Journal of Structural Biology 190 (2015) 200–214

Contents lists available at ScienceDirect

Journal of Structural Biology

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

SubspaceEM: A fast maximum-a-posteriori algorithm for cryo-EM single particle reconstruction

Nicha C. Dvornek [a,*], Fred J. Sigworth [b,c], Hemant D. Tagare [a,b,d]

[a] Department of Diagnostic Radiology, Yale School of Medicine, New Haven, CT 06510, USA
[b] Department of Biomedical Engineering, Yale University, New Haven, CT 06520, USA
[c] Department of Cellular and Molecular Physiology, Yale School of Medicine, New Haven, CT 06510, USA
[d] Department of Electrical Engineering, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 23 March 2015
Accepted 25 March 2015
Available online 31 March 2015

Keywords:
Cryo-electron microscopy
Single particle reconstruction
Maximum-likelihood
Maximum-a-posteriori
Expectation–maximization algorithm
Fast image processing

ABSTRACT

Single particle reconstruction methods based on the maximum-likelihood principle and the expectation–maximization (E–M) algorithm are popular because of their ability to produce high resolution structures. However, these algorithms are computationally very expensive, requiring a network of computational servers. To overcome this computational bottleneck, we propose a new mathematical framework for accelerating maximum-likelihood reconstructions. The speedup is by orders of magnitude and the proposed algorithm produces similar quality reconstructions compared to the standard maximum-likelihood formulation. Our approach uses subspace approximations of the cryo-electron microscopy (cryo-EM) data and projection images, greatly reducing the number of image transformations and comparisons that are computed. Experiments using simulated and actual cryo-EM data show that speedup in overall execution time compared to traditional maximum-likelihood reconstruction reaches factors of over 300.

1. Introduction

Single particle reconstruction from cryo-electron microscopy (cryo-EM) is an iterative process that infers the 3D structure of a macromolecule or a "particle" from many noisy 2D projections of the particles lying at random orientations. Currently popular approaches to single particle reconstruction are based on the maximum-likelihood principle or the related maximum-a-posteriori principle. All these methods are optimized using the expectation–maximization (E–M) algorithm (Dempster et al., 1977), which iterates between the expectation step (E-step) and the maximization step (M-step). The E-step calculates latent probabilities for every possible alignment between particle image and structure projection. The M-step then uses these latent probabilities to give weights for alignments in calculating the reconstruction. E–M algorithms have gained popularity because they perform more favorably compared to other approaches for data with low signal-to-noise ratio (SNR).

The maximum-likelihood principle and related methods have been applied to many cryo-EM reconstruction problems. The first application of the maximum-likelihood principle in cryo-EM was for aligning particle images to a single 2D reference (Sigworth, 1998). Since then, several extensions have been proposed. The maximum-likelihood framework has been applied to 2D multireference image refinement (Scheres et al., 2005b), single particle reconstruction (Doerschuk and Johnson, 2000; Yin et al., 2001, 2003), and reconstruction of structurally heterogeneous data (Scheres et al., 2007). Another extension incorporates prior probability functions (priors) which encourage smoothness of the 3D structure, thereby limiting the amount of noise that propagates through to the reconstruction (Scheres, 2012a,b). More recent priors include enforcing a sparse representation by using an adaptive, non-Fourier basis, which improves the SNR and hence the resolution of the reconstruction (Kucukelbir et al., 2012). Maximum-likelihood with priors is referred to as the Bayesian or maximum-a-posteriori approach.

While reconstruction methods based on maximum-likelihood and the E–M algorithm have several desirable properties, from a computational point of view, the E–M algorithm is very slow (Cheng and Walz, 2009; Sigworth et al., 2010). The computational bottleneck of the E–M algorithm is the calculation of the latent probabilities in the E-step. In its naive form, the computational complexity of the E-step is the number of particle images × the number of projection directions × the number of image rotations

* Corresponding author. Fax: +1 203 737 4273.
E-mail address: nicha.dvornek@yale.edu (N.C. Dvornek).

http://dx.doi.org/10.1016/j.jsb.2015.03.009
1047-8477/

and translations; this much computation is prohibitive. For example, Scheres (2012b) reports that a straightforward E–M reconstruction of 5053 GroEL particles required almost 25 days, even while using 56 CPUs in parallel. As data sets are larger by an order of magnitude or more, such execution times are clearly unacceptable.

Several heuristics have been proposed to accelerate E–M reconstruction algorithms (Sigworth et al., 2010). All of these heuristics depend on the observation that latent probabilities tend to be peaky; that is, the latent probabilities are high for relatively few alignments and very low for the rest. If the high probability alignments can be found quickly, then the rest can be ignored, speeding up the calculations. An early strategy used this idea by calculating latent probabilities only for those alignments whose probabilities in the previous iteration exceed some threshold (Scheres et al., 2005a). A more sophisticated strategy utilizes ideas from adaptive integration, in which the probabilities are calculated on a coarse sampling and then refined only where most of the probability mass is concentrated (Tagare et al., 2008, 2010). This adaptive E–M algorithm can be combined with GPU implementation (Tagare et al., 2010) and local orientation searches (Scheres, 2012b) for further speedup.

Another strategy is to exploit known symmetries of the particle. For example, Lee et al. (2007) exploits icosahedral symmetry of viruses to accelerate E–M reconstructions. The approach utilizes a fixed basis of spherical harmonics to efficiently sample rotations (Doerschuk and Johnson, 2000). Furthermore, a linear transformation of the data is applied to speed up E–M calculations.

