary strategies by allowing subpopulations to evolve in parallel (see description in section '*Parallel evolution*').

Encoding of a candidate solution

Each individual in the population represents an atomic model of the entire assembly, encoded as a linear string of real-valued genes representing translations and rotations of the constituent fragments. Individuals are composed of 4N genes $[\ldots, x_i, y_i, z_i, r_i, \ldots]$, $i = 1..N$ where N is the number of components, x_i, y_i, z_i represent the translation (in the space defined by the cryo-EM map) and r_i corresponds to an index in a list of rotational angles. This list provides a complete and uniform coverage of the 3D rotational space, reducing the search dimensionality (from 3 to 1) while at the same time avoiding gimbal lock problems. Each individual has associated a fitness value that quantifies the optimality of the solution they represent, i.e. the overlap between the multicomponent model and the cryo-EM map (see detailed description below).

The evolution starts with a population of n individuals randomly sampling the search space. In MOSAEC, this initial group of individuals is distributed over P threads that evolve

independently in parallel. Without loss of generality, we consider in the following that $P = 1$ and treat the case $P > 1$ in a later section. For each generation, the first step consists in selecting the individuals that are allowed to reproduce following a linear ranking scheme [3] in which higher mating probabilities are given to fittest individuals. The selected solutions undergo a process that simulates mating in which genetic operators such as recombination and mutation are applied to generate a population of offspring.

Recombination

The crossover operator enables the recombination of two "parent" individuals to create one or two offspring. The new individual(s) inherit(s) genes from the parents following a stochastic process that swaps/alters them following different schemes. For instance, the one-point crossover generates two offspring by swapping parental genes at only one location:

$$\begin{aligned} Parent1: &\quad [\ \ldots\ \ x_i\ \ y_i\ \ z_i\ \ r_i\ \ \ldots\ ] \\ Parent2: &\quad [\ \ldots\ \ X_i\ \ Y_i\ \ Z_i\ \ R_i\ \ \ldots\ ] \\ Offspring1: &\quad [\ \ldots\ \ x_i\ \ y_i\ \ Z_i\ \ R_i\ \ \ldots\ ] \\ Offspring2: &\quad [\ \ldots\ \ X_i\ \ Y_i\ \ z_i\ \ r_i\ \ \ldots\ ] \end{aligned}$$

while other schemes use multiple crossover locations, e.g. two-point or uniform crossover. Schemes may also generate only one offspring by applying arithmetic operations such as averaging. The recombination through crossover is based on the building block hypothesis which considers that better individuals may be generated from the best partial solutions of previous generations. This process enables a guided and efficient exploration of the search space.

MOSAEC stochastically applies each of these schemes.

Mutation

This operator takes a single individual and alters its genes creating a contiguous individual. Similar to the crossover, different schemes have been defined and used in MOSAEC, some randomly modifying the genes while other schemes only introduce small variations. Although such adjustments often model a bell curve, in MOSAEC they follow a Cauchy distribution:

$$C(\alpha, \beta, x) = \beta/(\pi \cdot (\beta^2 + (x - \alpha)^2)) \tag{1}$$

where α is the statistical median and $\beta > 0$ corresponds to the half-width at half-maximum. Similar to the normal distributions, Cauchy distributions have high probabilities to create small variations, however it also introduces larger changes which help the algorithm to escape from local optima.

New genetic operators in MOSAEC

In addition to the recombination and mutation, MOSAEC also introduces two new genetic operators to enhance the exploration and exploitation of the search space. A systematic operator applies stochastic mutations to all individuals in the population. Although computationally expensive, this operator was shown in our tests to be helpful for the identification of a global optima. The second novel operator introduced in MOSAEC applies ten Cauchy mutations to each gene of the fittest individual, thereby accelerating the local refinement.

Reinsertion

Following mating, $2 * n + 1$ new individuals are created: n from the reproduction pool via crossover and mutation, another n from the systematic mutation operator and eventually one from local search around the fittest individual. After evaluating their fitness, these individuals are merged with the n solutions of the original population, creating a pool of $3 * n + 1$ individuals, from which only the best n will be selected for the next generation.

MOSAEC applies a reinsertion scheme based on the elitist selection with fitness penalties for highly similar individuals. Classic elitist schemes conserve the fittest individuals typically without enforcing preservation of the genetic diversity. Maintaining a heterogeneous population is essential when solving optimization problems, in particular for complex cases that show multiple local optima. In MOSAEC, highly similar individuals are penalized if their gene distance (square root mean deviation of the gene values) is below a threshold inducing a decrease in fitness value (default by 10%).

Tabu Search

The exploration of the search space was enhanced in MOSAEC by introducing a tabu search strategy to prevent premature convergence to local optima. Such strategies are heuristics that combine local searches with adaptive memory to store the solutions [18]. MOSAEC considers a region as tabu, if the fittest individual has essentially not improved over the past ($T = 30$) generations. When a tabu region is introduced, the fittest individual is preserved in the list of optima and the region around it is considered prohibited and not allowed further exploration. MOSAEC introduces by default small tabu regions to prevent that they contain

more than one local optima. At the end of the run, the list of optima is examined and the top ten fittest individuals are refined.

Parallel evolution

Due to the stochastic nature of GAs, independent executions of the algorithm with the same initial population may result in the exploration of different regions of the search space. To take advantage of such a behavior, we modified the classic scheme of a GA to allow an independent evolution of subpopulations followed by a horizontal gene transfer. Identical subpopulations are distributed on different threads and are permitted to evolve for a small number of generations (100 generations by default). If our implementation is executed on a multi-core machine, such independent evolutions can run in parallel on different processing units. The user can choose the number of independent threads that will run in parallel, which typically should be the same as the number of cores available in the system. At the end of each cycle, the resulting subpopulations are merged and only the top individuals are selected (same number as in the initial subpopulation). This cycle is repeated until the total number of generations is achieved (Figure 1).

Fitness evaluation

Each individual in the population has a fitness value that quantifies the optimality of the solution it encodes. In MOSAEC, the fitness is assessed using the standard cross-correlation coefficient between the multicomponent atomic model and the volumetric map of the assembly as defined in eq. 2. ρ_{calc} and ρ_{em} are the direct space density distributions of the model

and of the cryo-EM map, $\overline{\rho}$ and $\sigma(\rho)$ are the average and, respectively, the standard deviation of a distribution ρ while T_i represent the transformation applied to the i^{th}, $i = 1..N$ component (both rotation and translation included). The density distribution ρ_{calc} has identical dimensions as ρ_{em} and was obtained by projecting the atoms of the model onto a 3D lattice followed by a Gaussian blurring. Similar cross-correlation coefficients are employed by others, see [43] for a review.

$$CCC(..., T_i, ...) = \frac{\int (\rho_{em}(\mathbf{r}) - \overline{\rho_{em}}) \cdot \left(\rho_{calc}(..., T_i, ..., \mathbf{r}) - \overline{\rho_{calc}(..., T_i, ...)}\right) d^3\mathbf{r}}{\sigma(\rho_{em}) \cdot \sigma(\rho_{calc}(..., T_i, ...))} \quad (2)$$

A coarse version of the cross-correlation coefficient was also implemented in MOSAEC to accelerate the execution. This score is computed following eq. 2 using coarse representations for ρ_{calc} and ρ_{em}. Topology-representing networks (TRN) were applied on the model to generate a simplified representation using feature points [45]. Such clustering techniques have been frequently employed in multi-resolution modeling of cryo-EM data [44, 41, 6, 7, 35]. These feature points were then projected onto the 3D lattice and low-pass filtered with a Gaussian kernel. Moreover, tri-linear interpolation can optionally be applied in MOSAEC to reduce the dimensions of the map ρ_{em}, for a further decrease in computational cost.

Fitness values in MOSAEC can be computed following the before mentioned forms of cross-correlation coefficients, whereby, according to our tests, even the coarse version is sufficient to identify global optima up to resolutions of 35-40Å. Note that contour enhancing filters, such as the Laplacian, were not applied in our validations, nor additional terms to penalize overlap between fragments.

Results

The performance of the method was assessed on multiple synthetic and experimental datasets. In this section, we present the results of this evaluation along with a study of the cross-correlation coefficient landscape in a simultaneous versus an independent registration.

Synthetic datasets

The benchmark for the validation of MOSAEC included simulated datasets of several biomolecular systems (Table 2). The component domains of these complexes were simultaneously docked into the volumetric map of the entire assembly, generated by Gaussian low-pass filtering to different resolutions. The best atomic model generated (measuring the highest cross-correlation coefficient during the run) was then compared with the native configuration of the assembly, as defined by the crystal structure.

Table 2: Biomolecular systems used for the validation of MOSAEC

Systems	PDB ID	# Atoms	# Parts	Refs
Oxido-reductase	1NIC	7908	3	[1]
Catalase	1QQW	16048	4	[27]
$I\kappa B\alpha/NF-\kappa B$ Complex	1IKN	4767	4	[23]
Helicase	1XMV	13338	6	[46]
GroEL	1OEL	26929	7	[8]

First, we present the progress of the best atomic model during a run for the pentamer Succinate Dehydrogenase (PDB ID 1NEK, [47]). This system was chosen to demonstrate the ability of the algorithm to explore a complex search space and to identify the global optima. Four fragments, of different size and shape, were registered into a 10Å-resolution synthetic map. Figure 2 shows the evolution of the best score over multiple iterations. Starting with

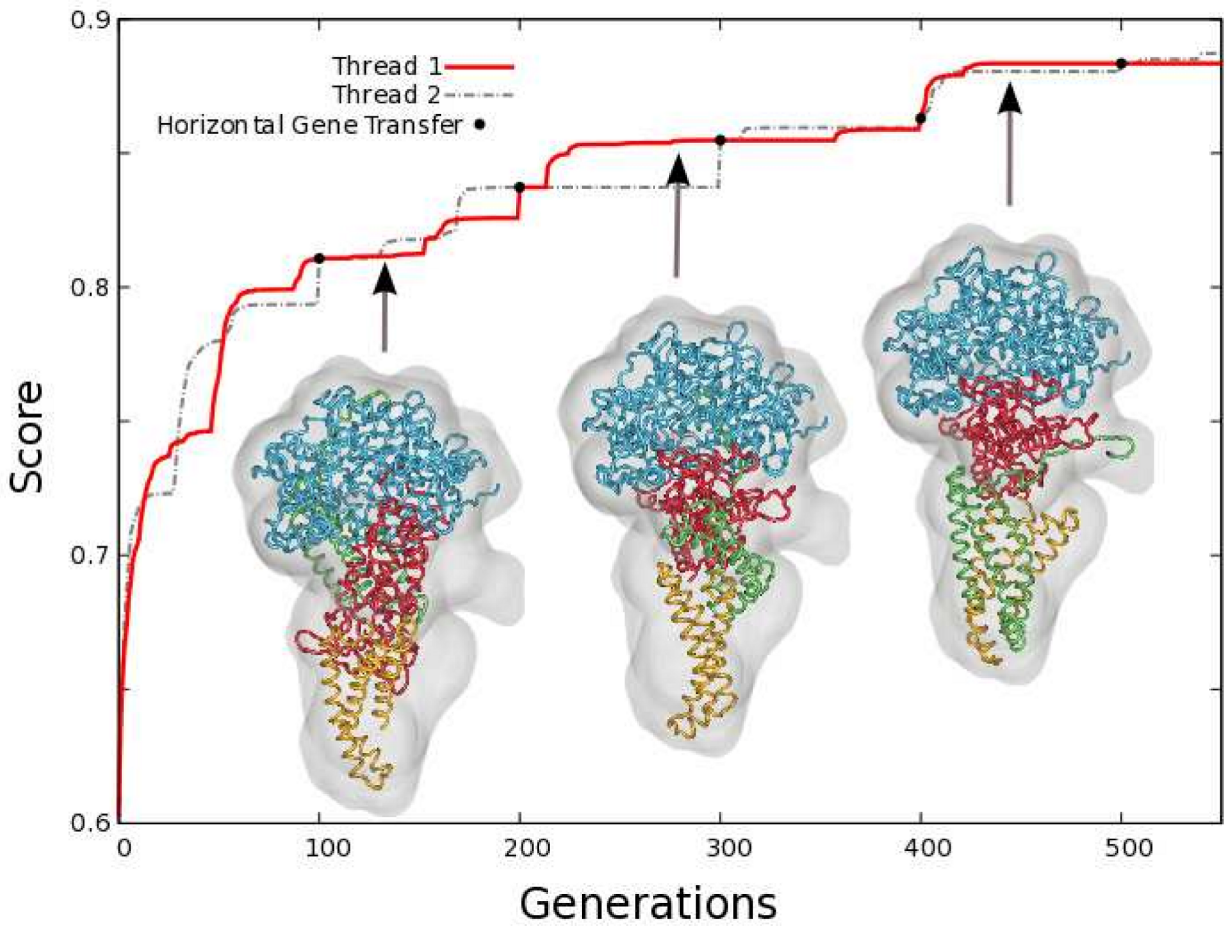

Figure 2: The evolution of the best score during a MOSAEC run in which four fragments were simultaneously docked into a 10Å resolution map of Succinate Dehydrogenase (PDB ID 1NEK).

