Published in final edited form as:

Methods Mol Biol. 2013 ; 955: 229–241. doi:10.1007/978-1-62703-176-9_13.

Modeling, Docking, and Fitting of Atomic Structures to 3D Maps from Cryo-electron Microscopy

Gregory S. Allen¹ and David L. Stokes^{1,2,3}

¹Skirball Institute of Biomolecular Medicine and Dept. of Cell Biology, New York University School of Medicine, New York, NY 10012

²New York Structural Biology Center, New York, NY, 10027

Abstract

Electron microscopy (EM) and image analysis offer an effective approach for determining the three-dimensional structure of macromolecular complexes. The versatility of these methods means that molecular species not normally amenable to other structural methods, e.g., X-ray crystallography and NMR spectroscopy, can be analyzed. However, the resolution of EM structures is often too low to provide an atomic model directly by chain tracing. Instead, a combination of modeling and fitting can be an effective way to analyze the EM structure at an atomic level, thus allowing localization of subunits or evaluation of conformational changes. Here we describe the steps involved in this process: building a homology model, fitting this model to an EM map, and using computational methods for docking of additional domains to the model. As an example, we illustrate the methods using an integral membrane protein, CopA, which functions to pump copper across the membrane in an ATP-dependent manner. In this example, we build a homology model based on the published atomic coordinates for a related calcium pump from sarcoplasmic reticulum (SERCA). After fitting this homology model to a 17 Å resolution EM map, computational software is used to dock a metal binding domain that is unique to the copper pump. Although this software identifies a number of plausible interfaces for docking, the constraints of the EM map steers us to select a unique solution. Thus, the synergy of these two methods allows us to describe both the location of the unknown metal binding domain relative to the other cytoplasmic domains and also the atomic details of the domain interface.

Keywords

Electron microscopy; structure modeling; protein-protein docking; P-type ATPases; computational biology

1. Introduction

Electron microscopy can be used to generate 3D structures of proteins using several different reconstruction strategies: electron crystallography, helical reconstruction, single particle averaging, or tomography. Due to the limited resolution (8–30 Å), however, it is often difficult to directly evaluate the conformation of the polypeptide chain. This limitation can be overcome by building a model based on related X-ray crystallographic or NMR structures and then fitting this model into the lower resolution EM map. Such a model is frequently useful for evaluating conformational changes due to different conditions for EM

³Corresponding author. stokes@nyu.edu.

sample preparation or for localizing accessory domains or subunits that were not present in the X-ray or NMR structure.

As an example of this modeling procedure, we have used an ATP-dependent copper pump from *A. fulgidus* called CopA. CopA belongs to the large family of P-type ATPases that couple the energy of ATP hydrolysis to the transport of ions across the membrane, thus generating ion gradients that are essential for the homeostasis of cells. X-ray crystallographic structures exist for Ca^{2+} -, Na^+/K^+ - and H^+ -ATPases from this family. However, CopA inhabits the P1b subclass of P-type ATPases that contains large insertions and deletions with respect to these existing structures. Of particular interest are the metal binding domains (MBD) on the N- and C-termini of CopA, which are homologous to a large family of soluble metal binding proteins. These MBDs are connected to the main body of CopA by flexible linkers, which allow the MBDs to interact with the other cytoplasmic domains responsible for binding and hydrolysis of ATP (Fig. 1). Previous work with CopA suggested that the N terminal MBD was involved in protein-protein interactions with one or more of the cytoplasmic domains (1).

For our modeling studies, we used X-ray crystallographic structures of related P-type ATPases, related metal binding proteins, and of isolated cytoplasmic domains of CopA. The shape of this model was constrained by a 12-Å resolution map of CopA that was determined by helical reconstruction of tubular crystals that were imaged by cryo-electron microscopy. We will describe our efforts first to model the structure of CopA with truncated N- and C-termini ($\Delta\text{N}\Delta\text{C}$ -CopA) and then to fit it to our EM map (1). Additionally, we will describe docking of an MBD to this $\Delta\text{N}\Delta\text{C}$ -CopA model in order to identify the orientation of the bound MBD that is consistent with our EM map and which provides the lowest energy domain interface.

2. Materials

2.1 EM maps

The EM maps were obtained by helical reconstruction of tubular crystals of CopA, as described elsewhere in this book and by Wu et al. (1). Alternative reconstruction strategies are possible depending on the nature of the sample (e.g., crystalline or a homogeneous preparation of isolated macromolecules). Depending on the software suite used for reconstruction, the user will obtain maps in a variety of different formats (e.g., SPIDER (2) or MRC (3)). These formats may be interconverted using em2em, which is a free component of the IMAGIC software suite (4). We have used the MRC format throughout and also encourage this practice.