This paper proposes a completely novel heuristic for speeding up the E–M algorithm. The idea is to represent the particle images and structure projections in two different, relatively low-dimensional subspaces that are adapted to the data. This representation is accurate, so the projections can be rotated, translated, and compared with the particle images simply by performing the corresponding operations on the subspace bases. Because the number of basis elements is much smaller than the number of images and projections, substantial speedup is possible. The approximation using the subspaces is integrated into the maximum-a-posteriori framework. We refer to the proposed algorithm as *SubspaceEM*.

The SubspaceEM algorithm is developed below in the spatial domain. However, as will become clear in Section 5, the SubspaceEM algorithm applies without change to the Fourier domain, where equally significant speedups can be expected. This is important because some E–M algorithms are implemented in the spatial domain (e.g., ML3D in Xmipp (Sorzano et al., 2004)) while others are interpreted in the Fourier domain (e.g., RELION (Scheres, 2012b)). The subspace idea applies to both.

In the following, we begin by describing the mathematical derivations and implementation of the proposed SubspaceEM algorithm. After the discussion of the algorithm, experiments are presented comparing SubspaceEM with the classic E–M algorithm in reconstructing both simulated and real cryo-EM data. The computational cost of the SubspaceEM algorithm is compared with a straightforward implementation of the E–M algorithm. The purpose of these experiments is to examine the potential of the SubspaceEM algorithm to reduce computational costs by orders of magnitude without loss in the quality of 3D reconstruction.

2. Mathematical methods

2.1. The maximum-likelihood and maximum-a-posteriori formulations

We first briefly review the maximum-likelihood and maximum-a-posteriori formulations for single particle reconstruction. An extensive overview of the theory of maximum-likelihood methods for cryo-EM analysis is available in Sigworth et al. (2010).

Suppose that S is a particle structure, mathematically represented as a set of density values on a grid in a three-dimensional cube. The structure is projected along D directions and a set of F contrast transfer functions (CTFs), belonging to F defocus classes, are applied to each projection. This results in $M = D \times F$ "filtered projected" structures, each with a specific defocus class. Let $\mathcal{P}_j$ denote the $M = D \times F$ projection operators composed with CTF operators. Then, each filtered projected structure is given by $\mathcal{P}_j(S)$. Reconstruction algorithms, including the E–M algorithm, compare particle images with these $M = D \times F$ filtered projected structures, but the comparison of any particle image is only with the filtered projected structures that have the same defocus class as the image. To express this in the following mathematics, we tag each filtered projected structure $\mathcal{P}_j(S)$ with its defocus class, denoted by C_j. The value of C_j is an integer from $1, \ldots, F$. We emphasize that C_j is not the CTF function, but an integer which indexes the defocus class. In the following, to simplify the terminology, we will call $\mathcal{P}_j$ the projection operator and $\mathcal{P}_j(S)$ a structure projection whose defocus class is C_j.

Next, let x_i, $i = 1, \ldots, N$, be the cryo-EM particle images. Similar to the structure projections, each particle image x_i is tagged with its defocus class, denoted by C_i^*. The defocus class C_i^* is also an integer from $1, \ldots, F$ and is assigned according to the CTF measured from the micrograph. The cryo-EM image formation model is that x_i is a projected structure of the same defocus class, rotated and translated, and further corrupted by zero mean additive noise. The identity of the projection direction is lost in the image formation process. Letting z_i denote the index of the unknown projection operator relating the structure S to the image x_i, the image formation process is

$$x_i = \mathcal{T}_{\tau_i}(\mathcal{P}_{z_i}(S)) + n_i, \quad \text{with } C_i^* = C_{z_i}. \tag{1}$$

In the above equation, $\mathcal{T}_{\tau_i}$ is the 2D transformation operator which rotates and shifts the image according to the transformation parameter $\tau_i = \{\phi_i, t_{ix}, t_{iy}\}$, which specifies the in-plane rotation angle ϕ_i and the translations t_{ix} and t_{iy} along the x and y image axes. Further, n_i is the additive white Gaussian noise with zero mean and standard deviation σ. Finally, $C_i^* = C_{z_i}$ states that the image x_i can only arise due to the action of a projection operator $\mathcal{P}_{z_i}$ whose defocus class C_{z_i} is identical to the image defocus class C_i^*.

Under the image formation model in Eq. (1), the conditional probability density that image x_i comes from structure S is

$$p(x_i|S, \sigma, z_i, \tau_i) = \begin{cases} \mathcal{N}(\mathcal{T}_{\tau_i}(\mathcal{P}_{z_i}(S)), \sigma^2) & \text{if } C_i^* = C_{z_i} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $\mathcal{N}(\mathcal{T}_{\tau_i}(\mathcal{P}_{z_i}(S)), \sigma^2)$ is the probability density function of a Gaussian random variable with mean $\mathcal{T}_{\tau_i}(\mathcal{P}_{z_i}(S))$ and variance σ^2.

The variables z_i and τ_i are nuisance or *latent variables* because their values are unknown and are not of particular interest. These variables are eliminated by marginalization, that is, by integrating them out. Let $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$, where α_j is the probability that $z_i = j$, let Ω_{τ_i} be the domain of possible values for the transformation parameter τ_i, and let $p(\tau_i)$ be the probability of τ_i. Then, marginalization results in

$$\begin{aligned} p(x_i|S, \sigma, \boldsymbol{\alpha}) &= \sum_{j=1}^{M} \int_{\Omega_{\tau_i}} p(x_i|S, \sigma, j, \tau_i) p(z_i = j) p(\tau_i) d\tau_i \\ &= \sum_{j=1}^{M} \alpha_j \int_{\Omega_{\tau_i}} p(x_i|S, \sigma, j, \tau_i) p(\tau_i) d\tau_i, \end{aligned} \tag{3}$$

where the sum over j marginalizes z_i and the integral marginalizes τ_i. Simplifying the notation by setting $\Theta = \{S, \sigma, \boldsymbol{\alpha}\}$ gives