a random distribution of the fragments, MOSAEC increases the scoring function within the first generations by placing all components inside the molecular envelope, but this placement is not optimal yet. As the evolution progresses further, the algorithm identifies the correct translation and rotation of each fragment, where often the large domains are found first, followed later by the smaller ones (see thumbnails in Figure 2). Identification of a native configuration is facilitated by the insertion of tabu regions as they enhance the investigation of unexplored areas within the search space.

Moreover, the independent parallel evolution of subpopulations, followed by horizontal

gene transfer, also enhances the sampling as different paths are explored at the same time. Indeed, we can observe in Figure 2 that different scores and local optima are reached in the parallel evolution, for example between generations 100 and 200. However, the horizontal gene transfer ensures that the best optima are conserved and that the diversity of the population is maintained.

In a second step, we put MOSAEC to a stringent test to assess the performance of the algorithm at different resolutions. The biomolecular systems presented in Table 2 were used for validation at resolutions ranging between 6Å and 40Å. These systems have different complexities, some only require the registration of three fragments while others have up to seven components. At each run, the root mean squared deviation (RMSD), measured in Ångström (Å), between the best atomic model and the native configuration was measured and plotted in Figure 3. These tests indicated that MOSAEC was successful in simultaneously docking multiple fragments up to 40Å resolution, with accuracies within one order of magnitude of the nominal resolution of maps.

Experimental datasets

The performance of the method was also assessed using experimental datasets. We performed a simultaneous registration of the bacterial ribosome and of the chaperonin GroEL.

The ribosome is the macromolecular assembly responsible for the protein translation, that enables the synthesis of polypeptide chains using the genetic information of the messenger RNA [32, 30]. Ribosomes are complexes of RNAs and proteins, and are organized into two subunits [48]. We carried out the simultaneous docking of these two fragments (PDB

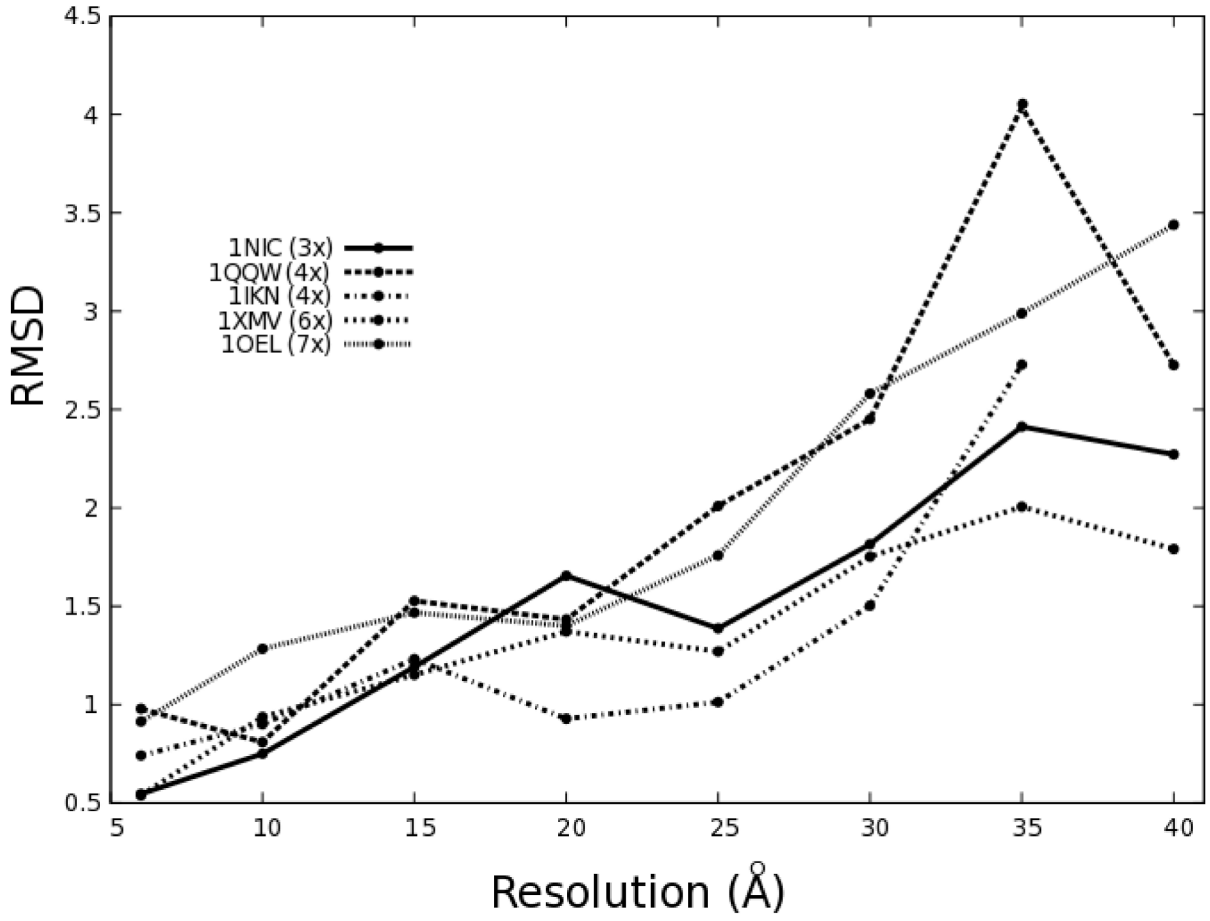

Figure 3: The accuracy of MOSAEC estimated in synthetic test cases at different resolutions. Root mean squared deviations (RMSD) were measured between the model generated and the known solution (Values computed for all atoms and shown in Ångström (Å)) .

IDs 1GIX, 1GIY, [48]) into the cryo-EM map of the assembly solved at 14 Å resolution (ID: emd-1005, [25]). MOSAEC successfully identified a native configuration (Figure 4), although only trace atoms were available for the crystal structures of the subunits. The model thus generated measures a 7.03Å RMSD from the one proposed by the authors of the map, but it improved the cross-correlation coefficient from 0.286 to 0.321 (measurement on alpha carbons and phosphates).

GroEL is a bacterial chaperonin that in association with co-chaperonin GroES is involved in the folding of proteins [38, 36, 15]. Our validation includes the cryo-EM map of GroEL

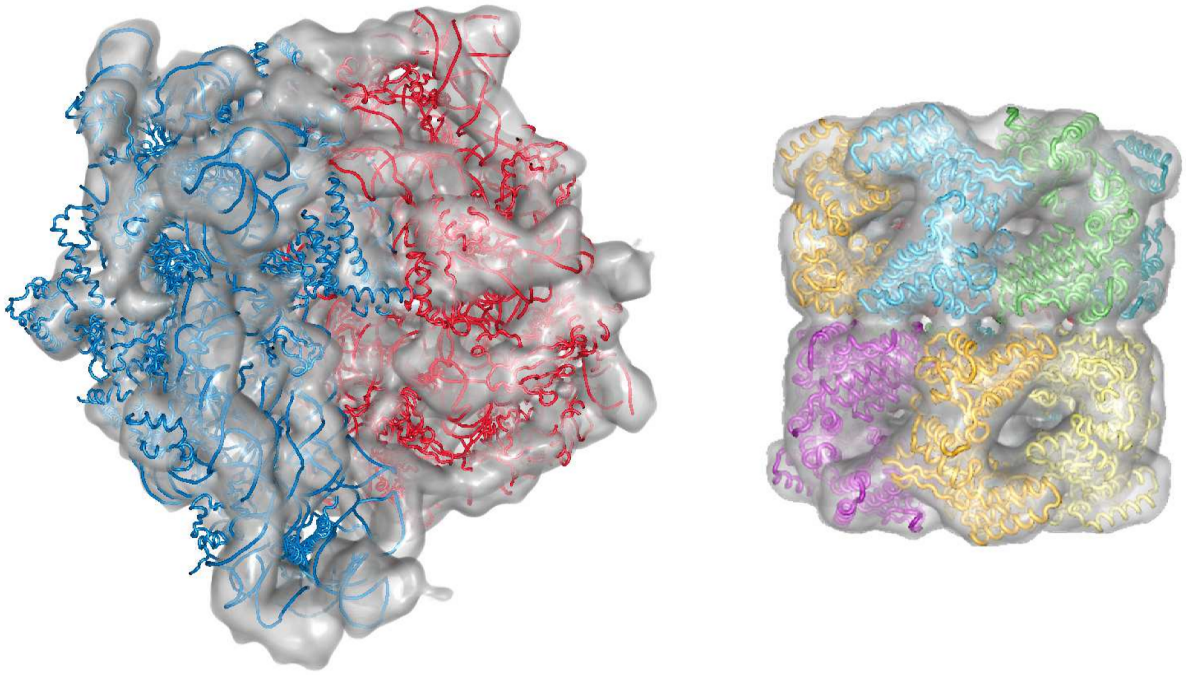

Figure 4: Experimental benchmark: (Left) ribosome - two subunits docked into a 14Å-resolution map (emd-1005); (Right) chaperonin GroEL - 14 monomers fitted into the 11.5Å resolution map (emd-1080)

alone as a double heptameric ring which displays a barrel-shape architecture. Fourteen monomers were simultaneously docked (PDB ID 1OEL [8]) into the 11.5Å resolution map (emd-1080,[29]). MOSAEC properly placed all these components, displaying a correlation coefficient of 0.947 with the experimental map (Figure 4).