2.2 Crystal/NMR structures

Crystallographic or NMR structures of relevant proteins or domains may be obtained from the Protein Data Bank (<http://www.rcsb.org>), which is maintained by the Research Consortium for Structural Bioinformatics (RCSB). Coordinates are generally downloaded in PDB format.

2.3 Sequence analysis software

There is a plethora of sequence analysis programs available. Most of these programs will be adequate for this application since we are using structures having high homology with our target. The following is a small sampling of available programs. For multiple alignments: ICM (5), ClustalW (6), MULTALIN; for pairwise alignments: PyMol (Schrödinger, LLC).

2.4 Modeling software

The following programs can automatically produce a structural model using a known X-ray crystallographic or NMR structure and a sequence alignment of the target molecule with this template structure: ICM, which is commercially available from Molsoft (5), Modeller or Modweb, which is available either as a web application or for download to a local workstation (7).

2.5 Visualization software

In order to display the EM maps as well as the models as they are developed, we have used Chimera (8), but PyMol (Schrödinger, LLC), Coot (9), or O (10) may also be used for this purpose.

2.6 Docking software

There are many docking programs available, but the following performed well in an analysis by Cross (11): ICM (5), GLIDE (12), Surflex (13), AutoDock (14), and UCSF DOCK (15). We used ICM, because we could perform all of the necessary tasks within this single integrated software suite.

2.7 Computer workstations

Modelling and docking were carried out on SGI-Linux workstations running 4 Intel Xeon 5160 cpus (3.0GHz) with 2 Gbyte RAM. Using this hardware, ICM modelling of CopA (664 residues) took ~2 hours; ICM docking of NMBD to cytoplasmic domains of CopA took ~12 hours.

3. Methods

The methods described below illustrate how to build a protein model into a density map obtained by electron microscopy, how to validate the models obtained, and how to use the models in a protein-protein docking experiment. We assume that at least one structure exists with significant homology to the target protein. The steps involved in this process are as follows. First an appropriate template structure must be chosen as a basis for building a model for the target protein. Next, the sequences of the template structure and the target protein must be aligned. A structural model for the target protein can then be built based on this sequence alignment. This model should be carefully validated based on common sense and with regard to existing data. Most importantly, the model should fit the EM density as closely as possible and we describe several means to optimize this fit. Finally, this model can be used to explore its interaction with known binding partners *in silico* by performing a protein-protein docking experiment.

3.1 Choose structure as a template for constructing a homology model

1. The first step is to use the target protein sequence to search for homologs using Psi-Blast (blastp) on Pubmed with default parameters (16). From the resulting set of potential homologs, focus only on those that have existing structures, because a homologous structure from the Protein Database (PDB) can be used as a template for building a model of the protein of interest. If structures of the homologous protein exist in multiple conformations, as was the case for Ca²⁺-ATPase from sarcoplasmic reticulum (SERCA), it is necessary to determine the most appropriate conformation for use as a homology model. Criteria for this choice include the conformation that is most consistent with the shape of the EM map and/or the conformation that is most consistent with the biochemical conditions used for crystallization (Fig. 2).

2. For building a model of CopA, we chose the E2 conformation of SERCA, which is represented by the PDB entry 1IWO (17). This conformation is induced by chelation of the primary transport ion (EGTA for chelating Ca^{2+}) and is indeed consistent with the conditions used for crystallization of CopA (BCDS for chelating Cu^{+}). The resolution of the SERCA structure was acceptable (3.1 Å) and the overall shape of this structure matched our EM map better than other conformations of SERCA in the PDB (*see* Note 1).

3.2 Align protein sequences of template structure with target protein

1. Obtain the protein sequences of the homologous template protein and the target protein by searching a public database such as UniProt (13).
2. Feed sequences into an alignment program such as Clustal, GCG, and ICM (*see* Note 2). It is beyond the scope of this chapter to discuss ideal alignment parameters for the various programs. We generally accepted the default parameters and were guided by the positive alignment of the known signature sequences of P-type ATPases. In difficult cases the related publications for the software should be consulted.

