Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 149 (2005) 191–203
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy

Maya Topf[a], Matthew L. Baker[b], Bino John[c], Wah Chiu[b], Andrej Sali[a,*]

[a] *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, 600 16th Street, Suite N472D, University of California, San Francisco, CA 94143, USA*
[b] *Verna and Marrs McLean Department of Biochemistry and Molecular Biology, National Center for Macromolecular Imaging, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA*
[c] *Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA*

Received 13 August 2004, and in revised form 5 November 2004
Available online 8 December 2004

Abstract

We explore structural characterization of protein assemblies by a combination of electron cryo-microscopy (cryoEM) and comparative protein structure modeling. Specifically, our method finds an optimal atomic model of a given assembly subunit and its position within an assembly by fitting alternative comparative models into a cryoEM map. The alternative models are calculated by MODELLER [J. Mol. Biol. 234 (1993) 313] from different sequence alignments between the modeled protein and its template structures. The fitting of these models into a cryoEM density map is performed either by FOLDHUNTER [J. Mol. Biol. 308 (2001) 1033] or by a new density fitting module of MODELLER (Mod-EM). Identification of the most accurate model is based on the correlation between the model accuracy and the quality of fit into the cryoEM density map. To quantify this correlation, we created a benchmark consisting of eight proteins of different structural folds with corresponding density maps simulated at five resolutions from 5 to 15 Å, with three noise levels each. Each of the proteins in the set was modeled based on 300 different alignments to their remotely related templates (12–32% sequence identity), spanning the range from entirely inaccurate to essentially accurate alignments. The benchmark revealed that one of the most accurate models can usually be identified by the quality of its fit into the cryoEM density map, even for noisy maps at 15 Å resolution. Therefore, a cryoEM density map can be helpful in improving the accuracy of a comparative model. Moreover, a pseudo-atomic model of a component in an assembly may be built better with comparative models of the native subunit sequences than with experimentally determined structures of their homologs.

Keywords: Protein structure prediction; Comparative modeling; Electron cryo-microscopy; Fitting

1. Introduction

The native structures of the individual proteins do not yield the "full picture" of the functional assemblies, such as viruses, ion channels, ribosomes, proteasomes, and other molecular machines (Alberts, 1998; Russell et al., 2004; Sali et al., 2003). To this end, the structures of whole assemblies are needed. One of the methods that can be applied to structural characterization of whole assemblies is electron cryo-microscopy (cryoEM) of single particles. Single-particle cryoEM can determine the structures of macromolecular complexes with molecular weights larger than approximately 150 kDa in different functional states and at increasingly higher resolutions (Frank, 2002; Ludtke et al., 2004; Zhou and Chiu, 2003). It is anticipated that the resolution of many single-particle cryoEM structures determined in the immediate future will be in the range of 5–10 Å.

* Corresponding author. Fax: +1 415 514 4231.
E-mail address: sali@salilab.org (A. Sali).
URLs: http://salilab.org, http://ncmi.bcm.tmc.edu/cbs.

1047-8477/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2004.11.004

While it is almost impossible to determine an atomic model only from density maps at 5–10 Å resolution, a wealth of information, such as spatial organization of domains, locations of long α-helices and large β-sheets (Jiang et al., 2001) as well as macromolecular dynamics (Ming et al., 2002; Tama et al., 2002; Wang et al., 2004), can still be obtained from such maps (Russell et al., 2004). Moreover, it has been shown that docking atomic-resolution structures of individual subunits into a cryoEM density map of the intact assembly can result in a useful pseudo-atomic model of the whole assembly (Zhou et al., 2001). Such models can yield significant insights into the structure and function of single proteins and their complexes (Golas et al., 2003; Kostyuchenko et al., 2003; Shin et al., 2003).