Scoring landscape in simultaneous versus independent registration

Although MOSAEC introduces a novel optimization technique, the scoring function used to assess the model is the classic density-based cross-correlation coefficient (used in similar forms by other programs [26, 44, 39, 33, 34]). This goodness-of-fit measure is computed in MOSAEC using all component fragments in the model (see eq. 2). Yet, one can indepen-

dently dock each fragment at a time using readily available techniques [44, 39, 33, 34] and assemble a complete model from the top scoring solutions. This model will not necessarily maximize eq. 2, but the additive measure:

$$CCC_{\Sigma}(T_1, \cdots, T_N) = \sum_{i=1}^{N} CCC(T_i) \tag{3}$$

where T_i is the transformation that includes both translation and rotation of the i^{th}, $i = 1..N$ fragment, and $CCC(T_i)$ corresponds to the cross-correlation coefficient as defined in eq. 2. In the following, we investigate such a strategy and compare it with the simultaneous registration procedure proposed in MOSAEC. The discrepancies between the two approaches are shown by plotting the score landscape of eq. 2 and eq. 3 when fitting the three domains of homo-trimer oxido-reductase (PDB ID 1NIC) into a 15Å-resolution map. The high dimensionality of such a tri-body registration problem prevents the exhaustive exploration of all (18) degrees of freedom and, moreover, renders it difficult to visualize the results. Hence, here we show the landscape obtained when the position of only one fragment is variable within the plane known to contain the solution (rotations are all scanned), while the other two components are held fix at predefined locations inside the map. These locked components either occupy the configuration of the crystal structure (Figure 5A) or are placed at the center of the map (Figure 5C).

The first scenario, depicted in Figure 5A, represent a simple optimization problem in which only the configuration of one fragment must be identified given that the remaining domains are already properly docked inside the assembly. The multi-body correlation CCC (eq. 2) shows a prominent peak at the correct docking position (Figure 5B), yet the maxima

of the additive correlation is observed far from this location (Figure 5C). These results indicate that optimization techniques may promptly identify the native placement of the fragment using the multi-body correlation, but will provide spurious solutions when scoring models with the additive measure CCC_{Σ}.

Figure 5D shows a more difficult test in which the two fixed fragments occupy non-optimal docking positions, at the interior of the map. The multi-body correlation displays three peaks - one for each of the identical monomers in the crystal structure (Figure 5E). Due to the placement of the fixed components, the mobile fragment has three optimal scores instead of only one, as it can occupy either one of the three correct positions. When using this multi-body correlation score, optimization techniques are able to identify the placement of the monomer at one of the correct docking positions even if the rest of the components are arbitrarily placed inside the envelope.

On the other hand, CCC_{Σ} shows one global optima at the center of the map, far from the correct docking locations (Figure 5F). Moreover, this global optima scores higher than the best model in Figure 5C. Such landscape prevents the additive sum CCC_{Σ} from identifying the proper docking position of the fragments, creating models that show considerable overlap between constituents. To prevent such incorrect models, additive measures can be paired with terms that penalize the overlap between fragments [28]. Such multi-term scoring functions typically require an extra parametrization step to identify the weights of each element in the equation.

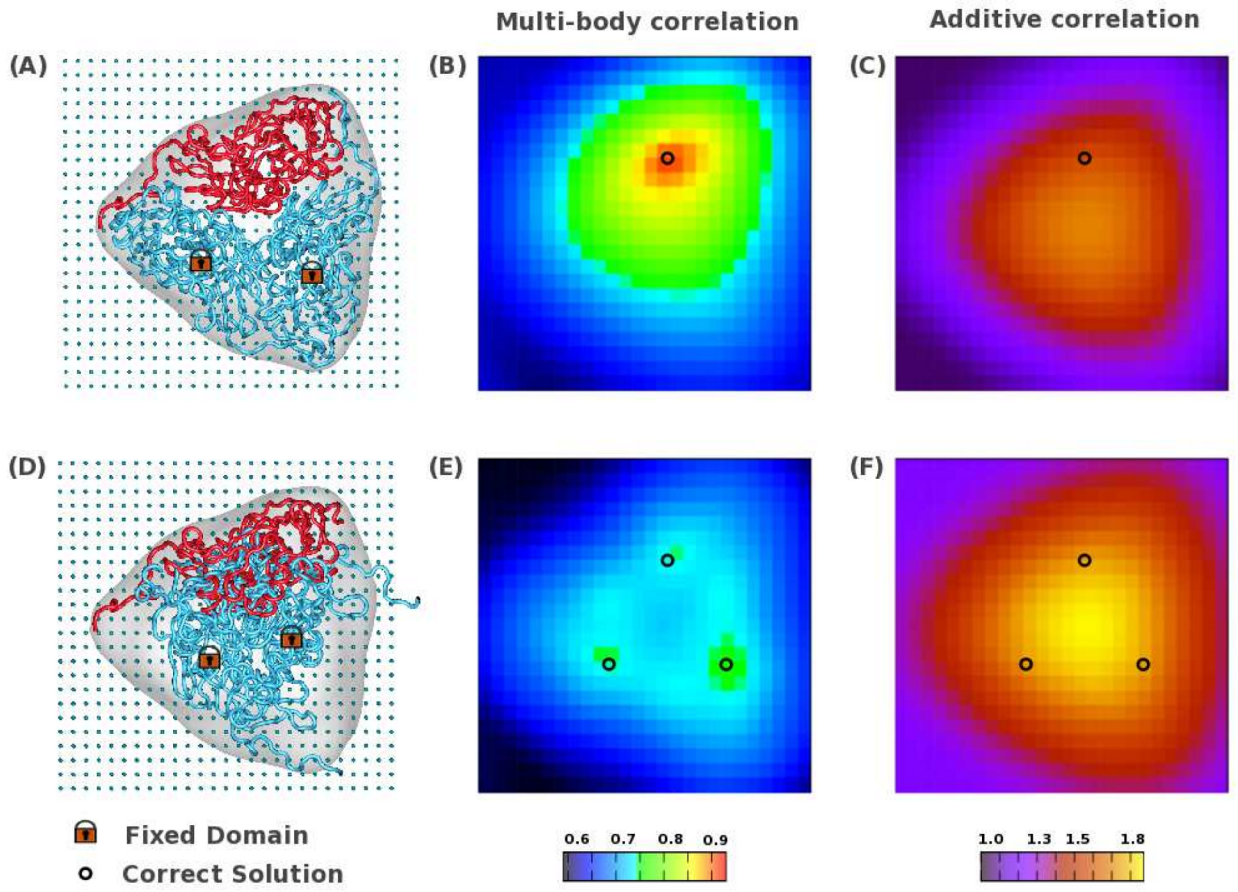

Figure 5: Scoring landscape of the multi-body correlation CCC and of the additive measure CCC_{Σ} for the homo-trimer oxido-reductase (PDB ID 1NIC). The landscape shows, for each grid position, the best score measured over all rotations (9° angular step size) in a scenario in which one fragment (red tube in A and D) is mobile on the grid and the other units are held fixed (blue tube) either in the crystallographic configuration (A) or at the center of the map (D).

Discussion

In this paper, we described a method for the simultaneous registration of multiple component atomic structures into cryo-EM volumetric maps of biomolecular assemblies. MOSAEC is a population-based optimization technique designed to explore the intricate and high-dimensional search space of the multi-body docking problem. This approach is derived from genetic algorithms and enhanced with parallel computing and tabu search strategies to enable

a better exploration of the scoring landscape.

MOSAEC successfully identified the spatial organization of constituent fragments within the cryo-EM envelope of the assembly. Our benchmark indicated that the algorithm is able to simultaneously register multiple component structures, identifying their placement and orientation with accuracies within one order of magnitude of the nominal resolution of the cryo-EM maps. Using the classic cross-correlation coefficient as a scoring function, such performance was observed for resolutions as low as 40Å. Maps with such low level of detail are typically beyond the reach of traditional docking methods that employ similar scores, but independently fit each component [10].

The successful registration was facilitated by the simultaneous docking of the constituent domains. The concurrent fitting of multiple structures indirectly introduces spatial constraints that guide the optimization towards identifying the correct configuration inside the complex. This additional information is especially beneficial at low-resolutions, where the volumetric maps have reduced interior detail and the boundaries between domains are ambiguous [43]. As opposed to other registration methods [39, 34, 28], these constraints are incorporated here solely by the shape of the scoring landscape and not by restraining the placement of the fragments to subregions of the search space.

Although the simultaneous registration favors the building of native atomic models, such an optimization procedure is computationally expensive. The calculation of the cross-correlation coefficient represents the most complex step of the approach, in particular for assemblies composed of a larger number of fragments, which require a more intensive sampling of the search space. To enable an efficient optimization, we employed a coarse scoring function (see Material and Methods section). This score allowed MOSAEC to successfully

register the biomolecular systems included in the benchmark (see Results section) with runtimes ranging from minutes to a few hours. For example, the seven monomers of GroEL were simultaneously fitted into a 20Å-resolution volumetric map in 139 min[¶] with an accuracy of 1.52Å RMSD from the known solution (700 individuals, 2000 generations and 4 parallel threads). These runtimes were obtained using the conservative default parameters of our software. However, tests indicated that smaller population sizes, a coarser representation of the data or of the score may still successfully identify the native configuration of the system, with up to a 36 fold speed up (3.8 min) and at the same time achieving an acceptable accuracy of 3.31Å RMSD (for a population size 100, 3.3 fold less feature vectors and no Gaussian blurring). Moreover, the deviations mentioned in this paragraph were computed before any optional refinement, which is available as a final step during a MOSAEC run in our software Sculptor.

Also, the optimization procedure was enhanced with parallel computing strategies accompanied by horizontal gene transfer. Such techniques were implemented both to exploit the multi-core architecture of current computers and to take advantage of the stochastic nature of genetic algorithms. Independent parallel evolutions are distributed on the available CPU cores to enable a more efficient exploration of the scoring landscape while investigating different pathways in the search space. The periodic horizontal gene transfer that follows each parallel evolution cycle ensures the conservation of the best individuals from each independent thread and the preservation of gene diversity in the population.

The previously mentioned outcomes were obtained using a default set of parameters that were estimated through empirical testing. The population size is the sole parameter that

[¶]runtime measured on a Dual-Core Intel Xeon processor 5140 @2.33 GHz

should be modified for each system to reflect the complexity of the assembly by setting its value proportional to the number of components to be registered (suggested scaling factor 100). All other parameters should otherwise be held constant as tests indicated that the algorithm is robust under changes in these values. Some parameters, such as the population size or the number of parallel threads, affect the sampling rate while others control the tabu search strategy influencing the amount of local optimization versus global search. The default values were selected to create a balance between sampling rate and runtime of the optimization, on one hand, and exploration and exploitation on the other.

The implementation of MOSAEC uses the C++ framework of our molecular modeling and visualization software Sculptor [7]. Sculptor provides a user-friendly graphical interface to set up the registration, to inspect intermediate results and to pause/restart/stop the optimization process when desired results were achieved. The interactive exploration of the intermediate results is possible in Sculptor due to the GA's characteristic to provide partial solutions to the problem during the optimization. Sculptor is freely available at `http://sculptor.biomachina.org`. In addition, we plan to develop a command-line version of the algorithm, to be distributed with the Situs program package.

To our knowledge MOSAEC is the first method to enable the simultaneous registration of multiple components on an essentially continuous search space. Without restricting the translations to a grid and with a rotational step size of just one degree, MOSAEC samples the scoring landscape in a continuous fashion making no assumptions about the shape of the system. The exploration of this search space is solely guided by the scoring function, a well established cross-correlation coefficient.

Acknowledgments

We thank Willy Wriggers for stimulating discussions and valuable advice regarding the project, Teresa Ruiz and Michael Radermacher for helpful comments and Manuel Wahle for kind input. The present work was supported by NIH grant R01GM62968, a grant from the Gillson-Longenbaugh Foundation, and startup funds from the University of Texas at Houston (to S.B.).

