Journal of Structural Biology 180 (2012) 519–530

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

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

RELION: Implementation of a Bayesian approach to cryo-EM structure determination

Sjors H.W. Scheres

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

ARTICLE INFO

Article history:
Received 24 July 2012
Received in revised form 3 September 2012
Accepted 6 September 2012
Available online 19 September 2012

Keywords:
Electron microscopy
Single-particle analysis
Maximum likelihood
Image processing
Software development

ABSTRACT

RELION, for REgularized LIkelihood OptimizatioN, is an open-source computer program for the refinement of macromolecular structures by single-particle analysis of electron cryo-microscopy (cryo-EM) data. Whereas alternative approaches often rely on user expertise for the tuning of parameters, RELION uses a Bayesian approach to infer parameters of a statistical model from the data. This paper describes developments that reduce the computational costs of the underlying maximum a posteriori (MAP) algorithm, as well as statistical considerations that yield new insights into the accuracy with which the relative orientations of individual particles may be determined. A so-called gold-standard Fourier shell correlation (FSC) procedure to prevent overfitting is also described. The resulting implementation yields high-quality reconstructions and reliable resolution estimates with minimal user intervention and at acceptable computational costs.

1. Introduction

Macro-molecular structure determination by single-particle analysis of electron cryo-microscopy (cryo-EM) images is a rapidly evolving field. Over the past two decades many reconstructions that reveal secondary structure elements have been obtained, e.g. see (Boettcher et al., 1997; Lau and Rubinstein, 2012; Lander et al., 2012), and recently several reconstructions to near-atomic resolution have been reported (Wolf et al., 2010; Liu et al., 2010; Yang et al., 2012). Improvements in electron microscopes and better computational tools for image processing have been important contributors to these successes. Moreover, on-going hardware developments such as direct-electron detectors (Milazzo et al., 2011; Brilot et al., 2012; Bammes et al., 2012) and phase-plates (Nagayama, 2011; Barton et al., 2011; Fukuda et al., 2012) are expected to improve data quality even further in the near future. This is likely to enhance the applicability of cryo-EM structure determination, as less noisy images will allow the visualization of smaller macro-molecular complexes.

The increased applicability of the technique is expected to attract new researchers to the field. Because conventional data collection and processing procedures often rely on user expertise, the needs for improved ease-of-use and automation are now widely recognized. More convenient data collection schemes are being developed through a combination of automated data acquisition software (Suloway et al., 2005) and improvements in the latest generation electron microscopes (Shrum et al., in press; Fischer et al., 2010). To cope with the large amounts of data from these experiments, semi-automated image processing pipelines and dedicated electronic notebooks have been proposed (Lander et al., 2009; Ludtke et al., 2003). Continuing developments in these areas are expected to increase the accessibility of cryo-EM structure determination to inexperienced users.

However, many cryo-EM projects still suffer from important hurdles in image processing that cannot be overcome by automation and increased volumes of data alone. Existing image processing procedures often comprise a concatenation of multiple steps, such as particle alignment, class averaging, reconstruction, resolution estimation and filtering. Many of these steps involve the tuning of specific parameters. Whereas appropriate use of these procedures may yield useful results, suboptimal parameter settings or inadequate combinations of the separate steps may also lead to grossly incorrect structures, thus representing a potential pitfall for newcomers to the field.

Recently, I described a Bayesian approach to cryo-EM structure determination, in which the reconstruction problem is expressed as the optimization of a single target function (Scheres, 2012). In particular, the reconstruction problem is formulated as finding the model that has the highest probability of being the correct one in the light of both the observed data and available prior information. Optimization of this posterior distribution is called maximum a posteriori (MAP), or regularized likelihood optimization. The Bayesian interpretation places the cryo-EM structure determination process on a firm theoretical basis, where explicit statistical assumptions about the model and the data, as well as the optimi-

E-mail address: scheres@mrc-lmb.cam.ac.uk

1047-8477/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2012.09.006

zation strategy itself, can be discussed and improved if deemed necessary. Whereas conventional refinement procedures employ many *ad hoc* parameters that need to be tuned by an expert user, the Bayesian approach iteratively learns most parameters of the statistical model from the data themselves.

