Journal of Structural Biology 188 (2014) 267–273

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Technical Note

Particle migration analysis in iterative classification of cryo-EM single-particle data

Bo Chen [a,1], Bingxin Shen [b,1], Joachim Frank [a,b,*]

[a] Department of Biology, Columbia University, New York, NY 10027, USA
[b] Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 11 October 2014
Accepted 15 October 2014
Available online 30 October 2014

Keywords:
Cryogenic electron microscopy
Single-particle reconstruction
Iterative classification
Class number
Bayesian agglomerative clustering
Maximum a posteriori
Ribosome

ABSTRACT

Recently developed classification methods have enabled resolving multiple biological structures from cryo-EM data collected on heterogeneous biological samples. However, there remains the problem of how to base the decisions in the classification on the statistics of the cryo-EM data, to reduce the subjectivity in the process. Here, we propose a quantitative analysis to determine the iteration of convergence and the number of distinguishable classes, based on the statistics of the single particles in an iterative classification scheme. We start the classification with more number of classes than anticipated based on prior knowledge, and then combine the classes that yield similar reconstructions. The classes yielding similar reconstructions can be identified from the migrating particles (jumpers) during consecutive iterations after the iteration of convergence. We therefore termed the method "jumper analysis", and applied it to the output of RELION 3D classification of a benchmark experimental dataset. This work is a step forward toward fully automated single-particle reconstruction and classification of cryo-EM data.

1. Introduction

Over the past three decades, cryogenic electron microscopy (cryo-EM) and single-particle three-dimensional (3D) reconstruction techniques have evolved into a powerful toolbox for determining biological macromolecular structures. There have been advances in computational tools, e.g. automated data collection (Suloway et al., 2005; Mastronarde, 2005) and the design of an image processing pipeline (Lander et al., 2009; Langlois et al., 2014; Scheres, 2012a), as well as significant improvements in hardware, particularly the direct electron detector (Milazzo et al., 2011; Bammes et al., 2012). Single-particle reconstruction for cryo-EM often deals with a large number of 2D images of biological macromolecules ("particles"). It employs automated particle selection, 2D alignment (transformation of the images, including in-plane rotation and translation, to bring them into optimal superimposition), and an iterative process of projection alignment and 3D reconstruction. In its original form (1990s–2000s), single-particle reconstruction required a homogeneous sample, in which all particles represent identical copies of the macromolecule (see Frank, 2010). Because the signal-to-noise ratio (SNR) of individual particles is low, it needs to be increased by averaging multiple particles, but averaging is applicable only if the particles represent the same view of replicas of macromolecules. Therefore, the original single-particle reconstruction method was limited by sample heterogeneity, i.e., particles in a sample representing compositionally and conformationally different biological macromolecules (Frank, 2006).

Recently developed classification methods have enabled resolving multiple structures/conformations of the macromolecules from cryo-EM data obtained from heterogeneous biological samples (e.g. Scheres et al., 2007; Fischer et al., 2010; Agirrezabala et al., 2012). The classification methods can be divided into two categories, supervised and unsupervised methods (Frank, 2006). Supervised classification utilizes two or more 3D density maps as references, and separates the particles based on their similarities to these references. Therefore, the supervised classification methods pose the danger of *reference bias*: In the extreme case, images of mere noise can result in an averaged image resembling the reference (Shaikh et al., 2008). Unsupervised classification, in contrast, groups the particles based on their mutual similarities. Although a low-resolution 3D map may be needed for the initial 2D alignment, unsupervised classification methods are largely immune to the reference bias problem.

* Corresponding author at: 650 West 168th Street, Black Building 221, New York, NY 10032, USA. Fax: +1 212 305 9500.
E-mail address: jf2192@columbia.edu (J. Frank).
[1] These authors contributed equally to this work and should be considered co-first authors.

http://dx.doi.org/10.1016/j.jsb.2014.10.006
1047-8477/

One important improvement in the classification methods is treating the class assignment of each particle (i.e., which structure/conformation of the macromolecule a particle represents) as a probability distribution among classes, instead of making an all-or-none class assignment. This idea forms the basis of the maximum-likelihood (ML)-based classification methods (Yin et al., 2003; Scheres et al., 2005, 2007, 2009). ML methods aim to find the values of a set of parameters which maximize the likelihood function of the parameters, given the observed data. In the case of cryo-EM and single-particle reconstruction, these parameters include each of the voxels of the 3D density map for each class, the class assignment, the projection angle relative to the 3D map, and the 2D rotation and translation of each particle image. The ML estimator is an intuitive and popular point estimator. However, because cryo-EM datasets are finite in size, noisy (SNR $\sim$ 0.1 for low-dose exposure) (Baxter et al., 2009) and lack the projection angle information, the multi-parameter ML estimator for cryo-EM data is susceptible to the *over-fitting* problem, i.e., treating noise as signal erroneously (Scheres, 2012b).