Manual docking tools (Beckmann et al., 2001; Beroukhim and Unwin, 1995; Hewat et al., 1998; Hoenger et al., 1998; Rayment et al., 1993; Sosa et al., 1997; Spahn et al., 2001; Voges et al., 1994), which are limited by the experience of the user, are slowly being replaced by more robust and objective docking strategies (Roseman, 2000; Wriggers and Chacon, 2001). The most widely used approach relies on a systematic maximization of the cross-correlation between the model density and the density map. In some earlier studies, this method was employed for local rigid-body refinements of the manual docking solutions, in both reciprocal space (Cheng et al., 1995; Hewat and Blaas, 1996; Kolatkar et al., 1999; Wikoff et al., 1994) and real space (Grimes et al., 1997; Stewart et al., 1993). Recently, routines that rely on a more thorough search over three translational and three rotational degrees of freedom in real space have been introduced, including COAN (Volkmann and Hanein, 1999), DOCKEM (Roseman, 2000), EMFIT (Rossmann, 2000), COLORES (Chacon and Wriggers, 2002; Wriggers et al., 1999), FOLDHUNTER (Jiang et al., 2001), the grid-threading Monte Carlo method (Wu et al., 2003), and 3SOM (Ceulemans and Russell, 2004).

Unfortunately, experimentally determined atomic-resolution structures of the isolated subunits in the complexes are frequently not available. In addition, even if they are available, the induced fit may severely limit their utility in the reconstruction of the whole assembly. In such cases, it might be possible to obtain useful models of the subunits in the correct structural state by comparative protein structure modeling (Baker and Sali, 2001; Jacobson and Sali, 2004; Marti-Renom et al., 2000). For example, partial pseudo-atomic models of the whole yeast (Beckmann et al., 2001; Spahn et al., 2001) and *Escherichia coli* ribosomes (Gao et al., 2003) were obtained by fitting into cryoEM maps comparative protein models calculated from the crystallographic structures of the prokaryotic ribosomal subunits.

Comparative modeling predicts the structure of a target protein sequence by (i) finding one or more related proteins with known structures (i.e., templates), (ii) aligning the target sequence to the template structure, (iii) building a model based primarily on the alignment from the previous step, and (iv) assessing the model (Marti-Renom et al., 2000). It is becoming increasingly applicable and accurate, in large part because of the structural genomics initiative. The structural genomics initiative aims to solve the structures 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 structures (Baker and Sali, 2001; Marti-Renom et al., 2000; Pieper et al., 2004; Sali and Kuriyan, 1999). The largest errors in comparative models result from incorrect sequence alignment and fold assignment, especially in models of the sequences that are only remotely related to their templates (i.e., at less than 30% sequence identity). Most pairs of detectably related protein sequences and structures are currently related at less than 30% sequence identity, with correspondingly large alignment errors (i.e., >20% of misaligned residues). Other errors include rigid-body shifts, errors in the modeling of loops, and errors in side-chain packing (Marti-Renom et al., 2000). It is usually possible to generate a set of models based on alternate templates and alignments that vary in the orientation of domains, packing of secondary structure elements, and conformation of loops. Selecting the best model from a model set can then be attempted through various methods for model assessment (Melo et al., 2002; Sippl, 1993).

Here, we evaluate the utility of cryoEM density maps at 5–15 Å resolution in assessing comparative protein structure models with alignment errors. We describe a method for fitting a given rigid model into a density map, implemented in MODELLER (Sali and Blundell, 1993) (Mod-EM) and an improved version of FOLDHUNTER (Jiang et al., 2003). This procedure is then tested with the aid of a benchmark data set, consisting of eight proteins of different folds with 300 different models each. In addition, we describe the criteria used to assess the correlation between the geometrical accuracy of a model and the quality of its fit into a given density map. Furthermore, we quantify the ability of Mod-EM and FOLDHUNTER to find the most accurate models by the quality of their fit into density maps at different resolutions and noise levels. We also compare model assessment by density fitting with model assessment by statistical potentials of mean force as implemented in the ProsaII program (Sippl, 1993). Finally, we discuss the implications of the results for comparative protein structure modeling and for improving the interpretation of cryoEM density maps.