Journal of Structural Biology 166 (2009) 67–78

Contents lists available at ScienceDirect

Journal of Structural Biology

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

A method for the alignment of heterogeneous macromolecules from electron microscopy

Maxim Shatsky [a,b,*,1], Richard J. Hall [b,1], Steven E. Brenner [a,b], Robert M. Glaeser [b]

[a] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA
[b] Physical Biosciences Division, Lawrence Berkeley National Laboratory, CA 94720, USA

ARTICLE INFO

Article history:
Received 13 June 2008
Received in revised form 13 December 2008
Accepted 18 December 2008
Available online 30 December 2008

Keywords:
Particle alignment
Heterogeneous data
2D alignment
EM reconstruction

ABSTRACT

We propose a feature-based image alignment method for single-particle electron microscopy that is able to accommodate various similarity scoring functions while efficiently sampling the two-dimensional transformational space. We use this image alignment method to evaluate the performance of a scoring function that is based on the Mutual Information (MI) of two images rather than one that is based on the cross-correlation function. We show that alignment using MI for the scoring function has far less model-dependent bias than is found with cross-correlation based alignment. We also demonstrate that MI improves the alignment of some types of heterogeneous data, provided that the signal-to-noise ratio is relatively high. These results indicate, therefore, that use of MI as the scoring function is well suited for the alignment of class-averages computed from single-particle images. Our method is tested on data from three model structures and one real dataset.

1. Introduction

Single-particle electron microscopy (EM) is used to resolve the three-dimensional (3D) structure of large macromolecular complexes (Frank, 2002a; van Heel et al., 2000). The reconstruction of a 3D structure from randomly orientated two-dimensional (2D) projections requires the assignment of both 2D translational alignment parameters and relative 3D Euler angles. Image alignment is used to assign translational parameters to raw images so that they can be classified and averaged in 2D, and to compare projections with an initial model so that Euler angles can be assigned (projection matching) (Penczek et al., 1994). Due to the large number of raw images and model templates used in a reconstruction, computationally efficient alignment methods based on the Fast Fourier Transform are widely used (Joyeux and Penczek, 2002). This use of the Fast Fourier Transform enables the processing of very large datasets, but limits the scoring function to cross-correlation or one of its variants.

The refinement of a 3D structure relies on the availability of an initial 3D model. This initial 3D model can be from a closely related structure generated by EM or X-ray crystallography, an ab initio reconstruction obtained using Random Conical Tilt (RCT) (Radermacher et al., 1987), or a reconstruction obtained using a common line based method (Penczek et al., 1996; van Heel, 1987). The initial 3D model is used as a template for the assignment of translational parameters and Euler angles. The raw data can then be used to calculate a new structure which again is used to assign translation and Euler angle parameters to the raw images. This process (projection matching) is used iteratively to improve the accuracy of parameter assignment, resulting in progressively higher resolution structures.

This basic projection matching approach relies heavily on the structural homogeneity of the macromolecule being studied. Problems arise when the macromolecule exhibits structural flexibility or alternative assemblies (Brink et al., 2004; Leschziner and Nogales, 2007; Orlova et al., 1999; Roseman et al., 2001; Saibil et al., 2001; Staley and Guthrie, 1998; Yang et al., 2002). Currently, the most successful approach for studying such samples is to biochemically isolate, or fix the structure in a single conformational state before microscopy (Frank, 2003); however, this is not always feasible. Several computational methods have been developed to study heterogeneous data, either by classifying the data into homogeneous subsets (Burgess et al., 2004; Fu et al., 2007; Hall et al., 2007; White et al., 2004), or by quantifying variability in the final reconstruction (3D-variance) (Penczek et al., 2006b). One approach for dealing with heterogeneous data uses maximum-likelihood-based classification (Scheres et al., 2007) whereby the Expectation Maximization steps reassign the raw images into several structurally distinct models. However, this approach is extremely computationally expensive and requires a large quantity

* Corresponding author. Address: Department of Plant and Microbial Biology, 461 Koshland Hall, University of California, Berkeley, CA 94720-3102, USA. Fax: +1 206 339 4619.
E-mail address: maxshats@compbio.berkeley.edu (M. Shatsky).
[1] These authors equally contributed to this work.

1047-8477/$ - see front matter © 2009 Published by Elsevier Inc.
doi:10.1016/j.jsb.2008.12.008

of experimental data. 3D-variance can be used as a tool to analyze heterogeneity in a reconstruction, enabling regions of high variability to be further investigated using other tools (Penczek et al., 2006a). Another approach uses cross-correlation of common lines (CCCL) (Hall et al., 2007) to segregate the 2D averages of raw particles (class-averages) into homogeneous 3D structures. This approach essentially clusters class-averages into groups that represent distinct conformational states. The methods of 3D-variance and CCCL rely on the accurate assignment of in-plane translation and Euler angle parameters, generally found by projection matching. However heterogeneity within the data has a detrimental effect on projection matching by reducing the accuracy of parameter assignment, and therefore lowering the quality of the final reconstruction and the ability of these methods to work.