To address the over-fitting problem, one can instead use the *maximum a posteriori* (MAP) estimator, a Bayesian approach to statistics. MAP estimation considers the experimenter's belief (prior knowledge) as well as the likelihood function of the parameters. In Bayesian statistics, the set of parameters is considered a quantity subject to variation that can be described by a probability distribution, called the *prior distribution*. The prior distribution is updated in light of the observed data to yield the *posterior distribution*, which is proportional to the product of the likelihood function and the prior distribution. As the size of the observed data increases, the MAP estimator gives more weight to the sample information, and less weight to the prior information (Casella and Berger, 2001).

The Expectation–Maximization algorithm is particularly well suited to find the ML/MAP estimator for a mathematically *incomplete* problem such as 3D reconstruction and classification of cryo-EM data, because cryo-EM data lack the information of class assignment and the projection angle for every particle. The Expectation–Maximization algorithm is based on the idea of alternately optimizing the set of parameters and the set of missing data (or hidden variables), while fixing the values of the other set. This optimization is performed iteratively, with its limit being the ML/MAP estimator for the original problem (Casella and Berger, 2001). The Expectation–Maximization algorithm for cryo-EM classification and 3D reconstruction has been implemented in ML3D (Scheres et al., 2007) and MLn3D (n stands for normalization) (Scheres et al., 2009) to find the ML estimator, and in RELION (REgularized Likelihood OptimizatioN) (Scheres, 2012b,a) to find the MAP estimator.

It has remained a question how to base the decisions made in the course of classification on the statistics of the data. The above-mentioned classification methods all can possibly give reliable solutions, if performed properly. The pitfall, however, is that they all involve various amounts of *subjective decisions* made by researchers with various degrees of experience. Subjective decisions may be involved in many steps of 3D classification, such as particle selection, particle alignment, 3D reconstruction, and filtering. The employment of subjective decisions can limit the use of these methods by inexperienced researchers, and should therefore be minimized. RELION has set a good example for reducing user discretion (Scheres, 2012b), although the user is still responsible for choosing the number of classes, number of iterations of the Expectation–Maximization algorithm, and the initial reference volume.

In this work, to further curb the role of subjective decisions in classification, we propose the jumper analysis based on the statistics of cryo-EM particles, to determine the iteration of convergence and the number of distinguishable classes. Specifically, the iteration of convergence, i.e., from which point onwards the 3D reconstructions become trustworthy and stable for user examination, is indicated primarily based on the probability distribution of all the particles over the iterations. The classes yielding similar 3D reconstructions are indicated by the migration behavior of the particles, i.e., change in class assignment, that occurs after the iteration of convergence. As we will show, this migration information can provide reliable criteria for determining which classes of particles represent the same conformation of the biological macromolecule, and can therefore be combined to obtain a better 3D reconstruction. We demonstrate the jumper analysis method by using the output of RELION classification on a well-characterized experimental cryo-EM dataset. Evidently, this analysis method can also be applied to other iterative classification schemes, e.g. the iterative classification algorithm implemented in FREALIGN (Lyumkis et al., 2013).