References

[1] E. T. Adman, J. W. Godden, and S. Turley. The structure of copper-nitrite reductase from *Achromobacter cycloclastes* at five pH values, with NO_2-bound and with type II copper depleted. *J. Biol. Chem.*, 270:27458–27474, 1995.

[2] B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92:291–294, 1998.

[3] J. E. Baker. Adaptive selection methods for genetic algorithms. In *Proceedings of the 1st International Conference on Genetic Algorithms*, pages 101–111, 1985.

[4] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–920, 2000.

[5] W. Baumeister and A. C. Steven. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.*, 25:624–631, 2000.

[6] S. Birmanns and W. Wriggers. Interactive fitting augmented by force-feedback and virtual reality. *J. Struct. Biol.*, 144:123–131, 2003.

[7] S. Birmanns and W. Wriggers. Multi-resolution anchor-point registration of biomolecular assemblies and their components. *J. Struct. Biol.*, 157(1):271–280, 2007.

[8] K. Braig, P. D. Adams, and A. T. Brünger. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. *Nature Struct. Biol.*, 2:1083–1094, 1995.

[9] H. Ceulemans and R. B. Russell. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *J. Mol. Biol.*, 338(4):783–793, May 2004.

[10] P. Chacón and W. Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.*, 317:375–384, 2002.

[11] M. S. Chapman. Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. *Acta Cryst. A*, 51(1):69–80, 1995.

[12] C. Darwin. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. London: John Murray, 1859.

[13] L. D. Davis and M. Mitchell. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, 1991.

[14] D. J. DeRosier and A. Klug. Reconstruction of three dimensional structures from electron micrographs. *Nature*, 217:130–134, 1968.

[15] W. A. Fenton and A. L. Horwich. Chaperonin-mediated protein folding: fate of substrate polypeptide. *Quart. Rev. Biophys.*, 36(2):229–256, 2003.

[16] H. Gao, J. Sengupta, M. Valle, A. Korostelev, N. Eswar, S. M. Stagg, P. Van Roey, R. K. Agrawal, S. C. Harvey, A. Sali, M. S. Chapman, and J. Frank. Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell*, 113(6):789–801, Jun 2003.

[17] J. I. Garzón, J. Kovacs, R. Abagyan, and P. Chacón. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*, 23(4):427–433, Feb 2007.

[18] F. Glover. Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.*, 13(5):533–549, 1986.

[19] A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science*, 292:1876–1882, 2001.

[20] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[21] J. H. Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI, 1975.

[22] R. Huber and M. Schneider. A group refinement procedure in protein crystallography using Fourier transforms. *J. Appl. Cryst.*, 18:165–169, 1985.

[23] T. Huxford, D. B. Huang, S. Malek, and G. Ghosh. The crystal structure of the IkappaBalpha/NF-kappaB complex reveals mechanisms of NF-kappaB inactivation. *Cell*, 95:759–770, 1998.

[24] W. Jiang, M. L. Baker, S. J. Ludtke, and W. Chiu. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.*, 308:1033–1044, 2001.

[25] B. P. Klaholz, T. Pape, A. V. Zavialov, A. G. Myasnikov, E. V. Orlova, B. Vestergaard, M. Ehrenberg, and M. van Heel. Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. *Nature*, 421:90–94, 2003.

[26] G. J. Kleywegt and T. A. Jones. Template convolution to enhance or detect structural features in macromolecular electron-density maps. *Acta Cryst. D*, 53:179–185, 1997.

[27] T. P. Ko, M. K. Safo, F. N. Musayev, M. L. Di Salvo, C. Wang, S. H. Wu, and D. J. Abraham. Structure of human erythrocyte catalase. *Acta Cryst. D*, 56(Pt 2):241–245, Feb 2000.

[28] K. Lasker, M. Topf, A. Sali, and H. J. Wolfson. Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J. Mol. Biol.*, 388(1):180–194, Apr 2009.

[29] S. J. Ludtke, J. Jakana, J. L. Song, D. T. Chuang, and W. Chiu. A 11.5 Å single particle reconstruction of GroEL using EMAN. *J. Mol. Biol.*, 314:253–262, 2001.

[30] K. Mitra and J. Frank. Ribosome dynamics: Insights from atomic structure modeling into cryo-electron microscopy maps. *Ann. Rev. Biophys. Biomol. Struct.*, 35:299–317, May 2006.

[31] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera – A visualization system for exploratory research and analysis. *J. Comp. Chem.*, 25(13):1605–1612, 2004.

[32] V. Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108:557–572, 2002.

[33] A. M. Roseman. Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Cryst. D*, 56:1332–1340, 2000.

[34] M. G. Rossmann, R. Bernal, and S. V. Pletnev. Combining electron microscopic with X-ray crystallographic structures. *J. Struct. Biol.*, 136(3):190–200, Dec 2001.

[35] M. Rusu, S. Birmanns, and W. Wriggers. Biomolecular pleiomorphism probed by spatial interpolation of coarse models. *Bioinformatics*, 24:2460–2466, 2008.

[36] H. Saibil. Molecular chaperones: containers and surfaces for folding, stabilising or unfolding proteins. *Curr. Opinion Struct. Biol.*, 10:251–258, 2000.

[37] A. Sali, R. Glaeser, T. Earnest, and W. Baumeister. From words to literature in structural proteomics. *Nature*, 422(6928):216–225, 2003.

[38] P. B. Sigler, Z. Xu, H. S. Rye, S. G. Burston, W. A. Fenton, and A. L. Horwich. Structure and function in GroEL-mediated protein folding. *Ann. Rev. Biochem*, 67:581–608, 1998.

[39] N. Volkmann and D. Hanein. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J. Struct. Biol.*, 125:176–184, 1999.

[40] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407(6802):327–339, 2000.

[41] W. Wriggers, R. K. Agrawal, D. L. Drew, A. McCammon, and J. Frank. Domain motions of EF-G bound to the 70S ribosome: Insights from a hand-shaking between multi-resolution structures. *Biophys. J.*, 79:1670–1678, 2000.

[42] W. Wriggers and S. Birmanns. Using Situs for flexible and rigid-body fitting of multi-resolution single molecule data. *J. Struct. Biol.*, 133:193–202, 2001.

[43] W. Wriggers and P. Chacón. Modeling tricks and fitting techniques for multiresolution structures. *Structure*, 9:779–788, 2001.

[44] W. Wriggers, R. A. Milligan, and J. A. McCammon. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.*, 125:185–195, 1999.

[45] W. Wriggers, R. A. Milligan, K. Schulten, and J. A. McCammon. Self-organizing neural networks bridge the biomolecular resolution gap. *J. Mol. Biol.*, 284:1247–1254, 1998.

[46] X. Xing and C. E. Bell. Crystal structures of *Escherichia coli* RecA in complex with MgADP and MnAMP-PNP. *Biochemistry*, 43:16142–16152, 2004.

[47] V. Yankovskaya, R. Horsefield, S. Törnroth, C. Luna-Chavez, H. Miyoshi, C. Léger, B. Byrne, G. Cecchini, and S. Iwata. Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science*, 299(5607):700–704, Jan 2003.

[48] M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. Crystal structure of the ribosome at 5.5 Å resolution. *Science*, 292:883–896, 2001.

Evolutionary Bidirectional Expansion for the Annotation of Alpha Helices in Cryo-Electron Microscopy Reconstructions*

*Manuscript submitted for publication in the Journal of Structural Biology

Mirabela Rusu[†,‡]

Willy Wriggers[§]

[†]School of Biomedical Informatics, University of Texas Health Science Center at Houston

[‡]To whom correspondence should be addressed: E-mail: mirabela.rusu@biomachina.org.

[§]Department of Physiology and Biophysics and Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1300 York Ave., New York, NY 10065. Permanent address: D. E. Shaw Research, 120 West 45th Street, New York, NY 10036

Evolutionary Bidirectional Expansion for the Annotation of Alpha Helices in Cryo-Electron Microscopy Reconstructions

Abstract

Cryo-electron microscopy (cryo-EM) enables the imaging of macromolecular complexes in near-native environments at resolutions that often permit the visualization of secondary structure elements. For example, alpha helices frequently show consistent patterns in volumetric maps, exhibiting rod-like structures of high density. Here, we introduce *HELEX* (HELix EXtractor) - a novel technique for the annotation of helical regions in cryo-EM data sets. *HELEX* combines a genetic algorithm and a bidirectional expansion with a tabu search strategy to locate and characterize helical regions. Our method takes advantage of the stochastic search by using a genetic algorithm to identify optimal placements for a short cylindrical template, avoiding exploration of already characterized tabu regions. These placements are then utilized as starting positions for the adaptive bidirectional expansion that characterizes the curvature and length of the helical region. The method reliably predicted helices with seven or more residues in experimental and simulated maps at intermediate (4-12 Å) resolution. The observed success rates, ranging from 70.6% to 100%, depended on the map resolution and validation parameters. For successful predictions, the helical axes were located within 2Å from known helical axes of atomic structures.

cryo-electron microscopy, intermediate resolution, extract alpha helices, annotate alpha helices, secondary structure elements

Introduction

The continuing progress in the field of cryo-electron microscopy (cryo-EM) due to the improvement of the instrumentation, data acquisition, and image processing techniques [2, 11] yields increasing numbers of biomolecular systems solved at intermediate to high resolution [9]. Our focus here is on the most abundant intermediate-resolution (6-10 Å) reconstructions that exhibit the characteristic density signatures of secondary structure elements.

There are already a number of existing tools for the annotation of such secondary structure elements in cryo-EM maps. HelixHunter is a semi-automatic approach that combines a thresholding-segmentation scheme with an exhaustive search using a short helical template [14]. In SSEHunter, a modified template search approach yielded α-helix and β-sheet probabilities for a coarse-grained representation of the map [1], which was then manually annotated by secondary structure type. EMatch is a more automated approach that combines a template search with a segmentation-linkage schema [18]. All of these template search techniques involve the discrete exploration of the cryo-EM map using a short cylindrical template, which is subject to a relatively coarse angular and translational sampling. More recently, helical regions were also predicted based on density gradient information [10]; however, the utility of the method was not yet demonstrated on experimental reconstructions. In addition to these algorithmic search approaches, it is still common practice to use manual identification of helical map regions in the modeling work flow. For example, α-helices were flexibly fit into low-resolution cryo-EM maps of transmembrane proteins[17], and folding topology was modeled from sequence-based secondary structure predictions [35, 19].

Here, we introduce the HELEX (HELix EXtractor) approach that annotates helical re-

gions in cryo-EM maps. Although significant contributions have already been made by other authors with respect to helix detection, we feel that there is still an opportunity to explore alternative methods. One of our aims was to enable a fully automatic exhaustive search with a novel, quasi-continuous sampling of orientations and translations that visualizes helices on the fly as they are being detected. Inspired by earlier filtering approaches [10, 6], we implemented a novel correction of map density variation for enhanced detection of helical densities. Another aim was to detect and follow the curvature of a helix.

The HELEX method combines a genetic algorithm (GA) for quasi-continuous sampling, a bidirectional expansion for following helical curvature and length, and a tabu search strategy for optimizing the exploration. Inspired by Darwinian evolution, GAs optimize a population of solutions allowed to evolve with operators such as mutation and crossover under the pressure of a scoring function [13, 12]. The evolutionary tabu search was introduced earlier for the simultaneous registration of multiple-component crystal structures with the cryo-EM map of their assembly [29]. HELEX uses a small cylindrical template for which three translations and two rotations are optimized. When sampling the cryo-EM map, the population of cylindrical templates evolves for several generations while maximizing the cross-correlation coefficient. The best scoring template is typically placed within a helical region, aligned to the helical axis. Further processing using a local bidirectional expansion then follows the curvature and determines the length of the helical region. Once identified, the helices are placed into a tabu list to avoid redundant exploration.

