Combining electron microscopy and comparative protein structure modeling
Maya Topf and Andrej Sali

Recently, advances have been made in methods and applications that integrate electron microscopy density maps and comparative modeling to produce atomic structures of macromolecular assemblies. Electron microscopy can benefit from comparative modeling through the fitting of comparative models into electron microscopy density maps. Also, comparative modeling can benefit from electron microscopy through the use of intermediate-resolution density maps in fold recognition, template selection and sequence-structure alignment.

Addresses
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California San Francisco, San Francisco, CA 94143, USA

Corresponding author: Sali, Andrej (sali@salilab.org)

Introduction
Atomic high-resolution structures of complete macromolecular complexes, such as viruses, ion channels, ribosomes and proteasomes, are needed for studying their assembly, function, regulation and evolution [1–3]. Although some assembly structures have been determined by X-ray crystallography [4] and NMR spectroscopy [3,4], their number is quite small compared to the thousands of different complexes in a typical cell. Thus, other methods for the structural characterization of assemblies at near atomic resolution are needed.

Electron cryo-microscopy (cryo-EM) can image complexes in their physiological environment and does not require large quantities of the sample [5–10]. Structures of macromolecular complexes with molecular weights greater than \(~150\) kDa can now be visualized in different functional states at intermediate resolution \((\sim5–15\ \text{Å})\) [11,12]. The corresponding cryo-EM maps are generally still insufficient for atomic structure determination on their own. However, they do enable accurate fitting of atomic-resolution structures of assembly components (e.g. domains, proteins and subcomplexes) into the lower-resolution density of the whole assembly [13–17]. Such fitting produces an atomic structure of the entire complex. In addition to rigid-body fitting [18–22,23], a growing number of ‘flexible’ fitting methods distort the component structures in order to optimize their fit into the density map [24–31].

Unfortunately, experimentally determined atomic-resolution structures of assembly components are frequently not available. Even when they are available, induced fit may severely limit their usefulness in the reconstruction of the complex. In such cases, it may be possible to get useful models of the components by comparative protein structure modeling (homology modeling) [32–35]. In comparative modeling, the structure of a target protein sequence is predicted by finding one or more related proteins of known structure (i.e. templates), aligning the target sequence to the template structures, building a model based primarily on the alignment from the previous step and assessing the model. Currently, \(~1.1\) million of the \(~1.9\) million known protein sequences [36] have at least one domain that can be modeled based on its similarity to one or more of the \(~30,000\) known protein structures [37]. Thus, the number of models that can be constructed with useful accuracy, at least comparable to the resolution of the cryo-EM maps, is almost two orders of magnitude greater than the number of available experimentally determined structures. Moreover, comparative modeling is becoming increasingly applicable and accurate. These improvements result partly from structural genomics initiatives, which aim to solve the structure of a member of most protein families by X-ray crystallography or NMR spectroscopy, such that most of the remaining proteins can be modeled with useful accuracy based on their similarity to the known structure [37–39]. They are also due to the growth in sequence databases, faster computers and improved comparative modeling methods [34,35,40].

In this review, we list recent advances in methods and applications that involve integrating comparative modeling and EM density maps to produce atomic structures of macromolecular assemblies. We start by describing how EM can benefit from comparative modeling through the fitting of comparative models into EM density maps. Next, we show how comparative modeling can benefit from EM through the use of intermediate-resolution density maps in fold recognition, template selection...
and sequence-structure alignment. Finally, we outline our approach to combined comparative modeling and fitting into a density map.

**Fitting comparative models into a density map**

It is becoming increasingly common to fit comparative models into EM density maps to produce atomic structures of macromolecular assemblies and yield insight into their functions [41]. A recent study showed that fitting a comparative model based on a remotely related template structure is generally better than fitting the template itself [23*]. Specifically, an accurate comparative model of a protein based on an experimentally determined atomic structure of a remotely related homolog generally has a higher correlation with its intermediate-resolution density map than the homolog itself.