2. Methods

2.1. Image formation model for 3D reconstruction and classification

Assume we collected N particles from a heterogeneous cryo-EM sample containing K structures. These K structures, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K$, differ from one another in composition and/or in conformation. Each particle $\mathbf{x}_i, i = 1, 2, \ldots, N$, is a 2D projection of one of the 3D structures $\mathbf{v}_{k_i}, k_i \in \{1, 2, \ldots, K\}$. According to the weak-phase-object approximation (WPOA), the image formation model in Fourier Space is:

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

where:

- $\mathbf{X}_i$ is the 2D Fourier transform of $\mathbf{x}_i, i = 1, 2, \ldots, N$. X_{ij} is the j-th component of $\mathbf{X}_i, j = 1, 2, \ldots, J$ and $J = D^2$. D is the number of pixels in one dimension.
- CTF_{ij} is the j-th component of the contrast transfer function (CTF) of particle X_i, which is assumed to be constant in the 3D classification and reconstruction process.
- $\mathbf{V}_{k_i}$ is the 3D Fourier transform of the 3D structure $\mathbf{v}_{k_i}, k_i \in \{1, 2, \ldots, K\}$. $V_{k_i l}$ is the l-th component of $\mathbf{V}_{k_i}, l = 1, 2, \ldots, L$ and $L = D^3$.
- The operation $\sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi_i} V_{k_i l}$ for all j extracts a central slice of $\mathbf{V}_{k_i}$ at orientation ϕ_i. According to the projection-slice theorem, this operation is equivalent to the real-space projection operation. ϕ_i is the Fourier-space equivalent of the real-space position parameter, which comprises the 3D rotation and 2D translation of particle $\mathbf{x}_i$ relative to $\mathbf{v}_{k_i}$ in real space. For convenience, ϕ_i is referred to as the orientation of $\mathbf{X}_i$.
- N_{ij} is the noise in Fourier space. Commonly used Wiener filters assume that the noise is independent and Gaussian distributed with mean 0 and variance σ_{ij}^2.

The goal of 3D classification and reconstruction is to find a solution for the model of 3D structures with parameter set $\mathbf{\Theta} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K\}$, given the observed data $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$. In Bayesian statistics, we find the MAP estimate of the parameter set $\hat{\mathbf{\Theta}}_{MAP}$ which maximizes the regularized likelihood:

$$\hat{\mathbf{\Theta}}_{MAP} = \arg\max_{\mathbf{\Theta}} p(\mathbf{X}|\mathbf{\Theta}) p(\mathbf{\Theta}), \tag{2}$$

where $p(\mathbf{X}|\mathbf{\Theta})$ is the conditional probability, or likelihood, of observing the data $\mathbf{X}$ given the parameter set $\mathbf{\Theta}$. $p(\mathbf{\Theta})$ is the prior probability distribution of the parameter set.

However, 3D classification and reconstruction is a mathematically *incomplete* problem, because cryo-EM data lack the information on class assignment and the orientation of every particle. Such missing information can be treated as *hidden variables*, and integrated out in the solution of the model. Let $\boldsymbol{\phi} = \{\phi_1, \phi_2, \ldots, \phi_N\}$ denote the orientation information of $\mathbf{X}$. Let $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ denote the class assignment of each particles, where $z_i = k$ if $\mathbf{x}_i$ is a projection of $\mathbf{v}_k$. The true values of $(\mathbf{z}, \boldsymbol{\phi})$, denoted as $(\tilde{\mathbf{z}}, \tilde{\boldsymbol{\phi}})$, are unknown. Furthermore, we treat the class assignment of each particle as a probability distribution among all K classes, instead of an all-or-none class assignment. Let $c_{ik} = p(z_i = k|\mathbf{X}_i, \boldsymbol{\Theta})$ denote the conditional probability of $\mathbf{x}_i$ being a projection of $\mathbf{v}_k$ given the parameter set $\boldsymbol{\Theta}$, where $c_{ik} \geqslant 0$ and $\sum_{k=1}^{K} c_{ik} = 1$. We define the most-probable class assignment for a particle $\mathbf{X}_i$ as:

$$\hat{z}_i = \arg\max_k(c_{ik}). \tag{3}$$

We can also estimate the probability of any one particle being a projection of $\mathbf{v}_k$ using $\hat{c}_k = \hat{p}(z = k|\boldsymbol{\Theta}) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{z}_i = k)$, where the indicator function $I(\hat{z}_i = k) = 1$, if $\hat{z}_i = k$; or 0, if $\hat{z}_i \neq k$.

