Journal of Structural Biology 174 (2011) 333–343

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Scoring functions for cryoEM density fitting

Daven Vasishtan, Maya Topf*

Institute of Structural and Molecular Biology, Crystallography, Department of Biological Sciences, Birkbeck College, University of London, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 10 November 2010
Received in revised form 22 January 2011
Accepted 31 January 2011
Available online 4 February 2011

Keywords:
Electron cryo-microscopy
Density fitting
Scoring functions
Cross-correlation function
Mutual information
Laplacian filtering
Envelope score

ABSTRACT

In fitting atomic structures into cryoEM density maps of macromolecular assemblies, the cross-correlation function (CCF) is the most prevalent method of scoring the goodness-of-fit. However, there are still many possible, less studied ways of scoring fits. In this paper, we introduce four scores new to cryoEM fitting and compare their performance to three known scores. Our benchmark consists of (a) 4 protein assemblies with simulated maps at 5–20 Å resolution, including the heptameric ring of GroEL; and (b) 4 experimental maps of GroEL at ~6–23 Å resolution with corresponding fitted atomic models. We perturb each fit 1000 times and assess each new fit with each score. The correlation between a score and the Cα RMSD of each fit from the "correctly" fitted structure shows that the CCF is one of the best scores, but in certain situations could be augmented or even replaced by other scores. For instance, our implementation of a score based on mutual information outperforms or is comparable to the CCF in almost all test cases, and our new "envelope score" works as well as the CCF at sub-nanometer resolution but is an order of magnitude faster to calculate. The results also suggest that the width of the Gaussian function used to blur the atomic structure into a density map can significantly affect the fitting process. Finally, we show that our score-testing method, when combined with the Laplacian CCF or the mutual information scores, can be used as a statistical tool for improving cryoEM density fitting.

1. Introduction

Single-particle reconstructions from cryo-electron microscopy (cryoEM) fill a niche in structure determination of biomolecules and their assemblies (Frank, 2009; Lawson et al., in press). Assemblies too large for NMR or too difficult to crystallise can be studied, and different conformational states can be revealed when these assemblies are observed in more native conditions. However, low resolution frequently makes obtaining atomic models based on cryoEM maps a non-trivial task. Most cryoEM maps have a resolution between ~5 and 15 Å (Chiu et al., 2005), and at these levels the unambiguous placement of atoms is not feasible. Prior information is needed, and the most obvious and useful data is available from crystallography, NMR or comparative modelling. It is for this reason that fitting atomic components of proteins and nucleic acids into the lower-resolution EM maps is the primary method for extracting pseudo-atomic models from those maps (Rossmann et al., 2005). Although this can be done manually using visualisation programs such as Chimera (Pettersen et al., 2004), such efforts are affected heavily by user bias. For maps of larger assemblies containing many different components, often in different conformations from those being fitted, the problem can also become intractably difficult for users. Automation of fitting can alleviate these problems, and also provide a less biased result.

There exists an extensive list of automated fitting programs that deal with various aspects of the task (Beck, in press; Fabiola and Chapman, 2005). Essential to all of them is the scoring function used to evaluate the goodness-of-fit, as well as a means of optimisation – an iterative algorithm that modifies the degrees of freedom incrementally to improve the fit. Optimisation methods vary significantly amongst these programs, depending on the information and data at hand. However, an accurate and sensitive means of describing the goodness-of-fit is almost always the same; for the programs currently in existence this is the *cross-correlation function* (*CCF*) between the map and the atomic structure blurred to match. Other scores include variations on the CCF, such as the Local CCF (Roseman, 2000), Laplacian-filtered CCF (Chacon and Wriggers, 2002) and core-weighted CCF (Wu et al., 2003). A different kind of fitting score was introduced in the 3SOM algorithm (Ceulemans and Russell, 2004), which is based on optimising the positions and orientations of vectors representing the surfaces of the maps. Recently, due to the need to fit multiple components in large assemblies and modify their conformations, new scores have been introduced that include the consideration of the physicochemical properties of the probe structure (Fabiola and Chapman, 2005), non-bonded interactions terms (Fabiola and

* Corresponding author. Tel.: +44 (0) 20 7079 0886.
E-mail address: m.topf@cryst.bbk.ac.uk (M. Topf).
URL: http://www.cryst.bbk.ac.uk/~ubcg67a (M. Topf).

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.01.012

Chapman, 2005; Rossmann, 2000), as well as the density envelope and geometric complementarity between the individual assembly components (Lasker et al., 2009).

Here we present four scores that are new to EM fitting, one that is frequently used in computer vision algorithms, one that is a standard in probability theory and two of which we have developed. We then compare them to three existing scores: the CCF, the Laplacian-filtered CCF (Chacon and Wriggers, 2002), and our implementation of the vector-based surface superimposition from the 3SOM algorithm (Ceulemans and Russell, 2004) (here referred to as the "normal vector score"). To evaluate the scores, we tested each of them on 4 experimental and 16 simulated cryoEM density maps at ~6–23 Å resolution. In Section 2, we describe the procedures used to calculate each score, as well as the procedures used to test their accuracy and usefulness. In Section 3, we present our findings and discuss the performance of each score. Finally, in Section 4, we summarise the application of our new scores to density fitting of macromolecular assemblies.

2. Methods

We implemented a total of seven different scores for measuring the goodness-of-fit of an atomic structure into a density map; the CCF, the Laplacian-filtered CCF (LAP), the difference least-squares function (DLSF), the envelope score (ENV), the normal vector score (NV), the Chamfer distance (CD) and the mutual information score (MI). Below we describe the scores and the procedures needed for their calculations.