Assembly structures that have recently been determined at low resolution (worse than 15 Å), which involved, at least partly, the fitting of comparative models into EM maps, include the actin–fibrin bundle [41], the membrane-bound coagulation factor FVIII–FIXa complex [42], the skeletal muscle Ca$^{2+}$ release channel [43], the yeast exosome [44], smooth muscle α-actinin [45], the cytosolic chaperonin CCF complexed with phodducin-like protein [46] and the karyopherin CRM1–exportin1 complex [47*]. In the study of CRM1–exportin1, integrating sequence analysis, comparative modeling, a low-resolution density map (i.e. a 22 Å resolution ‘negative-stain’ EM map) and mutagenesis data suggested cooperation during formation of the nuclear export complex. This study demonstrates how the hybrid approach can sometimes lead to important biological insights, even when the resolution of the density map is worse than 20 Å. Low-resolution maps, however, may not have sufficiently distinctive features for the unambiguous placement of the high-resolution assembly components [18,30,48]. In such cases, information about component proximity, derived from experiments such as affinity purification or immuno-EM [49], can be combined with the shape information encoded by the EM density to localize the components [50].

At intermediate resolution, the effectiveness of combining cryo-EM maps and comparative modeling is demonstrated by several ribosome studies [15,51–54,55**]. Due to the conservation of the ribosome across species, it is frequently possible to build relatively accurate comparative models of many ribosomal proteins. For instance, comparative models and crystal structures of ribosomal RNAs and proteins were fitted into the 12 Å resolution cryo-EM density map of a targeting complex, consisting of a mammalian signal recognition particle (SRP) bound to an active 80S ribosome carrying a signal sequence [53]. As a result, a molecular model of SRP in its functional state was obtained, showing how its S-domain contacts the large ribosomal subunit at the nascent chain exit site in order to bind the signal sequence. In another example, the ribosome-bound RACK1 protein was first localized on the ribosome using cryo-EM density maps at 15.6 Å resolution with and without the protein [55**]. Next, a comparative model of RACK1 was fitted into the 11.7 Å resolution map of the *Saccharomyces cerevisiae* 80S ribosome (Figure 1). The resulting model structure, supplemented by a representation of the charges on its surface, mutational analysis, and an atomic model of the 40S subunit derived by cryo-EM and comparative modeling, revealed the interactions of RACK1 with the ribosome and clarified its role in the control of eukaryotic translation.

In the subnanometer-resolution range, most examples of fitting comparative models into cryo-EM maps involve highly symmetric biological machines. The usefulness of this approach was recently demonstrated by constructing an atomic model of a complete clathrin lattice [56**] (Figure 1). Fitting crystallographic structures of two components and comparative models of the other seven components into the density map at ~8 Å resolution allowed the tracing of most of the 1675-residue clathrin heavy chain. The atomic structure explains how clathrin-coated vesicles adapt to cargos of different shapes and sizes.

**Fold assignment and template selection**

When the resolution of the density map is better than ~12 Å, it is possible to use it for fold assignment of the constituting domains (Figure 2). Such fold assignments are particularly helpful when a structural homolog of the target component cannot be detected by sequence-based or threading search methods [34,35]. Recent examples include viral [57] and membrane [58] proteins. At ~12 Å resolution, it is usually possible to recognize boundaries between the individual components in the complex. Secondary structure features, such as long α helices and large β sheets, can begin to be identified at ~10 Å resolution, and short helices and individual strands at ~4 Å [59]. When the fold is known, the density maps can be useful in selecting template structures for comparative modeling and fitting [23*]. Furthermore, even if a crystallographic structure of the isolated component exists, the density map may still be useful in selecting a template whose conformation better resembles the conformation of the target protein in the context of the assembly compared to the isolated component.

There are two classes of methods for identifying the structural motifs in a given density map. The methods in the first class identify secondary structure segments and their arrangement in space without recourse to known protein structures. Examples include a method for identifying the location of helices consisting of three or more turns (*Helicshunter* [19]); a method for assigning transmem-
brane helices in the sequences of integral membrane proteins to the approximate locations of the helices in the density map [60]; methods for locating regions belonging to β sheets (Sheetminer [61]) and for building Cα models of β sheets (Sheettracer [62]); a deconvolution method for enhancing secondary structure features in density maps [62]; and a method for determining protein topology from skeletons of secondary structure segments [63*]. These methods can be combined with secondary structure prediction based on sequence only [64–66] to improve the secondary structure assignments [67,68,69*,70]. In addition, the assigned relative spatial arrangement of helices can be matched against known protein structures to identify remote structural relationships [71–73]. This approach was used successfully for detecting previously determined structures similar to the rice dwarf virus outer shell protein RDV P8 [67] and the herpes simplex virus capsid protein VP5 (middle domain) [69*].