2.2. Iterative classification and convergence

For a mathematically incomplete problem like 3D classification and reconstruction, the Expectation–Maximization algorithm is suited to find its MAP estimator. In each iteration, this algorithm alternately optimizes the parameter set $\boldsymbol{\Theta}$ and the set of hidden variables $\{\mathbf{z}, \boldsymbol{\phi}\}$, while fixing the values of the other set. Generally, this algorithm converges to a local optimum, which depends on the initial values of the parameters. RELION has successfully implemented the Expectation–Maximization algorithm to find the MAP estimator $\hat{\boldsymbol{\Theta}}_{MAP}$ (Scheres, 2012b,a). The reader is referred to these papers and our summary of the method (Shen et al., 2014) for the detailed explanation of the algorithm. Thereafter we use the output of RELION 3D classification to demonstrate our jumper analysis of iterative classification.

We define the *iteration of convergence* as the smallest number of iterations from which onwards the 3D classification and reconstruction results become stable. In practice, we look for the iteration after which the likelihood function $p(\mathbf{X}|\boldsymbol{\Theta})$ reaches a plateau, because the optimization criterion is $p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$, and $p(\boldsymbol{\Theta})$ is negligible compared to $p(\mathbf{X}|\boldsymbol{\Theta})$ with a large dataset. Notably, after the iteration of convergence, the orientation distribution for most particles should be close to the delta function, so only a few of orientations need to be considered when calculating the particle statistics (Scheres, 2012a).

2.3. Agglomerative classification after the iteration of convergence

In this section, we propose a way to identify classes that should be combined by analyzing the change in most-probable class assignment of the particles after the iteration of convergence. The rationale is the following: if there is a sizeable portion of the particles commuting between two classes after the iteration of convergence, then these two classes are likely to have similar 3D reconstructions and can be combined. This prediction is validated both by visually examining the 3D reconstructions and calculating the difference map between each pair of 3D reconstructions. The particles in the classes with similar reconstructions can then be combined to obtain a better-quality reconstruction. We use a simplified, ideal situation to explain the rationale of this approach.

2.3.1. Ideal situation

Assume at iteration t after the iteration of convergence t^*, the 3D reconstructions from the K classes, $\mathbf{V}_{1,t}, \mathbf{V}_{2,t}, \ldots, \mathbf{V}_{K,t}$ are distinguishable from each other at the current resolution. Then each particle $\mathbf{X}_i$ has a distinct class assignment, $\hat{z}_{i,t} = \tilde{z}_i$, and $c_{ik,t} = 1$, if $k = \tilde{z}_i$; or 0, if $k \neq \tilde{z}_i$. Thus $\mathbf{X}_i$ only contributes to one reconstruction $\mathbf{V}_{\tilde{z}_i,t}$. Let $G_{k,t} = \{\mathbf{X}_i|\hat{z}_{i,t} = k\}$ denote all the particles with the most-probable class assignment being k at iteration t. Then the fraction population of class k is $\hat{c}_{k,t} = \#\ G_{k,t}/N$, where $\#\ G_{k,t}$ denotes the number of elements of $G_{k,t}$. Furthermore, if each of the K classes is homogenous, then the selected number of classes K equals the maximum number of distinguishable classes K^* in the dataset.

2.3.2. Situation with jumper particles

Since we do not know K^* from the start, in practice we usually perform the 3D classification multiple times, each time starting with the number of classes that is supposed to be greater than or equal to K^*. When the selected number of classes K is greater than K^*, at iteration t after t^*, there will be at least two classes with reconstructions indistinguishable at the current resolution, apart from translation and rotation offsets. For simplicity, we consider the situation where only two classes, $r, s \in \{1, 2, \ldots, K\}, r < s$, have identical reconstructions, and the other $(K-2)$ classes have reconstructions distinguishable from the rest classes. Then for $X_i \in \{X_i|\tilde{z}_i = r \text{ or } \tilde{z}_i = s\}, c_{ir,t} + c_{is,t} = 1$, and $c_{ik,t} = 0$, if $k \neq r, s$.

Looking at consecutive iterations after t^*, if $\hat{z}_{i,t} \neq \hat{z}_{i,t+1}$, we call $\mathbf{X}_i$ a *jumper* particle from class $\hat{z}_{i,t}$ at iteration t to class $\hat{z}_{i,t+1}$ at iteration $(t+1)$. We can approximate the most-probable class assignment $\hat{z}_{i,t}$ by using $(\hat{z}_{i,t}, \hat{\phi}_{i,t}) \approx \arg\max_{k,\phi} p(z_i = k, \phi|\mathbf{X}_i, \boldsymbol{\Theta})_t$, which is an output of the RELION program. This is because after the iteration of convergence, the orientation distribution for most particles is close to the delta function, and therefore $c_{ik,t} = \int_\phi p(z_i = k, \phi|\mathbf{X}_i, \boldsymbol{\Theta})_t \mathrm{d}\phi \approx \max_{k,\phi} p(z_i = k, \phi|\mathbf{X}_i, \boldsymbol{\Theta})_t$.