2.1. Convoluting an atomic structure into an EM density map

Given a *target map* to which we want to fit an atomic structure, all bar the ENV score require we first blur the structure to the map resolution in the following way:

(1) Impose a 3D grid with voxel size of 1 Å on the atomic structure to be fitted.
(2) For every non-hydrogen atom the density value of the nearest voxel is increased by the atomic number of that atom.
(3) Convolute the grid with a Gaussian function, using the *fourier_gaussian* function in the SciPy package (Jones, 2001). The sigma value for the Gaussian can be given by four different values (0.187, 0.356, 0.425 and 0.5) multiplied by the target map resolution (corresponding to: the Gaussian width of the Fourier transform falling to half the maximum at 1/resolution; the Gaussian width at 1/e maximum height being equal to the resolution; the full-width half-maximum being equal to the resolution; and the sigma value simply equal to half the resolution, respectively).
(4) Resample the grid to match the sampling of the target map using the *resample* method in the SciPy package.

2.2. Calculating the volume threshold

Four of the scores (CD, DLSF, ENV and NV) rely on an accurate and stable way to define the surface of a map. We characterise the surface as the set of points of density ρ_1, such that the volume corresponding to all density points less than ρ_1 is equal to the volume of the protein in question. The volume of the protein is calculated by an empirical result of the average volume globular proteins occupy (1.21 Å^3/kDa) (Harpaz et al., 1994). Since the amount of points with the exact value ρ_1 will typically be small, a second threshold $\rho_2 > \rho_1$ is used and set such that the number of density points $\rho_2 > \rho > \rho_1$ is equal to ~10% of the protein volume. Henceforth, this set of points is referred to as the 'volume threshold'.

2.3. Scores

We use the following equations to describe the specific scores:

2.3.1. Cross-correlation function (CCF)

A typical method of comparison between two sets of vectors is the measurement of the Euclidean distance between them. The four-dimensional Euclidean distance or the least-squares function (LSF), between the target (T) and the probe (P) maps is given by the difference in the densities of every two corresponding voxels:

$$\mathrm{LSF_{EM}} = \sum_{\mathrm{i}} \left(\rho_{\mathrm{i}}^{\mathrm{T}} - S\rho_{\mathrm{i}}^{\mathrm{P}}\right)^2 \tag{1}$$

where $\rho_{\mathrm{i}}^{\mathrm{P}}$ is the density at point i in the probe map, $\rho_{\mathrm{i}}^{\mathrm{T}}$ the density at the same point in the target map, and S is a scaling factor. Expanding Eq. (1) leads to:

$$\mathrm{LSF_{EM}} = \sum_{\mathrm{i}} \left(\rho_i^{\mathrm{T}^2} - 2S\rho_{\mathrm{i}}^{\mathrm{T}}\rho_{\mathrm{i}}^{\mathrm{P}} + S^2\rho_{\mathrm{i}}^{\mathrm{P}}\right)^2 \tag{2}$$

If we assume that the sums of the square densities of both maps (the first and last terms in Eq. (2)) are constant then they can be ignored. Since the scaling factor will also be constant, we can reduce the LSF to the CCF, given by:

$$\mathrm{CCF} = \sum_{\mathrm{i}} \rho_{\mathrm{i}}^{\mathrm{T}}\rho_{\mathrm{i}}^{\mathrm{P}} \tag{3}$$

where the maximisation of the CCF is equivalent to the minimisation of the LSF. In our method, the CCF is simply implemented using array multiplication of the probe and target maps in SciPy.

2.3.2. Laplacian-filtered CCF

Modification via filtering of the probe and the target densities can sometimes improve the performance of fitting scores. For example, a Laplacian filter, which is an approximation of the partial second derivative of a function, has been used (Chacon and Wriggers, 2002). The rationale behind this approach is to pick out a contour from the map that resembles the surface of the structure. This would approximate a more 'human style' of fitting by matching edges rather than the whole density. The Laplacian filter also acts to sharpen the maps, and therefore heighten the sensitivity of the CCF. The kernel for the filter is given by:

$$\begin{aligned}\nabla^2 a_{l,m,n} = &-6a_{l,m,n} + a_{l+1,m,n} + a_{l-1,m,n} + a_{l,m+1,n} + a_{l,m-1,n} \\ &+ a_{l,m,n+1} + a_{l,m,n-1}\end{aligned} \tag{4}$$

Thus, each voxel $a_{l,m,n}$ is modified as determined by the voxels that surround it. The Laplacian filter is implemented using the 'laplace' command in the *scipy.ndimage.filter* module. The filtered densities are then scored using the CCF as described above.

2.3.3. Difference least-squares function (DLSF)

The DLSF is similar to the LSF; whereas the LSF compares absolute density values, the DLSF compares the difference between pairs of values. In its full form, the equation is given by:

$$\mathrm{DLSF} = \sum_i \sum_{j>i} \left(\left(\rho_i^{\mathrm{T}} - \rho_j^{\mathrm{T}}\right) - \left(\rho_i^{\mathrm{P}} - \rho_j^{\mathrm{P}}\right)\right)^2 \tag{5}$$

Thus the difference between every pair of points in the experimental map is compared to the corresponding pair in the simulated map. Unfortunately, using every single possible pair of points would make this score far too computationally expensive (with $(i-1)$ calculations to make). We therefore use the following equation to calculate a partial DLSF score:

$$\mathrm{pDLSF} = \sum_{k,l\neq k} \left(\left(\rho_k^{\mathrm{T}} - \rho_l^{\mathrm{T}}\right) - \left(\rho_k^{\mathrm{P}} - \rho_l^{\mathrm{P}}\right)\right)^2 \tag{6}$$