3.3 Model building

1. To make a homology model, an automated building routine can be used within ICM directly after sequence alignment. Under the "Homology" tab, select "Build Model". The resulting model will have the sequence of the target protein that has been folded into a three-dimensional protein based on the structure of the template molecule. ICM then allows the user to choose among several possible conformations of variable loops between secondary structure elements (*see* Note 3). Alternatively, the Modeller web server (MODWEB) only requires that the user paste in the target protein sequence, after which the alignment and model building is performed non-interactively.
2. Inspect the homology model (Fig. 3). After automatic model building, the researcher must verify that the new model makes sense and that it also fits the target map. To do this, view the homology model using visualization software (e.g. Chimera) and specifically evaluate regions of the molecule that have low homology to the template molecule (*see* Notes 4 and ⁵).

¹When choosing between equivalent structures to use as a template, the highest resolution structure is preferable. Higher resolution structures are, by definition, better determined mathematically and thus more reliable as a template. α -helices are recognizable in x-ray maps with resolutions better than 5.5 Å resolution while β -sheets require <3.0 Å resolution (25). Thus, if the target contains significant amounts of β -sheet, the template model should have a resolution better than 3.0 Å to be confident of correct placement of modeled β -sheets; an all α -helical model could be based on a homologous structure of lower resolution. Nevertheless, higher resolution is valuable for providing information about the side chain configurations, which can prove valuable for docking experiments.

²In an alignment of the CopA and SERCA sequences, it was immediately apparent that CopA contains many common features but that there were discrepancies in the size of cytoplasmic domains and the numbers of transmembrane helices. These discrepancies arise from the fact that CopA belongs to a distinct subclass of P-type ATPases and indicate areas of potential problems in model building.

³After model building with ICM the user is given the possibility to choose among several different loop conformations for many loops throughout the model. Unless the EM density map has sufficient resolution to discriminate between the alternatives, we generally accept the default.

⁴In our case we had the opportunity to compare our new model with crystal structures of the three isolated cytoplasmic domains of CopA (2HC8 and 3A1D). This comparison indicated that the model building routine of ICM failed to make a reasonable N-domain for CopA, presumably due to the numerous deletions relative to SERCA. Therefore we opted to replace this modeled N-domain with the crystal structure 2HC8. Another possible approach would be to align and build each domain separately.

⁵When evaluating a model built by software it can be helpful to have a secondary structure prediction of the target sequence. Discrepancies between the new model and the prediction in either secondary structure or residue numbering should be carefully analyzed to determine which is likely correct.

3.4. Fitting the model to the EM map

1. Manual fitting of the homology model to the EM density map. Manual fitting is a straightforward procedure that involves using the visualization software to move the model as a rigid body until it resides within the envelope defined by the EM density map. This process generally involves selecting the model through the graphical user interface and then using the computer mouse to manipulate its position relative to the density map. It will also be necessary to rotate the view in order to evaluate the fit from different angles. Consult the software documentation for specific instructions for these processes. Shape complementarity is the main criterion that guides the placement of the model into the map, though properties of the macromolecule may also guide this process (e.g., the membrane domain should be placed within the corresponding region of the map).
2. Optimized manual fitting. Once a reasonable overall fit is achieved, the homology model can be broken into individual domains and each domain can be moved independently in order to best match the density map. Prior knowledge of domain boundaries in the map and flexible joints in the structure are helpful in this regard. One strategy for separating the domains is to edit the pdb file for the model and to save the coordinates for each individual domain into a separate file. Some software programs require appropriate header information within each individual pdb files (including secondary structure definitions), whereas other programs will work with a raw listing of the coordinates. Another approach is to select each domain through the graphical user interface and to move them independently using the computer mouse. Consult the software manual for up-to-date information about both of these procedures. With higher resolution maps (<8 Å resolution) secondary structure elements (e.g., α -helices) become apparent in the density map and provide a strong constraint for manipulating the homology model. The overall goal is to fill the EM density with the model both while minimizing protruding, poorly fitted regions and maintaining the topology of domain connections. This is necessarily a subjective process, but with some practice, good results may be obtained.
3. Automated fitting of models to density maps is provided by a few different software packages (e.g. Situs (18), Chimera, Modeller, Sculptor(19)). The procedures are generally more complicated and, depending on the nature of the project may not necessarily lead to great improvements. In particular, although these procedures may be successful for rigid-body docking, the introduction of conformational changes in the homology model during fitting is an especially difficult problem (20; 21). Nevertheless, these programs generally quantify the goodness of fit, e.g. with a cross-correlation coefficient, which provides a useful parameter for comparing several slightly different fits.
4. A good compromise between manual and fully automated fitting is provided by the 'Fit in Map' function in Chimera, which can be used to facilitate manual fitting. This function transforms the model, or some portion of it, into a map with a specified resolution. This model map is fit to the EM map using cross-correlation and quantifies the fit with the corresponding cross-correlation coefficient. Using 'Fit in Map' at various resolutions together during the process of manual fitting can help find a global minimum.
5. Once the optimal location of the model within the EM map has been established, CNS (Crystallography and NMR System) may be used to fine tune the model fitting by imposing geometry and symmetry restraints during simulated annealing (22). In the "Utilities" section of CNS, an input script called "em_map_to_hkl.inp" can be used to transform an EM map into a reflection file used in all