Furthermore, we can generalize the jumper particle analysis to multiple iterations in tandem after the iteration of convergence, and among any pair of classes. For iterations t_1 through $t_2, t^* \leqslant t_1 < t_2$, let $\hat{c}_{ik,t_1\sim t_2} = \sum_{t=t_1}^{t_2} I(\hat{z}_{i,t} = k)/(t_2 - t_1 + 1)$, then $\hat{c}_{ik,t_1\sim t_2}$ is another estimate of c_{ik} that is less dependent on the choice of iteration t.

Let $G_{r\to s,t} = \{\mathbf{X}_i|\hat{z}_{i,t} = r \text{ and } \hat{z}_{i,t+1} = s\}$ denote all the particles with the most-probable class assignment being r at iteration t AND being s at iteration $(t+1), r, s \in \{1, 2, \ldots, K\}, t \geqslant t^*$. We use a *transition matrix*, defined as $TM_{t_1\sim t_2} = \sum_{t=t_1}^{t_2} M_t/(t_2 - t_1 + 1)$, to indicate the probability of having jumper particles in each class, where the element of the $K \times K$ matrix $M_t, m_{sr,t} = \#\ G_{r\to s,t}/\#\ G_{r,t}$. In practice, the transition matrix is a sparse matrix, i.e., most of the elements off the diagonal have values close to zero, because most classes are distinguishable from the others.

We then rearrange $TM_{t_1\sim t_2}$ into $A_{t_1\sim t_2}$, a $K \times K$ sparse matrix with approximate minimum degrees using an agglomerative method (Amestoy et al., 1996, 2004). The class numbers $[1, 2, \ldots, K]$ in $TM_{t_1\sim t_2}$ are reordered into $[q_1, q_2, \ldots, q_K]$ in $A_{t_1\sim t_2}$, where $q_i \in \{1, 2, \ldots, K\}, i = 1, 2, \ldots, K$. After rearrangement, the classes that share a sizeable portion of jumper particles can be distinguished (Algorithm 1) by choosing an empirical cutoff value of 1/3 (i.e., the number of jumper particles is half the number of particles staying in these two classes), and these classes may yield similar reconstructions and be combined. Moreover, we can reduce the dimension of transition matrix data into a 2D bar diagram, where $B_{r\to s,t_1\sim t_2} = \sum_{t=t_1}^{t_2} \#\ G_{r\to s,t}/(t_2 - t_1 + 1)$, to visually examine the effectiveness of grouping the classes that have commuting jumper particles.

Algorithm 1. Identify classes that share a sizeable portion of jumper particles.

Data: $A_{t_1\sim t_2}, [q]$.
Result: All the distinct groups of classes $\{all\_grp\}$.
$curr\_grp \leftarrow \{q_K\}$;
$r \leftarrow K$;
while $r > 1$ **do**
  $b \leftarrow (a_{r,r} + a_{r-1,r-1} - a_{r,r-1} - a_{r-1,r})/(a_{r,r} + a_{r-1,r-1} + a_{r,r-1} + a_{r-1,r})$;
  if $b < cutoff$ **then**
    $curr\_grp \leftarrow \{curr\_grp, q_{r-1}\}$;
    $a_{r-1,r-1} \leftarrow a_{r,r} + a_{r-1,r-1} + a_{r,r-1} - a_{r-1,r}$;
    $a_{r-1,i} \leftarrow a_{r-1,i} + a_{r,i}, i = 1, 2, \ldots, r-2$;
    $a_{i,r-1} \leftarrow a_{i,r-1} + a_{i,r}, i = 1, 2, \ldots, r-2$;
  else
    $all\_grp \leftarrow \{all\_grp, curr\_grp\}$;
    $curr\_grp \leftarrow \{q_{r-1}\}$;
  end
  $r \leftarrow r - 1$;
end