This paper describes the implementation of the Bayesian approach to single-particle reconstruction in the stand-alone computer program RELION, which stands for REgularized LIkelihood OptimizatioN. The theoretical implications of the statistical approach represent a huge challenge for its implementation in a useful computer program. Various algorithmic developments are described that allow MAP optimization of single-particle reconstructions at an acceptable computational cost. Moreover, the theoretical framework provided by the Bayesian approach may yield valuable insights into outstanding questions. As an example of this, I will describe an approach that uses the statistical data model to estimate the accuracy with which individual particles may be aligned and to quantify the contribution of different frequencies to this. Finally, because in principle some degree of overfitting might still go by unnoticed in the previously proposed MAP optimization approach (Scheres, 2012), a new procedure is described that eradicates the possibility of overfitting by the use of so-called "gold-standard" FSC calculations (Henderson et al., 2012; Scheres et al., 2012). Application of RELION to both simulated and experimental data illustrates that reconstructions that are free from overfitting may be obtained in a highly objective manner, without compromising reconstruction quality and at acceptable computational costs.

2. Approach

2.1. Theoretical background

MAP refinement of cryo-EM single-particle reconstructions is based on the following linear model in Fourier space:

$$X_{ij} = \mathrm{CTF}_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl} + N_{ij}, \tag{1}$$

where:

- X_{ij} is the jth component, with $j = 1,\ldots,J$, of the 2D Fourier transform $\mathbf{X}_i$ of the ith experimental image, with $i = 1,\ldots,N$.
- CTF_{ij} is the jth component of the contrast transfer function for the ith image.
- V_{kl} is the lth component, with $l = 1,\ldots,L$, of the 3D Fourier transform $\mathbf{V}_k$ of the kth of K underlying structures in the data set. Multiple structures K may be used to describe structural heterogeneity in the data, and K is assumed to be known. All components V_{kl} are assumed to be independent, zero-mean, and Gaussian distributed with variance τ_{kl}^2.
- $\mathbf{P}^{\phi}$ is a $J \times L$ matrix of elements $\mathbf{P}_{jl}^{\phi}$. The operation $\sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}$ for all j extracts a slice out of the 3D Fourier transform of the kth underlying structure, and ϕ defines the orientation of the 2D Fourier transform with respect to the 3D structure, comprising a 3D rotation and a phase shift accounting for a 2D origin offset in the experimental image. Similarly, the operation $\sum_{j=1}^{J} \mathbf{P}_{lj}^{\phi^{\mathrm{T}}} X_{ij}$ for all l places the 2D Fourier transform of an experimental image back into the 3D transform.
- N_{ij} is noise in the complex plane, which is assumed to be independent, zero-mean, and Gaussian distributed with variance σ_{ij}^2.

Imagining an ensemble of possible solutions, the reconstruction problem is formulated as finding the model with parameter set Θ that has the highest probability of being the correct one in the light of both the observed data $\mathcal{X}$ and the prior information $\mathcal{Y}$. According to Bayes' law, this so-called posterior distribution factorizes into two components:

$$P(\Theta|\mathcal{X},\mathcal{Y}) \propto P(\mathcal{X}|\Theta,\mathcal{Y}) P(\Theta|\mathcal{Y}) \tag{2}$$

where the *likelihood* $P(\mathcal{X}|\Theta,\mathcal{Y})$ quantifies the probability of observing the data given the model, and the *prior* $P(\Theta|\mathcal{Y})$ expresses how likely that model is given the prior information. The likelihood is computed based on the assumption of independent, zero-mean Gaussian noise in the images, and one marginalizes over the orientations ϕ and class assignments k. The variance σ_{ij}^2 of the noise components is unknown and will be estimated from the data. Variation of σ_{ij}^2 with resolution allows the description of non-white, or coloured noise. The prior is based on the assumption that the Fourier components of the signal are also independent, zero-mean and Gaussian distributed with unknown and resolution-dependent variance τ_{kl}^2 (see Scheres, 2012 for more details). The model $\hat{\Theta}$, including all V_{kl}, σ_{ij}^2 and τ_{kl}^2, that optimizes the posterior distribution $P(\Theta|\mathcal{X},\mathcal{Y})$ is called the *maximum a posteriori* (MAP) estimate. Note that previously discussed ML methods in the Fourier domain (Scheres et al., 2007b) aimed to optimize $P(\mathcal{X}|\Theta,\mathcal{Y})$.