crystallographic refinement routines. Then, for example, symmetry could be imposed during the refinement in CNS. With a 17 Å map of CopA, we found that CNS only marginally improved the manual fitting of our homology model. At higher resolution, one could use rigid body fitting of the various domains to improve their fits.

6. Iterate these procedures. Analysis and refinement are necessarily iterative, especially if CNS is used. The researcher should monitor the outcome of each round to determine if changes are appropriate or if the results have converged to a stable answer. This process necessarily entails visual inspection of the model and evaluating its fit to the EM map.

3.5. Docking of additional components to the homology model

1. Obtain model for the element to be docked. For our work, we built a model for the N-terminal metal binding domain of CopA (NMBD), which has many homologous structures in the Protein Data Bank. We chose a copper metallochaperone (2QIF), which has an X-ray crystallographic structure with high resolution and which has high homology to NMBD. We then followed the procedure outlined in sections 3.2 and 3.3 to produce a homology model of the NMBD of CopA.
2. Docking of the two elements using ICM. In the protein-protein docking module of ICM, select one element as the receptor ($\Delta N\Delta C$ -CopA in our case) and the other element as the ligand (NMBD). Next select a region of the receptor for docking of the ligand. Based on this selection ICM calculates various potentials (electrostatic, hydrophobic, etc.) on a grid around the receptor (23). Solvent is accounted for by calculating its shielding effect on the solvent-exposed charges and including this in the grid potentials (24). The size of the grid may be varied, but the defaults generally produce satisfactory results. The output from our docking experiments is shown in Fig. 4. The table lists a number of alternative docking solutions sorted according to their energy (see Note 6). In our case, we found that the lowest energy solution gave a plausible location of NMBD that was consistent with extra density within the EM map.
3. Check validity of docking result relative to the EM map. Because the EM map was not considered during the docking experiment, the correspondence between the extra EM density and the location of the NMBD provides a powerful validation of the result (Fig. 5). Thus, the combination of docking and fitting act synergistically to cross-validate the model and to suggest not only a location, but also a specific binding interface that neither technique could independently produce.

References

1. Wu CC, Rice WJ, Stokes DL. Structure of a copper pump suggests a regulatory role for its metal-binding domain. *Structure*. 2008; 16:976–985. [PubMed: 18547529]
2. Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* 1996; 116:190–199. [PubMed: 8742743]
3. Crowther RA, Henderson R, Smith JM. MRC image processing programs. *J. Struct. Biol.* 1996; 116:9–16. [PubMed: 8742717]

⁶We used default settings for all our protein-protein docking experiments. The top solution or lowest energy pose for the ligand was reproducibly located in EM density that we had previously assigned to NMBD. This energy is nominally binding energy, but does not correspond to the real binding energy of the interaction due to the various assumptions and approximations made for the calculation. Rather, these energies only provide a measure of the relative strength of the various binding poses.