The Methods section provides a detailed description of the implementation of the algorithm. In the Results section, we present an extensive validation of HELEX on simulated and experimental maps with resolutions ranging from 4 to 14 Å. Finally, we describe com-

putational performance, advantages, and limitations in the Discussion section.

Methods

The work flow of HELEX shown in Figure 1 corresponds to the structure of this section. First, a novel local normalization filter for the cryo-EM map is introduced as a pre-processing step. Then, a detailed description of the genetic algorithm (GA), bidirectional expansion, and tabu search strategy is given. In the next step, we present the stop criteria and the post-processing of the helices. Finally, the validation procedure is described.

Local normalization of the cryo-EM map

A Gaussian-weighted local normalization was applied to the input map prior to launching HELEX. Such normalization is beneficial because it enhances the appearance of the helices and equalizes any uneven background density distributions in experimental cryo-EM maps (see Results). The filter is used only for helix detection, and no particular physical meaning is attributed to the resulting densities. For each voxel $\mathbf{r}$, the average $\overline{\rho_{em}(\mathbf{r})}$ and the standard deviation $\sigma(\rho_{em}(\mathbf{r}))$ of the densities are computed in the local neighborhood using weights that follow a Gaussian distribution. The parameter σ_W characterizes the spatial extent of the Gaussian and is given in voxel units. For the maps presented here, σ_W equals 1.5 voxel units for our simulated maps and 2.5 voxel units for our experimental maps. In practical applications the voxel spacing may not always follow the map resolution, so as a rule of thumb we suggest that σ_W should be equivalent to about half the nominal resolution of a map.

The locally normalized densities are then computed according to the formula

$$\rho'_{em}(\mathbf{r}) = \frac{\rho_{em}(\mathbf{r}) - \overline{\rho_{em}(\mathbf{r})}}{\sigma(\rho_{em}(\mathbf{r}))}\,, \tag{4}$$

where $\rho_{em}(\mathbf{r})$ is the original density at voxel $\mathbf{r}$ and $\rho'_{em}(\mathbf{r})$ represents the locally normalized density. The local normalization will amplify any exterior noise, so experimental maps that contain outside noise should be thresholded and/or segmented at the molecular surface density level to set exterior densities to zero. Here, the experimental maps were thresholded to the "suggested contour level for viewing the map" given by the EMDB Database[32]. Such thresholding does not affect the extraction of α-helices that correspond to higher density regions. No such thresholding or segmentation was applied to the simulated maps, which were created without noise. Although originally designed only for experimental maps, we observed that locally normalized simulated maps enhanced rod-like features, thus promoting the detection of α-helices. Consequently, the local normalization was applied to both experimental and simulated maps.

Genetic algorithm

Inspired by biological evolution, GAs use genetic operators such as mutation and crossover to optimize a fitness function in an iterative optimization [12]. GAs consider a population of candidate solutions and allow it to evolve over several generations according to an elitist scheme based on the principle of survival of the fittest. One reason for using the GA optimization for HELEX was that it allowed the approach to be integrated into our interactive molecular graphics software Sculptor[4] to visualize helices in real time as they

were identified. Another important reason for the GA was the possibility of supporting a quasi-continuous representation of translations and rotations.

In HELEX, each individual in the population represents a cylindrical template (radius = 2 Å, length = 20 Å) with the three translational and two rotational degrees of freedom as the free optimization parameters (the irrelevant rotation about the cylinder's main axis is ignored). These five degrees of freedom are encoded by four parameters [x,y,z; r_i], where x, y, and z represent the three dimensional translations and r_i is an index of the list of angles that uniformly sample the rotational degrees of freedom using an angular step size of 1 °. The angular sampling thus approaches a continuum, in contrast to earlier template convolution techniques that reported orientational steps of up to 15°. The fitness of each individual in the population is then estimated based on a cross-correlation coefficient that samples the cryo-EM map within the template cylinder mask.

Two genetic operators are considered during the GA evolution (see [29] for implementation details). The mutation modifies the transformation of the templates, allowing them to sample the cryo-EM map at different placements or with various orientations. Large mutations enable the template to explore the map, while small mutations have the effect of a more localized refinement. The mutation operator modifies randomly picked individuals by applying variations that follow a Cauchy distribution:

$$C(\beta, x) = \beta/(\pi \cdot (\beta^2 + x^2)) \tag{5}$$

where $\beta = 0.05$ corresponds to the standard deviation. Compared to a Gaussian, the Cauchy distribution is also biased to small variations, but it creates larger deviations with

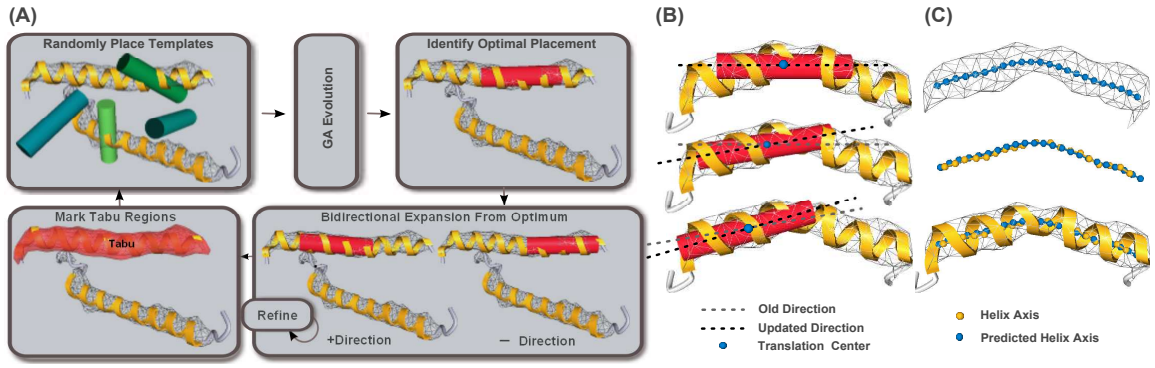

Figure 1: (A) HELEX work flow: A random initial population of cylindrical templates is allowed to evolve for several generations. The best scoring template is then used for the bidirectional expansion. The annotated region is included in the tabu list. A new GA run is executed, starting from new random distributions. (B) During the bidirectional expansion, the axis of the region is updated, allowing the template to follow the curvature of the helix. (C) Top: The predicted helix is described by the translation centers obtained in the bidirectional expansion. Middle: Comparison between the axes of the predicted helix and known helix. Bottom: The axis of the known helix is obtained by averaging four consecutive alpha carbons of the atomic structure. All molecular graphics in this paper were generated with Sculptor .[4]

higher probability, thereby promoting a better exploration of the search space.

The crossover operator enables the exchange of information between GA template individuals. New transformations are identified by swapping the translations and rotations of selected templates. We used a combination of crossover schemes where [x,y,z; r_i] were either swapped at one or multiple points, or modified using arithmetic operations. This crossover operator [29] not only affords efficient exploration of the cryo-EM map, but it is also particularly beneficial in the case of bundles of parallel helices where the orientation is conserved.

Initially, the cylindrical template population randomly samples the cryo-EM map outside of any tabu regions (Figure 1A, top left). This population is then allowed to evolve under selective pressure for several generations until an optimal solution is found (a solution is considered optimal when no further improvements are achieved over several generations).

Bidirectional expansion

The template with the optimal placement usually covers part of a helical region, aligned with its main axis (e.g., Figure 1A, top right), but it does not capture the full length of the helix or its curvature. Starting from this optimal placement and using the template's main axis as an indicative direction, a bidirectional expansion is performed to determine curvature and length of the helix.

The bidirectional expansion is performed in two steps, using an 8 Å-long cylinder with a radius of 1 Å. First, a local refinement of the translations and rotations is performed at the current placement of the template. In the second step, the template is translated in one and then the other direction along the axis of the optimal solution (Fig. 1A, bottom right). These two steps are iterated along the axis of the helical region until the score at the current position falls below a certain percentage of the initial score. By default, this limiting score threshold is set to 70% of the initial (highest) score computed within the current region. The iterative annotation is based on the assumption that the short template should maintain a rather constant score when moved inside a helical region. Therefore, as the template reaches the end of the region, the score decreases considerably and the expansion is stopped.

We note that the score threshold acts as an adjustable tolerance for deviations from the ideal rod shape due to experimental noise or reconstruction artifacts. In this work, we used a relatively stringent 80% threshold for idealized (simulated) maps and more permissive thresholds of 55-75% for experimental maps. The threshold levels were determined based on the observed performance of the algorithm (see the Results and Discussion sections).

Atomic structures of proteins exhibit both straight and bent helices. Therefore, HELEX

considers the general case in which helices may be curved. At each translation of the template, the orientation is subject to refinement, following the curvature of the region (see Fig. 1B). The translation center is stored as an axis point of the predicted helix (Fig. 1C top). This predicted axis closely follows the known axis of the atomic structure (Fig. 1C bottom), as is evident from a side-by-side comparison (Fig. 1C center). The parameterization of the cylinder length and radius was chosen so that a linear point density of $\sim$1.5 Å was achieved in order to approximately match the point spacing of the known axis.

Tabu search

Once a helical region is characterized by the bidirectional expansion, it is appended to the tabu list and eliminated from further exploration. A tabu region is defined about each translation center of the template, i.e., the axis of the predicted helix. The radius of each exclusion sphere is set to 6 Å to generate overlapping spheres for adjacent axis points, marking the entire length and width of the helix. During the evolution, the templates are not allowed to be placed within such tabu regions, i.e., their centers may not be closer than 6 Å to the points in the tabu list. This strategy prevents the algorithm from revisiting occupied regions, thereby promoting more efficient exploration of other helical regions in the map.

The previously described GA and bidirectional expansion steps identify one helical region at a time and therefore need to be iterated several times until all helical regions are identified. Each such iteration starts with a new random population of short cylindrical templates, while the tabu regions are preserved between iterations.

Stop criteria

The algorithm is iterated until a stop criterion is met. Without loss of generality, it can be assumed that the number of helices to be identified, denoted by N, is known either from prior structural data or sequence-based secondary structure prediction algorithms. For a given N, the algorithm will stop exploring the cryo-EM map when it has identified $3 \cdot N$ helices. More than N helices are investigated to allow for some imperfect ranking of the results, thereby yielding a better exploration of the search space. If the number of helices is not known a priori, the algorithm is stopped once the map has been extensively explored as assessed by a coverage rate, when it is impossible to place more templates into the map due to the tabu regions, or when only short helices are annotated for multiple consecutive iterations. In this case, more than $3 \cdot N$ helices may be identified.

Ranking of predicted helices

The outcome of the algorithm is a list of helices described by the translation centers of the template during the bidirectional expansion (Fig. 1C). To facilitate the exploration of the results, the list is sorted in a post-processing step using a correlation-weighted length

$$L_{Helix} = \frac{\left[\overline{CC_{Expansion}}\right]^2 \cdot \left[CC_{Interior}\right]^2}{\left[CC_{Exterior}\right]^2} \cdot Len\,, \tag{6}$$

where Len represents the length of the helix (computed as the sum of distances between adjacent points on the axis of the predicted helix). The squared correlations in the fraction

emphasize the scoring function relative to the length of the helix: $\overline{CC_{Expansion}}$ is the mean normalized cross-correlation measured by the short cylindrical template during the bidirectional expansion. $CC_{Interior}$ is the normalized cross-correlation of a 1 Å radius cylinder following the predicted axis. $CC_{Exterior}$ is the normalized cross-correlation of a 5 Å radius cylinder following the predicted axis. The heuristic L_{Helix} is high for long helices that have high density around the axis and low density at the exterior.