Optimisation of $P(\Theta|\mathcal{X},\mathcal{Y})$ may be achieved by the expectation–maximization algorithm (Dempster et al., 1977), in which case the following iterative algorithm is obtained:

$$V_{kl}^{(\mathrm{n}+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{ik\phi}^{(\mathrm{n})} \sum_{j=1}^{J} \mathbf{P}_{lj}^{\phi^{\mathrm{T}}} \frac{\mathrm{CTF}_{ij} X_{ij}}{\sigma_{ij}^{2\,(\mathrm{n})}} d\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{ik\phi}^{(\mathrm{n})} \sum_{j=1}^{J} \mathbf{P}_{lj}^{\phi^{\mathrm{T}}} \frac{\mathrm{CTF}_{ij}^2}{\sigma_{ij}^{2\,(\mathrm{n})}} d\phi + \frac{1}{\tau_{kl}^{2\,(\mathrm{n})}}}, \tag{3}$$

$$\sigma_{ij}^{2\,(\mathrm{n}+1)} = \frac{1}{2} \sum_{k=1}^{K} \int_{\phi} \Gamma_{ik\phi}^{(\mathrm{n})} \left| X_{ij} - \mathrm{CTF}_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}^{(\mathrm{n})} \right|^2 d\phi, \tag{4}$$

$$\tau_{kl}^{2\,(\mathrm{n}+1)} = \frac{1}{2} \left| V_{kl}^{(\mathrm{n}+1)} \right|^2, \tag{5}$$

where $\Gamma_{ik\phi}^{(\mathrm{n})}$ is the posterior probability of class assignment k and orientation assignment ϕ for the ith image, given the model at iteration number (n). It is calculated as follows:

$$\Gamma_{ik\phi}^{(\mathrm{n})} = \frac{P(\mathbf{X}_i|k,\phi,\Theta^{(\mathrm{n})},\mathcal{Y}) P(k,\phi|\Theta^{(\mathrm{n})},\mathcal{Y})}{\sum_{k'=1}^{K} \int_{\phi'} P(\mathbf{X}_i|k',\phi',\Theta^{(\mathrm{n})},\mathcal{Y}) P(k',\phi'|\Theta^{(\mathrm{n})},\mathcal{Y}) \mathrm{d}\phi'}, \tag{6}$$

with:

$$P(\mathbf{X}_i|k,\phi,\Theta^{(\mathrm{n})},\mathcal{Y}) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^{2\,(\mathrm{n})}} \exp\left(\frac{\left| X_{ij} - \mathrm{CTF}_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}^{(\mathrm{n})} \right|^2}{-2\sigma_{ij}^{2\,(\mathrm{n})}} \right), \tag{7}$$

and $P(k,\phi|\Theta^{(\mathrm{n})},\mathcal{Y})$ may be used to express prior information about the distribution of the hidden variables k and ϕ. In practice, the integrations over ϕ are replaced by (Riemann) summations over discretely sampled orientations, and translations are limited to a user-defined range. Also, the power of the signal, τ_{kl}^2, and of the noise, σ_{ij}^2, are estimated as 1D vectors, varying only with the resolution of Fourier components j and l.

The iterative algorithm in Eqs. (3)–(7) is started from an initial estimate for $\mathbf{V}_k$: the starting model. If $K > 1$, multiple different starting models are obtained by random division of the data set in the first iteration. The user controls the number of models K that is to be refined simultaneously. Initial estimates for τ_{kl} and σ_{ij} are calculated from the power spectra of the starting model and individual particles, respectively.

It is important to note that the algorithm outlined above is a local optimizer. Thereby, the outcome of the refinement depends on the suitability of the starting model, and grossly incorrect starting models may lead to suboptimal results. Typically, to reduce