4. van Heel M, Harauz G, Orlova EV, Schmidt R, Schatz M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* 1996; 116:17–24. [PubMed: 8742718]
5. Cardozo T, Totrov M, Abagyan R. Homology modeling by the ICM method. *Proteins.* 1995; 23:403–414. [PubMed: 8710833]
6. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics.* 2002; Chapter 2(Unit 2 3)
7. Sanchez R, Sali A. Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods Mol Biol.* 2000; 143:97–129. [PubMed: 11084904]
8. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* 2004; 25:1605–1612. [PubMed: 15264254]
9. Emsley P, Lohkamp B, Scott W, Cowtan K. Features and development of Coot. *Acta Crystallographica Section D - Biological Crystallography.* 2010; 66:486–501.
10. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica A.* 1991; 47:110–119.
11. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J Chem Inf Model.* 2009; 49:1455–1474. [PubMed: 19476350]
12. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem.* 2004; 47:1750–1759. [PubMed: 15027866]
13. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem.* 2003; 46:499–511. [PubMed: 12570372]
14. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit.* 1996; 9:1–5. [PubMed: 8723313]
15. Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, Rizzo RC. Development and validation of a modular, extensible docking program: DOCK 5. *J Comput Aided Mol Des.* 2006; 20:601–619. [PubMed: 17149653]
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
17. Toyoshima C, Nomura H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature.* 2002; 418:605–611. [PubMed: 12167852]
18. Wriggers W, Milligan RA, McCammon JA. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* 1999; 125:185–195. [PubMed: 10222274]
19. Rusu M, Birmanns S. Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions. *J. Struct. Biol.* 2010; 170:164–171. [PubMed: 20056148]
20. Trabuco LG, Villa E, Mitra K, Frank J, Schulten K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure.* 2008; 16:673–683. [PubMed: 18462672]
21. Hinsén K, Reuter N, Navaza J, Stokes DL, Lacapere JJ. Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.* 2005; 88:818–827. [PubMed: 15542555]
22. Brunger AT. Version 1.2 of the Crystallography and NMR system. *Nat Protoc.* 2007; 2:2728–2733. [PubMed: 18007608]
23. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. *Protein Sci.* 2002; 11:280–291. [PubMed: 11790838]
24. Fernandez-Recio J, Totrov M, Abagyan R. Screened charge electrostatic model in protein-protein docking simulations. *Pac Symp Biocomput.* 2002:552–563. [PubMed: 11928507]
25. McRee, DE. *Practical Protein Crystallography.* Academic Press; San Diego: 1999.

Figure 1.

Topology of the proteins used for modeling. **(A)** The Ca^{2+} -ATPase from sarcoplasmic reticulum (SERCA) provided the template for homology modeling. SERCA consists of ten transmembrane domains and two large cytoplasmic loops. In the 3D structure of SERCA, these cytoplasmic loops form three domains, which are denoted A-, N-, and P-domains. **(B)**. The Cu-ATPase from *Archeaoglobus fulgidis* (CopA) has eight transmembrane helices. Six of these helices and the two cytoplasmic loops are analogous to the catalytic core of the SERCA molecule. Specialized metal binding domains with the conserved CxxC motif are

present on both the N-terminus (NMBD) and C-terminus (CMBD). Two additional transmembrane helices are also associated with the N-terminus of CopA.

Figure 2. Selecting the appropriate conformation for homology modeling. In the case of SERCA, the various different X-ray crystal structures show that the linker between the N- and P-domains is flexible. **(A)** The cytoplasmic domains from the EM density map of CopA at 17 Å resolution (1). The locations of the A-, N-, and P-domains are indicated. **(B)** After superimposing the P-domains from three atomic structures from SERCA, their N-domains are seen to adopt different positions. The position of the N-domain in the red model, which corresponds to 1IWO, is most consistent with the EM density map. PDB codes for the three structures are 1IWO (red), 1SUH (pink), 1VFP (olive). The corresponding A-domains have been omitted for clarity.

Figure 3.

Homology model for CopA. **(A)** The crystal structure of SERCA in the E2 conformation (1IWO) provided the template for building the homology model for CopA. SERCA does not have an NMBD, so a separate template (inset) was used, namely a copper metallochaperone (2QIF) which has high sequence homology with NMBD. **(B)** The resulting homology models were produced by ICM. Note that CopA lacks a number of loops in the three cytoplasmic domains but that the overall configuration of these domains is preserved. Some of the transmembrane helices appear to have been misfolded and will need manual rebuilding for optimal fitting to the map. Domains have been colored as follows: N-domain green, P-domain yellow, A-domain red, transmembrane domain blue, metal binding domain purple. SERCA also has two N-terminal helices associated with its A-domain (light blue).

Figure 4.

Docking Results in ICM. The graphical user interface for ICM has several windows for viewing relevant information. The main display window allows interactive visualization of the model. The top-left window allows selection of objects for viewing in the main display window. The bottom-left window shows a table of docking results sorted by energy value; clicking an individual row displays that solution in the visualization window. The bottom-right window displays a plot of each docking solution *versus* energy (on the abscissa). Low energy solutions (e.g., red square) that are well separated from the remainder indicate a successful docking experiment.

Figure 5. Result of fitting the model into the EM map. **(A)** Side view of the cytoplasmic domains from the final model fitted into the 17 Å resolution EM map. **(B)** Top view of these cytoplasmic domains, obtained after a 90° rotation about the horizontal axis. The oblique black line represents the boundary with a two-fold related molecule that has not been fitted with a corresponding model. This model represents the lowest energy structure obtained from the docking experiment and appears to produce an optimal fit with the map.