Validation

We designed a range of tests on simulated and experimental maps to assess the sensitivity and accuracy of the predictions afforded by HELEX. Experimental maps were selected from cases where closely matching atomic structures were available from X-ray crystallography. To be consistent with the resolution convention in earlier publications in this field, we created simulated maps by low-pass filtering of atomic structures using the *pdb2mrc* tool of the EMAN package [22] (resolution values using our *pdb2vol* (Situs) and Sculptor tools would have been smaller by a factor of 1.285; see [33]).

The N top-ranked solutions (where N represents the known number of helices) were selected for validation, and their performances were quantified using point-based and helix-based measures. Point-based measures compare the points on the predicted axes with points on the known axes (resulting from averaging coordinates of four consecutive alpha carbons in the crystal structure):

- The point sensitivity (pSe) is defined as the percentage of known points that were correctly predicted by HELEX.

- The point positive predictive value (pPPV) is defined as the percentage of predicted points that correspond to known axis points.

- The root mean square deviation (RMSD) of corresponding predicted and known axis points quantifies the separation of the axes.

For these point measures (pSe, pPPV, and RMSD), a predicted point is considered to match a known point if they are found within 4 Å of each other. This tolerance was set in order to accommodate experimental maps that show minor conformational differences from the crystal structure [34].

Helix-based measures directly compare the geometric properties of the predicted and known helices:

- The ΔTurns value is defined as the number of mismatched helical turns between two helices.

- The helix sensitivity (hSe) is the percentage of known helices detected among the top N predicted helices. For this measure, a known helix is considered to a true positive (TP), i.e., detected by HELEX, if a predicted helix at least partially overlaps and aligns with its axis. For example, a reported hSe of 70.6%, where N = 17, implies that 12 out of the 17 helices of the crystal structure were found among the top 17 predicted solutions. Typically, other known helices are identified as well, but they are ranked lower in the list of solutions and not considered for the hSe value.

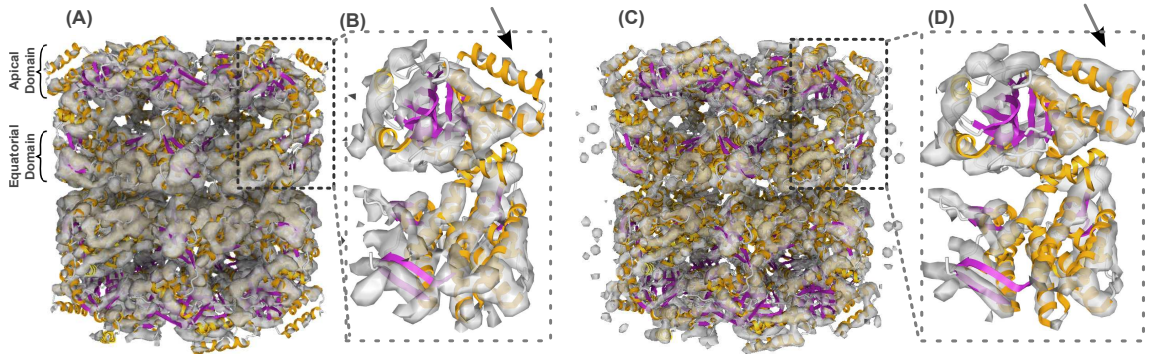

Figure 2: Gaussian-weighted local normalization applied to a 6 Å resolution experimental map of the chaperonin GroEL (EMD-1081, [23]). (A-B) The map shows higher density values in the equatorial than in the apical domain (isolevel 0.59744). (C-D) After Gaussian-weighted local normalization, the map depicts comparable density value across the map. Arrows indicate area of interest. The crystal structure of GroEL is shown as a reference in ribbon representation, with α-helices depicted in yellow.

Results

This section is organized as follows: first, the local normalization of densities is demonstrated using an experimental map of the chaperonin GroEL [23]; then, a typical helix extraction outcome is shown for an idealized (simulated) 10 Å resolution GroEL map [5]. To assess the performance of HELEX more systematically, we performed a series of tests using simulated maps and six experimental maps at variable resolutions.

Local normalization

Experimental cryo-EM volumes may suffer from uneven density distributions due to conformational disorder and alignment artifacts, with higher density exhibited at the core and a lower density at the surface. For example, a GroEL map solved at 6 Å-resolution [23] shows high densities in the equatorial domain, whereas the apical domain appears weaker (Fig. 2A,B). Therefore, to normalize the features across the map, the Gaussian-weighted

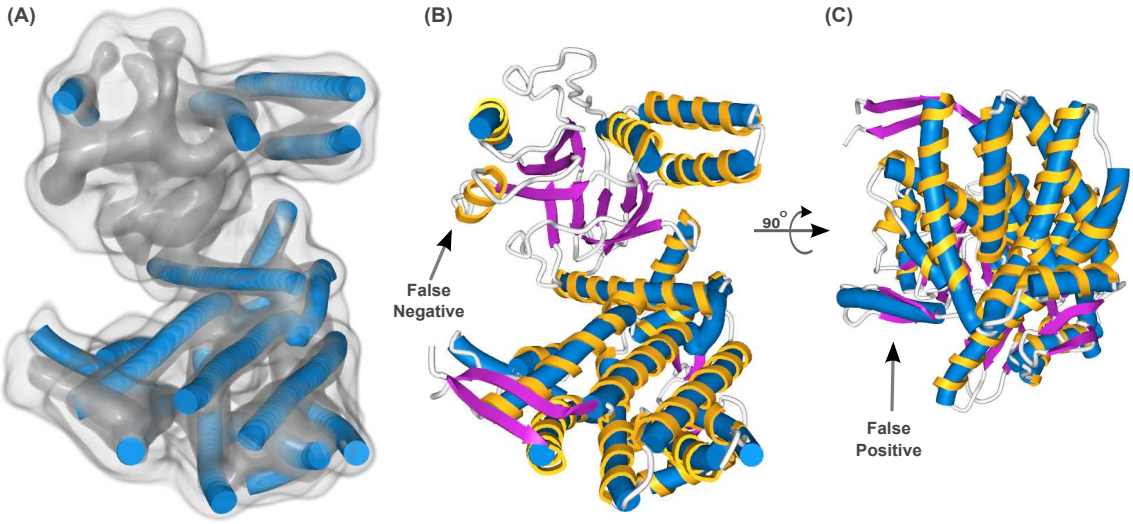

Figure 3: (A) A simulated map obtained by low-pass filtering a GroEL monomer to 10 Å resolution is presented along with the helices predicted by HELEX (represented as blue tubes). (B) Side and (C) bottom views of HELEX results (blue cylinders) overlapping the target crystal structure (α-helices represented as yellow ribbons).

local normalization was applied (see Methods). The resulting filtered volume (Fig. 2C,D) shows a more uniform distribution of density, bringing the equatorial and apical domains to a comparable level (indicated by arrows in Fig. 2B,D). This balanced distribution of filtered densities was conserved at different isosurface values, as observed by visual inspection in a molecular graphics program (data not shown).

Application example

To demonstrate a typical HELEX application, a GroEL monomer (from PDB ID: 1OEL, [5]) was low-pass filtered to 10 Å resolution with a voxel size of 2 Å (Fig. 3A), and the helix extraction was executed on a locally normalized map using an expansion threshold of 80% (see Methods). HELEX identified all 17 known helices of seven or more residues, placing 16 in the top 17 scoring solutions (Fig. 3B-C) and the remaining one at rank 18. The total run time for this example was 13.5 min (using a Quad-Core AMD Opteron processor 2360 SE at

2.5GHz).

The ranking of the results using the empirical L_{Helix} value (eq. 6) performed well in this example. In the case of GroEL, the N top results were divided into 16 true positives (the actual helices) and one false positive. False positives may be found in rod-like regions that do not correspond to alpha helices, for example (anti)parallel β-sheets (Fig. 3C). On the other hand, false negatives represent correct helices that are found lower in the ranking (such as rank 18 here), below the N top predictions. Our tests have shown that these false negative helices either are of smaller size ($\sim$7-10 residues) or may lack the characteristic rod-like shape.

To measure the performance of HELEX, we compute the hSe value, i.e., the percentage of true positive helices predicted by HELEX, to hSe = 94.1%. The agreement between the predicted and known axis points was characterized by the measures pSe = 98.2%, pPPV = 86.4%, and RMSD = 0.65 Å. As a control, we repeated the calculation in the absence of the local normalization, which was expected to perform less favorably (see Methods). The resulting performance measures without local normalization were hSe = 76.4%, pSe = 87.3%, pPPV = 66.3%, and RMSD = 1.1 Å.

Performance as a function of resolution

For a systematic benchmark of HELEX we generated simulated maps from monomer GroEL structures (17 long helices of seven or more residues; 57 kDa total molecular weight; [5]), succinate dehydrogenase (33 long helices; 118 kDa;[36]), and photosynthetic reaction center (34 long helices; 132 kDa;[3]). The α-helical secondary structure content ranged from 37% to 51% for these structures. Each structure was low-pass filtered to resolutions ranging from

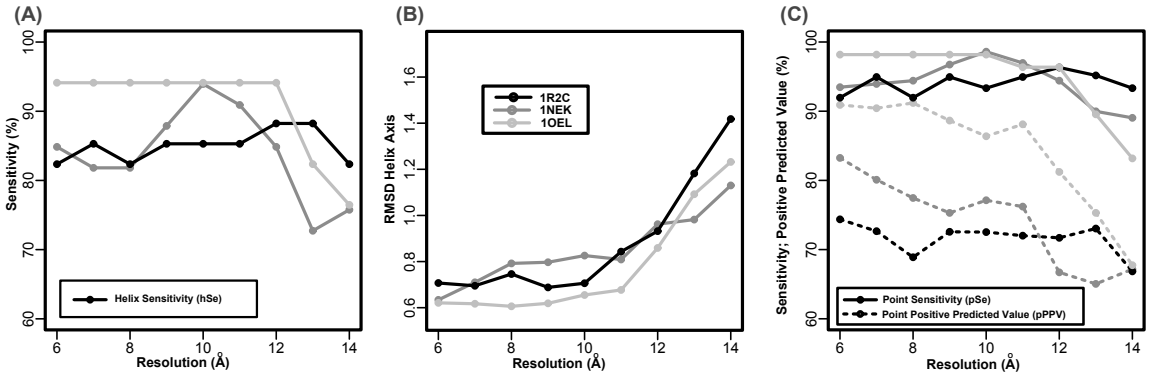

Figure 4: HELEX performance validation as a function of resolution for simulated maps of GroEL (PDB ID: 1OEL), succinate dehydrogenase (PDB ID: 1NEK), and photosynthetic reaction center (PDB ID: 1R2C). (A) Helix sensitivity hSe (see text). (B) RMSD of the helical axes. (C) The pSe and pPPV measures (see text) based on the axis points.

6 Å to 14 Å (2 Å voxel size), using *pdb2mrc* [22]. A local normalization and an expansion threshold of 80% were applied (see Methods). Figure 4 shows the helix and point-based performance measures hSe, pSe, pPPV, and RMSD as a function of the resolution.