3. Results and discussion

3.1. Benchmark experimental data of 70S ribosome

We used a standard benchmark cryo-EM dataset of 70S ribosome (Baxter et al., 2009) to illustrate the procedure of the jumper analysis. The benchmark dataset contains 10,000 ribosome particles, among which 5000 were classified by supervised classification as 70S ribosome containing elongation factor G (EF-G), and the other 5000 as 70S ribosome containing no EF-G in the original work. The 70S ribosome containing no EF-G was observed to be in a classical, non-rotated global conformation, and contains three tRNAs in the aminoacyl (A), peptidyl (P), and exit (E) sites. The 70S ribosome containing EF-G was observed to be in a rotated global conformation (i.e., the small subunit (30S) is rotated relative to the large subunit (50S), compared to the non-rotated global conformation), and also contains an E-site tRNA. Scheres performed RELION 3D classification on this benchmark dataset, and discovered a small class of particles representing the 50S subunit, demonstrating the intrinsic capability of non-supervised classification method, such as RELION, to detect unanticipated class(es) in a heterogeneous dataset (Scheres, 2012b).

We chose $K = 6$ as the number of classes to start the RELION 3D classification, because we knew that there are at least three classes: 50S subunit, 70S containing EF-G, and 70S containing three tRNAs but no EF-G, and wanted to be able to accommodate potential new classes. We ran the classification for 60 iterations to demonstrate how to determine the iteration of convergence.

3.2. Determining the iteration of convergence

The purpose of determining the iteration of convergence is twofold: (1) to determine when to stop the iterative classification, and (2) to save the effort of manually examining the reconstructions of all classes from every iteration. We first note that the sum of likelihood functions of all classes reaches a plateau after a certain iteration, in this example iteration 24 (Fig. 1a). In practice, we wrote a MATLAB function to examine the change of the sum of likelihood function, to determine the iteration of convergence. If the change is less than 5% of the sum of likelihood function for 5 consecutive iterations, we deem that the classification has converged at the first of the five consecutive iterations. In addition, the likelihood function of each class is less ideal than the sum of likelihood functions of all classes for determining the iteration of convergence, because the likelihood function of different classes may reach a plateau after different iterations, and have more fluctuations than the sum of likelihood (Fig. 1b).

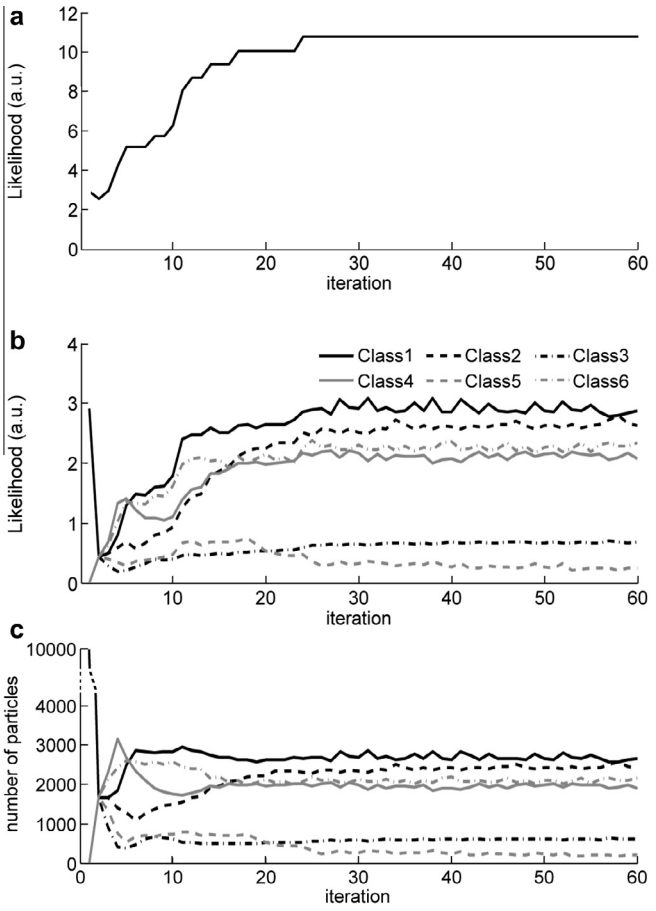

Fig.1. Likelihood function and number of particles as a function of iteration number. (a) The sum of likelihood functions of all six classes in each iteration. The value of likelihood is in arbitrary unit (a.u.). Note that the sum of likelihood functions of all classes reaches a plateau after iteration 24. (b) The likelihood function of each class in each iteration. (c) The number of particles of each class in each iteration. Label scheme is the same as in panel (b).

