Journal of Structural Biology 184 (2013) 93–102

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Computational methods for constructing protein structure models from 3D electron microscopy maps

Juan Esquivel-Rodríguez [a], Daisuke Kihara [a,b,c,*]

[a] Department of Computer Science, College of Science, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907, USA
[c] Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Available online 21 June 2013

Keywords:
Electron microscopy
Structure fitting
Macromolecular structure modeling
Electron density map
Computational algorithm

ABSTRACT

Protein structure determination by cryo-electron microscopy (EM) has made significant progress in the past decades. Resolutions of EM maps have been improving as evidenced by recently reported structures that are solved at high resolutions close to 3 Å. Computational methods play a key role in interpreting EM data. Among many computational procedures applied to an EM map to obtain protein structure information, in this article we focus on reviewing computational methods that model protein three-dimensional (3D) structures from a 3D EM density map that is constructed from two-dimensional (2D) maps. The computational methods we discuss range from *de novo* methods, which identify structural elements in an EM map, to structure fitting methods, where known high resolution structures are fit into a low-resolution EM map. A list of available computational tools is also provided.

1. Introduction

Electron density maps from cryo-electron microscopy (cryo-EM) have been used to model macromolecular structures for almost two decades (Volkmann et al., 2000; Ludtke et al., 2004; Mitra et al., 2005). Each step in structure determination by cryo-EM has steadily improved, allowing scientists to determine structures with higher resolutions (Rossmann et al., 2005). The number of structures solved by EM is also increasing, resulting in over 1600 EM maps available in the EM Data Bank (EMDB) (Lawson et al., 2011) (Fig. 1). The entries in EMDB include many important biological macromolecules, such as GroEL and virus capsids (Zhou, 2008). As shown in Fig. 2, resolutions of the EM maps in EMDB range from 3.1 Å to ~9 nm. It has been reported recently that the resolution of structures determined by EM is approaching those determined by X-ray crystallography (Zhang et al., 2008).

The structure determination by cryo-EM involves several stages as overviewed in Fig. 3. Once 2D density maps are obtained by single particle cryo-EM for a sample of macromolecules, they are subjected to computational image processing, refinement, and 3D structure reconstruction. A 3D density map is constructed from individual 2D maps that capture different poses in the sample. Different 2D views of the same pose can be grouped together to form clusters that putatively represent the same molecular orientation. If a considerable degree of structural heterogeneity is present in the sample, finding common features to group the 2D projections into clusters becomes more challenging, and the quality of the clusters directly impacts the resolution of the reconstruction (Förster and Villa, 2010). Several experimental steps in the process, such as the centrifugation and the freeze-and-thaw steps, are crucial for obtaining samples that contain structurally homogeneous particles. While homogeneity in the sample can create higher resolution maps, the ability to handle heterogeneous samples is one of the advantages that cryo-EM has over other structural determination techniques such as X-ray crystallography and NMR spectroscopy. Other factors that influence the quality of the 3D map include electron beam alignment, compensating for specimen drift, and making corrections for defocusing. An iterative refinement process can be applied for 2D map alignment that considers these factors.

Electrons can cause radiation damage to biological samples and thus impacts the attainable resolution. Low-contrast images and challenging signal-to-noise ratios are the main problem that computational methods have to deal with in terms of image

Abbreviations: 3DZD, 3D Zernike descriptor; CATH, Class, Architecture, Topology, Homologous superfamily. Acronym for the CATH protein structure database; DEN, Deformable Elastic Network; EM, electron microscopy; EMDB, Electron Microscopy Data Bank; ENM, Elastic Network Model; MC, Monte Carlo; MD, molecular dynamics; MDFF, molecular dynamics flexible fitting; NMA, normal mode analysis; NMFF, Normal Mode Flexible Fitting; NMR, Nuclear Magnetic Resonance; PDB, Protein Data Bank; RMSD, Root Mean Square Deviation; SVM, support vector machine.

* Corresponding author at: Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA. Fax: +1 765 496 1189.
E-mail address: dkihara@purdue.edu (D. Kihara).

1047-8477/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2013.06.008

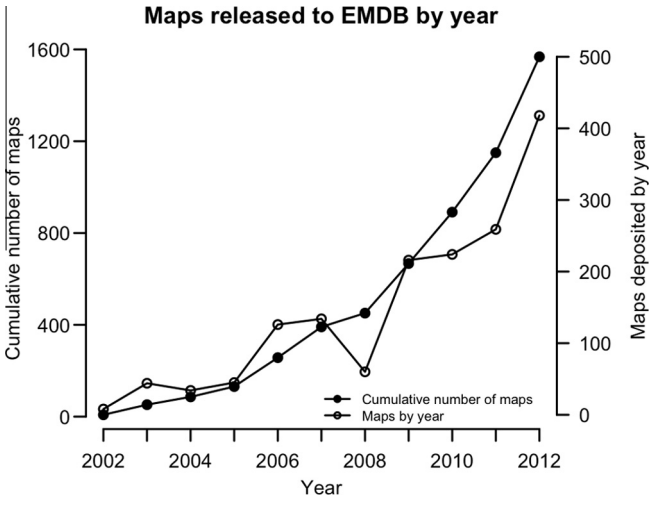

Fig.1. The growth in the number of EM Maps in the Electron Microscopy Data Bank. The total number of entries at each year (horizontal axis), starting in 2002, is shown as a continuous line with filled circles (left vertical axis). Additionally, a continuous line with unfilled circles shows the number of entries deposited each year (right vertical axis).