Overall, HELEX detected α-helices reliably up to $\sim$12 Å resolution, beyond which the performance suffered from blurring of interior map detail. In all three cases the hSe was above 80% for resolutions in the 6-12 Å range (Fig. 4A). The geometric accuracy of the prediction was better than 1 Å in that same resolution range (Fig. 4B). Moreover, the pSe estimated on the axis points was better than 90%, indicating that any missed helices were short, as they accounted for only a small number of points (Fig. 4C). The pPPVs were systematically lower than the corresponding pSe values, indicating a HELEX bias to predict slightly longer helices than justified by the known structure. A more detailed inspection revealed that some axis points were predicted at the ends of known helices, where occasionally the backbone showed helix-like organization. We assumed that such minor false positive predictions would be preferable to false negatives. If desired, a user could reduce the resulting helix length by increasing the default score threshold in the bidirectional expansion (see Methods).

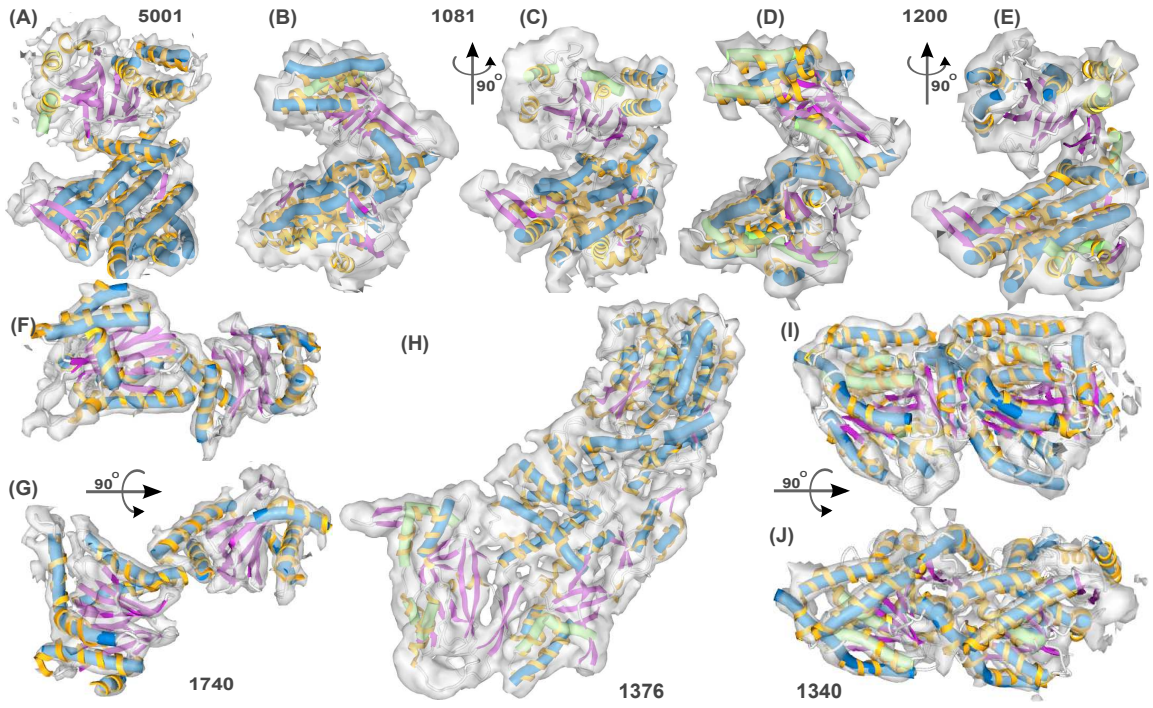

Figure 5: HELEX predictions for the experimental cryo-EM maps of GroEL (EMDB ID: 5001, 1081, 1200), 20S proteasome (EMDB ID: 1740), rice dwarf virus (EMDB ID: 1376), and kinesin (EMDB ID: 1340). The predictions are depicted using tube representations, blue for the helices predicted in the top N (column 'Helix Count' in Table 3) scoring solutions, and green for false negatives ranked lower in the list of solutions. Cryo-EM maps are shown in gray transparent surfaces, and the corresponding crystal structures (column 'PDB ID' in Table 3) use a yellow ribbon representation for the α-helices.

Experimental validation

The above tests were based on idealized maps in the absence of noise. To assess the performance of the algorithm on realistic cryo-EM maps in the presence of noise and 3D reconstruction artifacts, we chose six benchmark maps for which atomic structures were available as a control. Specifically, we used maps of GroEL at resolutions of 4.2 Å [21], 6.0 Å[23], and 7.8 Å[31], a map of the 20S proteasome at 6.8 Å resolution [27], a map of rice dwarf virus at 7.9 Å resolution[20], and a map of kinesin at 9 Å resolution[30]. Figure 5 presents an overview of the six benchmark systems.

Local normalization and expansion thresholds ranging from 55% to 75% were applied (see

Methods and Table 3). In some cases, the expansion thresholds were adjusted from the 70% default value to optimize the hSe measure, depending on the quality of a particular map. Similar to the simulated test cases, different performance measures were assessed, which are presented in Table 3. As in the simulated test cases, HELEX predicted helices that were slightly (∼1-2 turns) longer compared to those in the crystal structure.

Historically, GroEL was solved at increasing resolution by cryo-EM, whereas initial reconstructions at 11-13 Å showed only the overall shape of the chaperonin [28, 24], secondary structure elements were detected in intermediate resolution maps[23, 31], and a recent map at 4.2 Å resolution even afforded a trace of the protein backbone [21]. HELEX was executed here for a single monomer (extraction of the monomer was performed by masking the map using a docked atomic structure). Figure 5A-E shows the outcome of HELEX for the three investigated GroEL maps, depicting in blue the helices detected in the top N=17 solutions and in green the helices found lower in the list. Overall, the observed hSe values varied between 70.6% and 82.4% (Table 3). Similar values were also identified for pSe (68.2-86.3%). The axis RMSD values were below 2.02 Å.

The GroEL tests show that the accuracy of HELEX depends on the quality of a particular map. As expected, the performance was best for the 4.2 Å map (as judged by TP, hSe, pPPV, RMSD, and ΔTurns measures). However, it came as a surprise that the lowest (7.8 Å) resolution map surpassed the performance of the 6.0 Å map in pPPV, RMSD, and ΔTurns values. The challenging 6.0 Å map, despite its relatively high resolution, required a particularly low expansion threshold of 55% to accommodate visible variations in the rod

System (EMD ID)	Res/Vox (Å)	Helix Count	TP	hSe	pSe	pPPV	Δ Turns	RMSD (Å)	Total TP (Rank)
GroEL (5001)	4.2/1.06	17	14	82.4	86.3	84.42	0.95	1.07	16 (32)
GroEL (1081)	6.0/2.08	17	12	70.6	68.6	64.5	1.39	2.02	14 (31)
GroEL (1200)	7.8/2.28	17	12	70.6	68.2	70.4	1.15	1.70	17 (42)
Proteasome (1740)	6.8/1.38	11	11	100	100.0	91.8	0.46	1.12	11 (11)
Rice Dwarf Virus (1376)	7.9/1.49	33	26	78.8	85.5	68.6	1.67	1.14	33 (61)
Kinesin (1340)	9.0/2.00	23	19	82.6	79.2	65.2	1.49	1.30	23 (39)
Res - Resolution; Vox - Voxel size; Helix Count- total number of helices with 7 or more residues; TP - True Positive; hSe - Sensitivity on the helix count; pSe - sensitivity computed on the axis points; pPPV - positive predictive value computed on the axis points; Turns off - number of turns difference between the predicted helices and the actual helices.; Total TP - total number of actual helices predicted; Rank - at which rank was the last helix found where 0 represents the best scoring helix.									

Table 3: Results of HELEX for experimental systems

densities.

The other three systems in the experimental benchmark measured helix-based hSe values ranging from 78.8% to 100%, and axis point-based pSe values ranging from 79.2% to 100.0% (Table 3). HELEX performance was essentially perfect for the 20S proteasome at 6.8 Å resolution (Fig. 5F,G), where the algorithm predicted all 11 helices with an RMSD of 1.12 Å and ΔTurns of 0.46. In comparison, the slightly lower resolution rice dwarf virus and kinesin cases exhibited RMSD values of 1.14 Å and 1.30 Å, respectively, and ΔTurns values of 1.67 and 1.49, respectively, when compared to the crystal structure (Fig. 5H-J).

Discussion

HELEX combines a genetic algorithm, a bidirectional expansion, and a tabu search strategy to annotate helical regions in cryo-EM maps. The genetic algorithm performs a global search to identify fragments of helical regions, while the bidirectional expansion determines the curvature and length of the entire region. As the algorithm annotates the helices, they are placed in the tabu list to prevent them from being revisited later in the search.

Similar to earlier approaches by other groups, HELEX was designed under the assumption that α-helices can be identified as rod-like densities in cryo-EM maps. However, HELEX also considers the possibility of curvature. The bending of a helix is characterized in the bidirectional expansion by using a short cylindrical template that traces the rod-like feature.

HELEX reliably detected α-helices in simulated maps of 6-12 Å resolution and in experimental maps of 4-9 Å resolution, with a true positive accuracy ranging from 70.6% to 100%, as estimated in experimental settings. Although low resolution maps did fare worse, we did

not observe a clear correlation between performance and map resolution in experimental maps, as the results in the 4-7 Å resolution range depended on the particular experimental system and on the reconstruction quality.

False negatives often correspond to helices of short length or to those that fail to show the characteristic cylindrical rod shape. Such false negatives are often still detected but ranked lower in the solutions list. A visual inspection of the results may allow for the selection of such helices according to prior knowledge of the system. To reduce the risk of such false negatives, a user could lower the score threshold in the bidirectional expansion. The default value of the expansion threshold parameter was set to 70% for our experimental test cases, but it may be lowered to 50-60% for challenging maps that exhibit noisy helical features. Decreasing the value of the parameter entails a reduction in the pPPV and an increase in the length of predicted helices, so there is a tradeoff between tolerance and helical length. In our experimental test cases we were able to optimize the threshold based on the observed hSe values using known atomic structures. If crystal structures are unavailable, docking models may be substituted, or the user may use sequence-based secondary structure prediction[15, 8, 26, 25, 16, 7] as a control.

Although rod-like densities typically correspond to α-helices, other structural elements may display similar patterns and generate false positives. Examples include the (anti)parallel β-sheet, which exhibits a smaller rod radius than the helices. Inspection of HELEX results would permit the manual removal of such false positives. Moreover, several α-helices may occasionally be found in sequence, which become blended into a long cylinder without clear ends between distinct sub-helices. Such situations require additional information regarding the α-helix composition of the system, such as sequence-based secondary structure prediction,

in order to identify the ends of any sub-helices. Sequence-based prediction methods may also be helpful in situations where the number of helices, N, is not known a priori.

A significant number of new parameters were introduced for the algorithmic components of HELEX. These parameters were derived based on geometric considerations and were tested empirically, as is typical for a proof-of-concept paper. We found that the performance of HELEX was robust for the systems under investigation, but a more systematic refinement of the parameter values could be performed in future research. Based on our experience, these parameters should not require any fine-tuning by the user, except for the above-mentioned expansion threshold that controls the tolerance to density variations.

HELEX incorporates parallel computing strategies to take advantage of the multi-core architecture of current workstations. Multiple genetic algorithm runs are launched in parallel. Although independent of each other, the parallel threads use a common tabu list. Each parallel run is finalized by a bidirectional expansion and by a global update of the tabu list before being re-launched until one of the stop criteria is met. Due to the multi-threaded approach, a typical HELEX run takes only minutes on a modern workstation. The run time may be decreased by reducing the sampling in the genetic algorithm or in the bidirectional expansion. Our default parameters currently favor sampling over efficiency to ensure a near-continuous exploration of the map. Specifically, we employ an angular step of 1° in contrast to earlier template convolution algorithms that employ larger angular steps up to 15°.