3.3. There are jumper particles after the iteration of convergence

The number of particles of each class may fluctuate even after the iteration of convergence, as exemplified in Fig. 1c. After iteration 24, Class 5 and Class 3 contain about 200 and 500 particles, respectively, much fewer than the other four classes, which each contains $2000 \sim 3000$ particles. Furthermore, the numbers of particles of some pair of classes may be anti-correlated. From iteration 24 to 60, the correlation coefficient between the numbers of particles of each pair of classes is: $\rho_{1,2} = -0.56, \rho_{1,3} = -0.45, \rho_{1,4} = -0.48, \rho_{1,5} = 0.48, \rho_{1,6} = -0.62, \rho_{2,3} = 0.72, \rho_{2,4} = -0.00, \rho_{2,5} = -0.74, \rho_{2,6} = 0.05, \rho_{3,4} = -0.14, \rho_{3,5} = -0.81, \rho_{3,6} = 0.31, \rho_{4,5} = -0.21, \rho_{4,6} = -0.01, \rho_{5,6} = -0.26$. As shown in the next section, the correlation coefficient is not a robust indicator to determine which classes may be combined.

There are several reasons for the existence of jumper particles: (1) projection angle step size is too large. The projection angle in real space is parameterized by three Euler angles (Scheres, 2012a), and the first two Euler angles are discretized using the HEALPix framework (Gorski et al., 2005) to achieve an approximately uniform sampling. As the projection angle step size

decreases, fewer particles will fall between the sampled projection angles. Therefore, more particles will find their most-probable class assignment and orientation with high confidence, i.e., having high $\max_{k,\phi}\ p(z_i = k, \phi|\mathbf{X}_i, \mathbf{\Theta})$. (2) Limited number of particles in some class results in low-quality reconstruction of the class, likely because of missing projection angles and/or noisy averaged images in some projection angles. (3) A portion of the particles contain local conformational/compositional differences compared to the averaged 3D reconstruction. Such differences may not be distinguishable at the current resolution, partly due to the small number of particles containing these features.

3.4. Using jumper analysis to find the classes that may be combined

Particles commuting between classes after the iteration of convergence indicate that these classes may have similar reconstructions, and in that case their particles can be combined to yield a better-quality reconstruction. We use the jumper analysis to find the classes that may be combined. To reduce the dependence of the analysis on the choice of iteration, we monitor the change of particle assignment along several iterations after the iteration of convergence, rather than just between two consecutive iterations.