The methods in the second class are based on the fitting of structural units larger than secondary structure segments, such as whole folds. Recently, a method for fold assignment at intermediate resolution, SPI-EM, has been developed [74*]. The method combines the fitting of many high-resolution domains from the CATH database.
[75] into a given cryo-EM density map at intermediate resolution with a probabilistic analysis to determine the superfamily of the component. However, as the resolution decreases, some superfamily folds become indistinguishable. In such cases, the method will require additional constraints from other sources, such as bioinformatics or experiment.

**Improving comparative models**

It was recently shown that intermediate-resolution cryo-EM density maps are helpful for improving the accuracy of comparative protein structure modeling [23*]. More specifically, the cross-correlation between a comparative model and the corresponding density map is highly correlated with the accuracy of the model; in other words, a more accurate model fits the EM density map more tightly. Therefore, cryo-EM maps are useful not only for deriving atomic models of assemblies, but also for reducing comparative modeling errors.

The largest errors in comparative models result from incorrect sequence alignment and fold assignment, especially in models of sequences that are remotely related to their templates (i.e. less than 30% sequence identity). Other errors include rigid-body shifts, errors in the modeling of loops and errors in sidechain packing [34]. It is usually possible to use different programs and options to generate a set of models based on alternative templates and alignments that vary in the orientation of domains, the packing of secondary structure elements and the conformation of loops. Selecting the most accurate model from a model set can then be accomplished by combining [23*] density fitting with various methods of model assessment, such as statistical potentials of mean force [76,77].

An example of such an approach relies on the previously developed ‘moulding’ protocol, which iterates over alignment, model building and model assessment [78] (Figure 3). The underlying genetic algorithm depends on the operators that evolve the alignments by mutations and cross-overs, as well as on a fitness function that corresponds to model assessment by statistical potentials. A combined comparative modeling and density fitting protocol (M Topf et al., unpublished) is implemented by adding the cryo-EM density fitting score to the fitness function. A benchmark shows that intermediate-resolution density maps help reduce the $C_\alpha$ root mean square error of comparative models by $\sim 27\%$. Further improvements of combined comparative modeling and cryo-EM fitting are likely to be achieved by exploring the conformational and configurational degrees of freedom using flexible fitting, elastic deformation and real-space refinement methods [24–30].
Conclusions

Our broad objective is to maximize the coverage, accuracy, resolution and efficiency of the structural characterization of macromolecular assemblies [2,3]. This aim is likely to be achieved by integrating techniques that consider various types of information, including density maps from cryo-EM, atomic structures from crystallography and NMR spectroscopy, and atomic models from protein structure prediction. A major class of such hybrid methods involves the fitting of molecular components into cryo-EM density maps of large assemblies. Given the increasing number of molecular machines under investigation using cryo-EM techniques, it is likely that fitting comparative models into intermediate-resolution density maps will increase significantly in the near future.

Acknowledgements

We would like to thank Frank Alber, Fred Davis, Damien Devos, Dmitry Korkin, MS Madhusudhan, Matthew L Baker and Wah Chiu for assistance in preparing this manuscript. We are also grateful for the support of the National Science Foundation (grant EIA-032645), National Institutes of Health (grant R01 GM54762), Human Frontier Science Program, The Sandler Family Supporting Foundation, SUN, IBM and Intel.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest


Comparative modeling and EM density fitting

Topf and Sali


A structural model of the human karyopherin CRM1 is suggested based on a combination of X-ray crystallography, EM and comparative modeling. The architecture of CRM1 resembles that of the import receptor transportin1, with 19 HEAT repeats and a large loop implicated in the binding of the small GTPase Ran. The integrated structural information provides insights into karyopherin-mediated nuclear export signal recognition.


A new method for determining protein topology by defining loop connectivity based on skeletons of secondary structure segments obtained from intermediate-resolution cryo-EM density maps. The method involves a knowledge-based geometry filter and an energy-based evaluation. It is tested successfully on a wide range of protein architectures.


Kinoshita K, Kidera A, Go N: Structural analysis of the 8.5 A resolution cryo-EM map of the major capsid protein (VP5) of herpes simplex virus type 1 is combined with sequence-based bioinformatics to give its architectural model. The model provides insights into the strategies used to achieve viral capsid stability.


SPI-EM is a new method for determining the fold of a protein domain using its EM density map at intermediate resolution. This aim is achieved by fitting protein structural domains representative of the Protein Data Bank into an EM map of the specimen, followed by a probabilistic analysis of the results. The approach is demonstrated with the aid of simulated and experimental multidomain density maps.