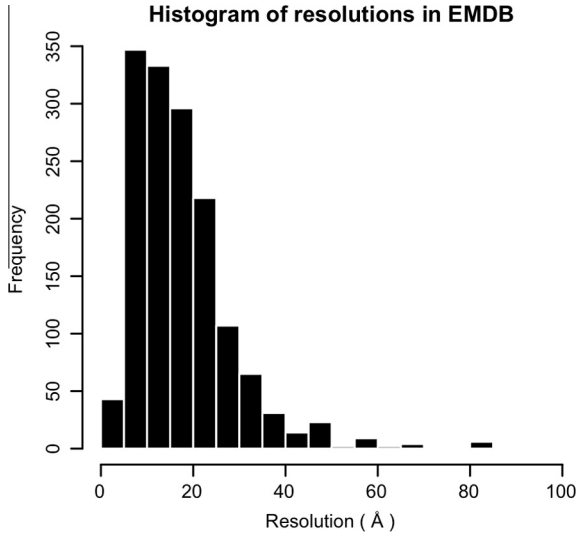

Fig.2. Distribution of EM map resolutions in the EMDB. Entries are eliminated if the resolution is not provided.

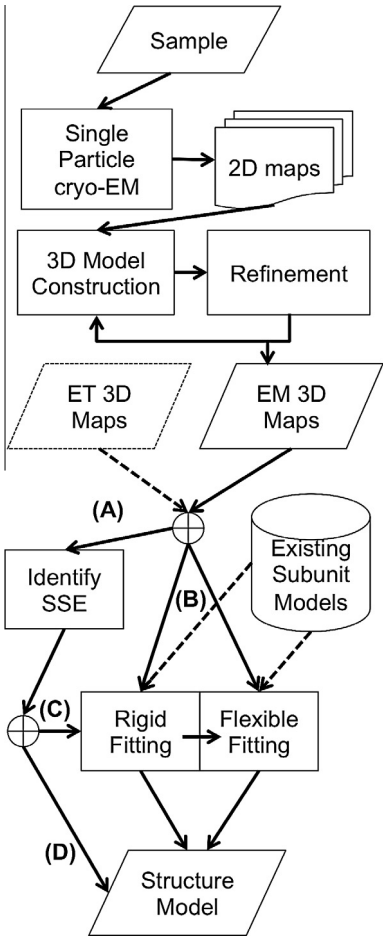

Fig.3. Steps involved in constructing structure models of proteins from an EM map. The first steps involve experiments such as sample preparation and single-particle cryo-EM data collection. Once a 3D EM map is constructed, computational methods are applied to build structural models, which range from determining secondary structure elements to rigid-fitting and flexible fitting approaches. (A) α-helices start to be identifiable in an EM map if its resolution is 10 Å or higher and they can be clearly identified at 6 Å resolution, while β-sheets can be identified in a map at around 5 Å. To follow this route in the diagram the input EM maps should meet the resolution. (B) In order to perform structural fitting in a map the user needs to have atomic-detailed models of the subunits to fit. These can come either from X-ray, NMR spectroscopy, or computational modeling. (C) Some methods use the identified SSEs as input to their rigid fitting algorithm. (D) Coarse-grain models that provide only a backbone trace can be directly derived from identified SSEs in the EM map.

processing (Chiu et al., 2005). Ruprecht and Nield (Ruprecht and Nield, 2001), as well as Zhou more recently (Zhou, 2008, 2011), have presented in-depth discussions about the factors that contribute to higher quality EM maps. A high-resolution map is a requirement for constructing a high-quality atomic-level model.

Once a 3D electron density map has been determined, different types of computational methods can be applied to obtain the 3D structure information of biological macromolecules. The effectiveness and types of methods used depend on the density map resolution or additional information available for the macromolecules being studied. The aim of the methods ranges from the identification of secondary structure elements to the modeling of full-atom structures (Fabiola and Chapman, 2005; Topf and Sali, 2005; Lindert et al., 2009b; Beck et al., 2011).

In this review, we focus on discussing computational methods and tools used for constructing 3D structure models from a 3D EM map (the bottom half in Fig. 3). First, we describe methods to identify local structures, particularly secondary structures in a 3D EM density map, without assuming the availability of an atomic-detailed protein structure to fit in. The following three sections analyze methods for fitting atomic-detailed structures of proteins into an EM map. We begin with analyzing different scoring functions that evaluate the quality of a fit. Then, we explain the main characteristics of methods for fitting high-resolution structures into an EM map that do not explicitly consider protein flexibility (rigid fitting). What follows is a discussion on methods that account for protein structure flexibility in structure fitting (flexible fitting). While some of the methods clearly belong to one of the aforementioned sections, readers should note that the classification of methods is not always clear-cut because they have multiple components that belong to different classifications. Finally, in the last section, we discuss examples of actual applications using these methods.

2. Identification of secondary structure elements in an EM map

If the resolution of an EM map is below 10 Å, the secondary structures of proteins can be identified in the map (Beck et al., 2012). Normally α-helices start to be identified in a map at a