Journal of Structural Biology 175 (2011) 264–276

Contents lists available at ScienceDirect

Journal of Structural Biology

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

BCL::EM-Fit: Rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement

Nils Woetzel [a,c], Steffen Lindert [a,c], Phoebe L. Stewart [b,c], Jens Meiler [a,c,*]

[a] Department of Chemistry, Vanderbilt University, TN 37212, USA
[b] Department of Molecular Physiology and Biophysics, Vanderbilt University, TN 37212, USA
[c] Center for Structural Biology, Vanderbilt University, TN 37212, USA

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form 28 April 2011
Accepted 28 April 2011
Available online 4 May 2011

Dedicated to Dr. Brigitte Heink on occasion of her retirement.

Keywords:
Cryo-electron microscopy
CryoEM
Geometric hashing
Real space
Monte Carlo Metropolis
Fitting
Docking

ABSTRACT

Cryo-electron microscopy (cryoEM) can visualize large macromolecular assemblies at resolutions often below 10 Å and recently as good as 3.8–4.5 Å. These density maps provide important insights into the biological functioning of molecular machineries such as viruses or the ribosome, in particular if atomic-resolution crystal structures or models of individual components of the assembly can be placed into the density map. The present work introduces a novel algorithm termed BCL::EM-Fit that accurately fits atomic-detail structural models into medium resolution density maps. In an initial step, a "geometric hashing" algorithm provides a short list of likely placements. In a follow up Monte Carlo/Metropolis refinement step, the initial placements are optimized by their cross correlation coefficient. The resolution of density maps for a reliable fit was determined to be 10 Å or better using tests with simulated density maps. The algorithm was applied to fitting of capsid proteins into an experimental cryoEM density map of human adenovirus at a resolution of 6.8 and 9.0 Å, and fitting of the GroEL protein at 5.4 Å. In the process, the handedness of the cryoEM density map was unambiguously identified. The BCL::EM-Fit algorithm offers an alternative to the established Fourier/Real space fitting programs. BCL::EM-Fit is free for academic use and available from a web server or as downloadable binary file at http://www.meilerlab.org.

1. Introduction

Cryo-electron microscopy (cryoEM) (Lepault et al., 1983) has evolved in the past decade as an important tool to obtain medium resolution structures of biological macromolecular assemblies in the form of density maps. One challenge is to dock high resolution experimental structures, obtained by X-ray crystallography (Kendrew et al., 1958) and nuclear magnetic resonance (NMR) (Wüthrich, 1990), or models of individual proteins into these density maps to arrive at quasi atomic-detail representations of the macromolecular assembly. This procedure identifies regions of conformational change and regions that can be assigned to proteins of uncharacterized structure or which are characterized only in isolation.

Several protocols have been developed to fit atomic structures, usually obtained by X-ray crystallography or NMR, into low and medium resolution density maps (Fabiola and Chapman, 2005; Wriggers and Chacón, 2001). The computational problem amounts to determining six degrees of freedom, three rotational and three translational. Exhaustive searches systematically seek within this six-dimensional parameter space to optimize the cross correlation coefficient (*CCC*), which consumes significant amounts of computational time (Korostelev et al., 2002; Roseman, 2000). Computational time can be reduced by the use of a fast Fourier transformation accelerated translational search as implemented in the "Colores" program within the SITUS package (Wriggers et al., 1999). In this approach only the three rotational degrees of freedom are searched in an exhaustive fashion in real space, while the translational degrees of freedom are searched in Fourier space. For both algorithms the step size impacts the speed of the calculation, but also the reliability and quality of the solution. An optimal local fit can be found with Chimera. It provides the benefit of a graphical user interface and an implementation of gradient refinement (Goddard et al., 2007). This refinement is only local and re-

Abbreviations: EM, electron microscopy; MCM, Monte Carlo/Metropolis; $RMSD_{C\alpha}$, root mean square distance or deviation of the Cα-atom coordinates; *CCC*, cross correlation coefficient; Voxel, volume pixel.

* Corresponding author at: Vanderbilt University, Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology, 465 21st Ave. South, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232-8725, USA. Fax: +1 615 936 2211.

E-mail addresses: nils.woetzel@vanderbilt.edu (N. Woetzel), jens.meiler@vanderbilt.edu (J. Meiler).

URL: http://www.meilerlab.org (J. Meiler).

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.04.016

quires that the initial placement is closer to the correct solution than the protein diameter. Gradient based local minimization has been implemented on general purpose graphical processing units (GPGPU) showing speed ups of at least 30 with the same accuracy as a CPU version (Woetzel et al., 2011).

To further increase the speed of fitting, vector quantization was introduced (Wriggers and Birmanns, 2001). Single molecule data is represented by k so-called codebook vectors for high resolution protein structure data and low resolution density maps. In a search within the k! permutations the best fit is identified by the lowest residual $RMSD_{C\alpha}$ after superimposition. This "QDOCK" method in the SITUS program is fast and reliable for rigid body docking and can be used for flexible docking as well. Difficulties arise however, if the density map contains different and multiple protein structures.