During iterative optimization, HELEX often determines high-ranking helices first. Such characteristics prompted us to integrate HELEX into the interactive modeling software Sculptor [4]. The user can investigate the results on the fly as they are generated and stop the execution early if desired. HELEX was included in Sculptor version 2.1, available

at http://sculptor.biomachina.org.

Acknowledgments

We thank Stefan Birmanns, Zbigniew Starosolski, and Manuel Wahle for helpful discussions and for reading the manuscript. This work was supported in part by a grant from National Institutes of Health (R01GM62968).

References

[1] M. L. Baker, T. Ju, and W. Chiu. Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, 15(1):7–19, 2007.

[2] W. Baumeister and A. C. Steven. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.*, 25:624–631, 2000.

[3] R. H. G. Baxter, N. Ponomarenko, V. Šrajer, R. Pahl, K. Moffat, and J. R. Norris. Time-resolved crystallographic studies of light-induced structural changes in the photosynthetic reaction center. *Proc. Natl. Acad. Sci. USA*, 101(16):5982–5987, 2004.

[4] S. Birmanns, M. Rusu, and W. Wriggers. Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. *J. Struct. Biol.*, 173(3):428–35, 2011.

[5] K. Braig, P. D. Adams, and A. T. Brünger. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. *Nature Struct. Biol.*, 2:1083–1094, 1995.

[6] P. Chacón and W. Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.*, 317:375–384, 2002.

[7] J.-M. Chandonia and M. Karplus. New methods for accurate prediction of protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*, 35(3):293–306, 1999.

[8] C. Cole, J.D. Barber, and G.J. Barton. The Jpred 3 secondary structure prediction server. *Nucl. Acids Res.*, 36(suppl 2):W197–W201, 2008.

[9] Y. Cong and S. J. Ludtke. Single Particle Analysis at High Resolution. *Methods in Enzymology*, 482:211–235, 2010.

[10] A. Dal Palù, J. He, E. Pontelli, and Y. Lu. Identification of alpha-helices from low resolution protein density maps. *Computational Systems Bioinformatics Conference*, pages 89–98, 2006.

[11] J. Frank. Single-particle imaging of macromolecules by cryo-electron microscopy. *Ann. Rev. Biophys. Biomol. Struct.*, 31:303–319, 2002.

[12] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[13] J. H. Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI, 1975.

[14] W. Jiang, M. L. Baker, S. J. Ludtke, and W. Chiu. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.*, 308:1033–1044, 2001.

[15] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202, 1999.

[16] K. Karplus, K. Sjölander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm, and C. Sander. Predicting protein structure using hidden Markov models. *Proteins: Structure, Function, and Bioinformatics*, 29(s 1):134–139, 1997.

[17] J. A. Kovacs, M. Yeager, and R. Abagyan. Computational prediction of atomic structures of helical membrane proteins aided by EM maps. *Biophys. J.*, 93(6):1950–1959, 2007.

[18] K. Lasker, O. Dror, R. Nussinov, and H. Wolfson. Discovery of protein substructures in EM maps. *Algorithms in Bioinformatics*, pages 423–434, 2005.

[19] S. Lindert, R. Staritzbichler, N. Wötzel, M. Karakaş, P. L. Stewart, and J. Meiler. EM-Fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure*, 17(7):990–1003, 2009.

[20] X. Liu, W. Jiang, J. Jakana, and W. Chiu. Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-Path Simulated Annealing optimization algorithm. *J. Struct. Biol.*, 160(1):11–27, 2007.

[21] S. J. Ludtke, M. L. Baker, D.-H. Chen, J.-L. Song, D. T. Chuang, and W. Chiu. De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure*, 16(3):441–448, 2008.

[22] S. J. Ludtke, P. R. Baldwin, and W. Chiu. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.*, 128(1):82–97, 1999.

[23] S. J. Ludtke, D. H. Chen, J. L. Song, D. T. Chuang, and W. Chiu. Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure*, 12:1129–1136, 2004.

[24] S. J. Ludtke, J. Jakana, J. L. Song, D. T. Chuang, and W. Chiu. A 11.5 Å single particle reconstruction of GroEL using EMAN. *J. Mol. Biol.*, 314:253–262, 2001.

[25] J. Meiler and D. Baker. Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci. USA*, 100(21):12105–12110, 2003.

[26] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.*, 7(9):360–369, 2001.

[27] J. Rabl, D. M. Smith, Y. Yu, S.-C. Chang, A. L. Goldberg, and Y. Cheng. Mechanism of gate opening in the 20S proteasome by the proteasomal ATPases. *Mol. Cell*, 30(3):360–368, 2008.

[28] N. A. Ranson, G. W. Farr, A. M. Roseman, B. Gowen, W. A. Fenton, A. L. Horwich, and H. R. Saibil. ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell*, 107:869–879, 2001.

[29] M. Rusu and S. Birmanns. Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions. *JSB*, 170:164–171, 2010.

[30] C. V. Sindelar and K. H. Downing. The beginning of kinesin's force-generating cycle visualized at 9-Å resolution. *J. Cell Biol.*, 177(3):377–385, 2007.

[31] S. M. Stagg, G. C. Lander, J. Pulokas, D. Fellmann, A. Cheng, J. D. Quispe, S. P. Mallick, R. M. Avila, B. Carragher, and C. S. Potter. Automated cryoEM data acquisition and analysis of 284 742 particles of GroEL. *J. Struct. Biol.*, 155(3):470–481, 2006.

[32] M. Tagari, R. Newman, M. Chagoyen, J. M. Carazo, and K. Henrick. New electron microscopy database and deposition system. *Trends Biochem. Sci.*, 27:589, 2002.

[33] W. Wriggers. Conventions and work flows for using Situs. *Acta Cryst. D.*, 2012. Invited review, in preparation.

[34] W. Wriggers, R. K. Agrawal, D. L. Drew, A. McCammon, and J. Frank. Domain motions of EF-G bound to the 70S ribosome: Insights from a hand-shaking between multi-resolution structures. *Biophys. J.*, 79:1670–1678, 2000.

[35] Y. Wu, M. Chen, M. Lu, Q. Wang, and J. Ma. Determining protein topology from skeletons of secondary structures. *J. Mol. Biol.*, 350(3):571–586, 2005.

[36] V. Yankovskaya, R. Horsefield, S. Törnroth, C. Luna-Chavez, H. Miyoshi, C. Léger, B. Byrne, G. Cecchini, and S. Iwata. Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science*, 299(5607):700–704, Jan 2003.

Research Project Summary and Future Directions

Solving the structure of biomolecular systems at high-resolution is often a difficult, if not impossible, task. Even when structural data is available, which is in itself a difficult and time consuming procedure, the level of details or the data complexity may prevent the building of an atomic model for the biomolecular system. Modeling techniques have thus been introduced to enable the interpretation of the data and the building of atomic models.

The three manuscripts presented as part of this dissertation are concerned with such modeling tasks that enable the interpretation of cryo-EM reconstructions. The computational methods introduced here take as input the cryo-EM map, and additional information, when available, and either provide information or directly generate an atomic model. Although experimental validation is required to confirm the models and the predictions derived from them, one can still analyze the molecular architecture to identify important sites, e.g. interaction regions with/between different components, or binding pockets for targeted therapeutics. Such pieces of information can then be used to plan further experiments or design the validation procedure.

The first manuscript describes a heuristics to generate the atomic model of the Rift Valley fever virus. A classic example of an icosahedral virus, the Rift Valley fever virion has its genetic material protected by an envelope composed of two different types of glycoproteins. The approach presented in this first manuscript is concerned with identifying

the spatial organization of these glycoproteins relative to each other inside the envelope. The proposed model, although not yet experimentally validated, is supported by existent molecular data regarding fusion peptide sites (involved in the attachment to the host cell) or epitopes (responsible for the interaction with antibodies). The Rift Valley fever virus is a typical example of multi-resolution modeling task that involves the integration of structural data from various experimental techniques. Due to their health and economical impact, viruses are typically studied at structural level in order to generate vaccines or identify anti-viral drugs. However their large dimension and pleiomorphic nature typically prevent their structural elucidation at high level of detail. In fact, cryo-EM is often applied to solve the low-resolution structure of the entire virion. Then, such data is interpreted relative to the high-resolution structure of the component proteins. The manuscript describes such a multi-resolution modeling approach, applied here for the Rift Valley fever virus, yet that may be employed by experimental biologists to identify the atomic model of other viruses of interest.

The second manuscript introduces an evolutionary tabu search strategy for the simultaneous registration of multiple atomic structures into the low-resolution envelope of their assembly. If in the first manuscript a multi-body refinement was applied, in the second, a multi-body global registration technique is presented. The former starts with an initial position and refines the placements and orientation of the components, while the latter seeks both to place and refine these positions. The two techniques consider the atomic structures of multiple components, simultaneously, in a high-dimensional optimization. Such approaches are beneficial at low-resolutions, where the boundaries between the different component are not easily discerned and the scoring functions are not discriminative enough to place the different component independently. By considering all fragments simultaneously, additional

spacial constrains are introduced indirectly, enabling the proper placement of the components relative to each other inside the assembly.

In this second manuscript, only the rigid-body transformations, translations and rotations, are optimized. However, the biomolecular systems are dynamic and the interactions between components often entail flexible deformations. As future direction, such conformational variability will be considered as a parameter in the optimization. In fact, structural changes can be identified using statistical approaches such as the principal component analysis that will allow the characterization of the deformation space. A list of discrete configurations may thus be built, and an index in this list can be considered as a parameter to be optimized by the genetic algorithm.

The final manuscript applies the evolutionary tabu search strategies in combination with the bidirectional expansion to annotate alpha helices in cryo-EM reconstructions. The approach successfully detected 70-100% helical regions in intermediate-resolution cryo-EM reconstructions as assessed in experimental systems. The sensitivity of the technique was quantified for the automatic detection, yet a manual investigation of all the results may enable higher detection rates. In the future, the extracted helical regions will be used for further processing to permit the *de novo* identification of atomic models in intermediate cryo-EM reconstructions. The approach described in this third manuscript predicts the helical regions and provides, as output, their axes. Such axes will be used to place a short helix template and thus generate an atomic model of the entire helical region.

The computational techniques presented here cover a wide range of problems that arise at different resolutions in cryo-EM. This work represents a starting point for further research that will consider more complex models to better capture the problem. Although applied here

on cryo-EM data, such techniques may be employed on data collected from other biophysical sources, e.g. cryo-electron tomography. In fact, the latter method was already tested and will be further investigated on tomographic reconstructions to annotate and characterize actin filaments in filopodia.

The goal of this dissertation is not only to introduce automatic computational techniques to solve various multi-resolution modeling problems, but also to render these approaches available to the large scientific community in a user friendly environment. Each of the techniques presented here is available in the molecular modeling Sculptor.

The field of structural biology is evolving at high speeds towards generating increasing amounts of high-resolution data regarding the molecular architecture of biological systems. Improvements in instrumentation, experimental and computational methodology have allowed the solving of even larger or more dynamic systems. Often, such efforts are made possible by integrating the structural data obtained from the different biophysical sources. The modeling approaches described in this dissertation are introduced to facilitate the integration and interpretation of low- and intermediate-resolution data. They represent computational tools to be used by structural biologist to enable the high-resolution interpretation of cryo-EM reconstructions.

Mirabela Rusu, M.Eng., M.S.

July 2011