The cross-correlation function is widely used to optimize alignment between two images. It has been shown to be very successful in aligning highly noisy images to model templates. It is of most use when the particles are structurally homogeneous. Heterogeneous data are problematic, as templates from a single model cannot be fully representative of the data being studied; thus, multiple models can be used in the processing of heterogeneous data (Brink et al., 2004; Falke et al., 2005; Valle et al., 2002). In this approach, translation, relative Euler angles, and conformational state are assigned to particle projection images according to the template for which they have the highest cross-correlation. This approach was successful on multiple occasions when processing EM data of the ribosome (Frank, 2003), but relies on the availability of multiple starting models that account for the heterogeneity within the sample.

Here we present a computationally efficient alignment method that is not limited in the scoring function which is used. The method identifies basic image features that are used to define local directions that are subsequently used to superimpose two images. The identified image features allow a more efficient sampling of transformational space than an exhaustive search approach. Our data indicate that the application of a scoring function based on the Mutual Information (MI) improves the alignment of heterogeneous particles when aligning to a single model, given that the signal-to-noise ratio of the images being aligned approaches that which is expected for class-averages. Mutual information measures the amount of shared information between two images. It has been extensively used in medical imaging for image registration (alignment) (Maes et al., 1996; Pluim et al., 2003; Viola and Wells, 1995). For example, two images of roughly the same brain area obtained using two different techniques, such as PET, which captures functional processes, and CT, which captures structural information, can be aligned based on their shared information rather than on a linear relationship between gray level distributions. In the field of electron microscopy mutual information has been previously used for the alignment of 3D volumes that differ in resolution (Telenczuk et al., 2006).

We demonstrate the performance of MI as a scoring function using three model structures and one real dataset. As model structures we use the 50s and 70s subunits of the *Escherichia coli* ribosome (Gabashvili et al., 2000; Matadeen et al., 1999), the multiprotein splicing factor SF3b (Golas et al., 2005), and the Kelnow fragment of DNA polymerase I (Beese et al., 1993). The ribosome structure was used in the development of the method and demonstrates the ability of mutual information to utilize a single model for the alignment of heterogeneous data that show alternative molecular assemblies. SF3b is used to demonstrate the model independent nature of our alignment method, while the Klenow fragment of DNA polymerase I is used as a test for the type of heterogeneity observed in the real cryo-EM data of eIF3-IRES complex (Siridechadilok et al., 2005), i.e., a constant structure with a protruding flexible domain. Cryo-EM data from human translation initiation factor eIF3 bound to the flexible internal ribosome entry site (IRES) from the hepatitis C virus (HCV) (Siridechadilok et al., 2005) are used to show that our new method is successful for use with experimental images of biological macromolecules.

Our results indicate that mutual information leads to a better alignment than cross-correlation when the signal-to-noise ratio (SNR) is at the level usually found in class-averages, e.g., SNR of 1:1. Therefore, the best use of the MI based alignment is in the reconstruction approaches that first compute class-averages from raw data and then align class-averages to a model (Elad et al., 2008; Hall et al., 2007). Thus, we expect that our approach will improve the alignment step of such methods and consequently will lead to a better reconstruction of multiple models from heterogeneous data.

2. Methods

Given a set of template images, $\{m_i\}$, consisting of evenly spaced 2D projections of an initial 3D model, and a set of raw images, $\{r_i\}$, boxed from an electron micrograph, *projection matching* is the problem of finding, for each raw image r_i, the template image m_i to which it is most similar. The raw images, $\{r_i\}$, have unknown in-plane translation and rotation parameters relative to the model projections $\{m_i\}$. It is therefore necessary to first solve the *alignment* problem; i.e., given two images m_i and r_j find a superposition such that the similarity between two images is maximized.

A similarity measure based on the dot product of image pixels is widely used in image processing. To integrate the dot product with the image alignment problem, the cross-correlation function (CCF) is used to describe the dot product optimization. In case of translations, the CCF is defined as:

$$\mathrm{CCF}(m_i, r_j, t) = \sum_k m_i(k) r_j(k+t),$$

where t is a translation applied to image r_j and k runs over the image pixels. For simplicity we omit discussion of the dot product optimization that involves rotations. A CCF variant, the normalized cross-correlation function (NCCF), is usually more appropriate for image comparison since image brightness may vary among the samples. NCCF is defined as:

$$\mathrm{NCCF}(m_i, r_j, t) = \sum_k \frac{(m_i(k) - \bar{m})(r_j(k+t) - \bar{r})}{\sigma_m \sigma_r},$$

where $\bar{m}$ and $\bar{r}$ are image mean values, and σ_m and σ_r are standard deviations. Mean and standard deviation values can be computed either based on all image pixels or only within the area where the dot product is being applied. In the later case the NCCF is called locally normalized cross-correlation function (LNCCF). One reason for the widespread use of the CCF and its variants is their computationally efficiency. Using the Fast Fourier Transform (FFT), the translation that maximizes the CCF can be found in $O(n\log(n))$, where n is the image size. The CCF and NCCF measure similarity between two images based on their pixel dot product. However, if we wish to apply a different scoring function, like mutual information used below, a naïve approach would be to perform an exhaustive search over all possible rotations and translations, which is computationally expensive.

We propose an alignment technique based on feature extraction that avoids an exhaustive search and is able to efficiently find a high scoring superposition between two images. Feature based alignment approaches are extensively used in computer vision (Forsyth and Ponce, 2002). Using features that can range from simple forms such as corners and edges to more complex structures such as lines, curves, object boundaries and areas, a superposition between two images is computed by alignment of the defined fea-