Protein structures obtained by X-ray crystallography often differ from the form of the protein observed in the cryoEM experiment. This can be the case if the protein was modified to facilitate crystallization or if a comparative model was built from a crystal structure of a homologous protein. In these cases the atomic model might not reflect all of the structural and dynamical properties observed in the cryoEM map. Therefore, flexible docking protocols were developed to overcome the limitations of rigid body fitting. For example, structural alignments of one protein to proteins in the same super family can be used to sample different conformations and improve the *CCC* (Velazquez-Muriel and Carazo, 2007). Alternatively, normal mode based fitting varies the coordinates of the structure within reasonable limits while docking (Tama et al., 2004). Molecular dynamics approaches have also been tested to optimize the fit of an atomic structure into electron density maps (Schröder et al., 2007; Trabuco et al., 2009). Flexible docking can also be achieved by defining hinges between domains and varying the orientation between them using QDOCK in the SITUS package. Methods such as molecular dynamics, conjugate-gradient minimization, and Monte Carlo optimization can be integrated with different scoring functions in an iterative protocol that combines the strengths of each individual approach (Topf et al., 2008).

The present work implements for the first time a "geometric hashing" algorithm (Wolfson and Rigoutsos, 1997) termed BCL::EM-Fit for the task of fitting atomic-detail protein models into cryoEM densities. Geometric hashing was developed in the robotics field, where feature-recognition and pattern-matching give computers the ability to connect real life objects to abstract computational representations. This technique is already used in structural biology to identify similar binding sites in proteins (Shulman-Peleg et al., 2004). A second step in the BCL::EM-Fit approach involves a Monte Carlo (Metropolis and Ulam, 1949)/Metropolis (Metropolis et al., 1953) (MCM) small perturbation protocol to refine the initial fits by maximizing the *CCC*. The time and robustness of BCL::EM-Fit compares favorably with the widely used Fourier/real space fitting program "COLORES" in the SITUS package (Wriggers and Birmanns, 2001). Benchmark results are presented with simulated density, as well as examples that demonstrate fitting with experimental GroEL density (Stagg et al., 2008) and of adenovirus capsid protein crystal structures into experimental cryoEM density maps (Saban et al., 2006).

2. Methods

2.1. Geometric hashing re-casted for searching density maps with protein structures

The following paragraph gives a general overview of the steps required before a more detailed description of the present implementation is given. The basic idea of geometric hashing was developed for image recognition in robotic applications. Critical points of a complex image (features) are extracted into a feature cloud. A large number of possible rotations and translations of this feature cloud are encoded *a priori* in a hash map (Wolfson and Rigoutsos, 1997) which later allows a rapid search for objects within this image. For BCL::EM-Fit the 3D image will be the cryoEM density map. The objects to be recognized will be protein structures which will also be represented as feature clouds. Each combination of a rotation (three degrees of freedom) and translation (three degrees of freedom) of the feature cloud is a transformation with six degrees of freedom.

The general scheme for generating the geometric hash is to define many possible transformations for the density map feature cloud and store these in a memory-efficient, rapidly searchable hash map. In this process the features are "quantized", i.e. not the actual position of a feature but only the specific space bin that contains the feature is stored. This procedure not only saves memory and accelerates the search, it also limits the search to a finite (but large) set of all possible transformations. Further it compensates for experimental noise in the density map and protein structure. In the recognition step this hash map is searched with a feature cloud representation of the protein to be docked. It is expected that one of the original transformations puts the feature cloud of the density map in good overlap with the feature cloud of the protein. This can be recognized by the number of shared features, i.e. features that end up in the same space bin.

This procedure speeds up the search as not the complete image but only the features deemed important are considered. Further, not every possible transformation is considered but only a finite subset. In contrast to robotics the problem of scaling the image is absent for feature-recognition in a distance invariant cryoEM density because the units of length in the density map and atomic models are the same. Further, 3D images have an increased complexity over 2D pictures that a robot usually sees using a single camera, which changes the protocol slightly compared to plain 2D picture recognition.

2.2. Extraction of feature cloud from density map intensities (Fig. 2a)

The user inputs a density map that will be completely encoded as a point cloud for rapid fitting. If the user wants to fit into a specific segment of the density map, it is necessary to extract that from the original map in a pre-processing step. In order to generate a representation of the features in the density map two pieces of information are used (Fig. 2a): the absolute intensity of a voxel and the intensity difference to its neighboring voxel, a gradient. The higher the intensity the more likely it is that a structurally compact region such as a secondary structure element can be found in the respective position of the density maps. The higher the intensity gradient the more likely the edge of a secondary structure element can be found here. Often there is an intensity drop at the edge of secondary structure elements due to less rigid amino acid side chain atoms. The edge regions are usually close to backbone atoms of secondary structure elements and encode most of the information within the density map. In order to define the total number of features extracted from a density map Eq. (1) was derived empirically:

$$N_{\text{points}} = N_{\text{Voxel Atoms}} \times \frac{V_{\text{Voxel}}}{Max(\frac{\pi}{6} d_{fd}^3, V_{Voxel}, \rho_{\text{Atoms Protien}}^{-1})} \quad (1)$$

where $N_{\text{Voxel Atoms}}$, Number of voxels the atoms would occupy when mapped to grid of the density map; V_{Voxel}, Volume of voxel; $\frac{\pi}{6} d_{fd}^3$, Volume that one point occupies according to feature distance; V_{Vox}, Volume that one point occupies according to a Voxel's volume;