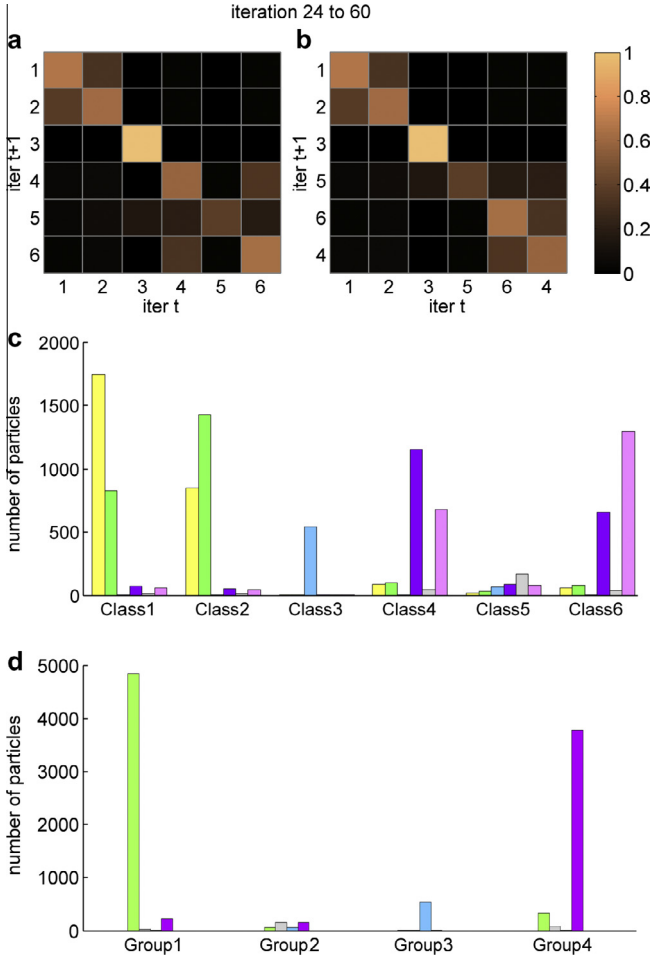

Fig.2. Transition matrix and bar diagram after the iteration of convergence. (a,b) Transition matrix before ($TM_{24\sim60}$, panel (a)) and after ($A_{24\sim60}$, panel (b)) agglomerative rearrangement, respectively. The order of classes in panel (b) is represented by $[q]$ in the text. Heatmap color scheme from dark to light corresponds to values 0–100% in the transition matrix. There are notable size of commuting jumper particles between Class 1 and Class 2, and between Class 6 and Class 4. (c,d) Bar diagram before (c) and after (d) combining classes into new groups. Class 1 and Class 2 are combined into Group 1. Class 5 is assigned as Group 2. Class 3 is assigned as Group 3. Class 4 and Class 6 are combined into Group 4.

We now introduce the *transition matrix* to show the fraction of jumper particles in each class along several iterations (Fig. 2a). The elements on the diagonal represent the fraction of the particles that remain in each class in two consecutive iterations, averaged from iteration 24 to 60. The elements off the diagonal represent the fraction of jumper particles between each pair of classes. From this transition matrix, we are able to recognize readily that Class 3 and Class 5 are distinct classes, whereas Class 1 and Class 2 share ~40% of particles, and Class 4 and Class 6 share ~35% of particles. However, the transition matrix is more difficult to analyze by eye as the number of classes increases. We therefore applied an agglomerative algorithm to reorganize the transition matrix (Amestoy et al., 1996, 2004), so that the high values off the diagonal in the transition matrix are close to the diagonal (Fig. 2b). We then identify the two classes that share the highest portion of particles and combine them, and do this iteratively (Algorithm 1). By choosing a cutoff value of 35%, we combined Class 1 and Class 2 into new Group 1, Class 5 as new Group 2, Class 3 as new Group 3, and combined Class 4 and Class 6 into new Group 4.

To examine the effectiveness of grouping classes, we further reduced the dimension of the data in the transition matrix by using the *bar diagram*. The bar diagram illustrates the average number of particles that stay in a class, or jump to another class, during the given iterations. The bar diagram in this example (Fig. 2c) clearly shows that Class 1 and Class 2 share a sizeable portion of the particles in these two classes, and so do Class 4 and Class 6, whereas Class 3 shares almost no particles with the other classes, and Class 5 have too few particles to yield a reliable reconstruction. After combining the classes, the new bar diagram (Fig. 2d) shows that the new groups do not share a sizeable portion of particles, indicating successful grouping. Furthermore, the new bar diagram indicates that the new Group 1 and Group 4 may have similar reconstructions, because they still share a detectable number of particles.

3.5. Examining the maps to confirm the jumper analysis results

To verify the conclusion of the jumper analysis, we examined the reconstructions of the classes at the iteration of convergence (Fig. 3a–f). The Class 3 map is a ribosome large subunit, distinct from the other classes. The Class 1 and Class 2 maps are both 70S ribosome complexes containing EF-G, different from the Class 4 and Class 6 maps, which are 70S ribosome complexes containing three tRNAs but no EF-G. Class 5 yields a low-quality reconstruction that is not comparable with the other classes. This classification result is in good agreement with the study by Scheres (Scheres, 2012b). However, another study by Lyumkis et al. using FREALIGN identified a class of 70S ribosome complex with strong density for A- and P-site tRNAs and weak density for E-site tRNA, indicating that there may be 70S ribosome lacking E-site tRNA in this dataset (Lyumkis et al., 2013). The fact that such class of 70S containing only A- and P-site tRNAs was not identified in our experiment is likely due to differences in the performance of the classification algorithms with the given choices of classification parameters.

Furthermore, we calculated the difference maps between Class 1 and Class 2, and between Class 4 and Class 6, showing that each pair of maps are identical at the current resolution (Fig. 3g,i). In contrast, the Class 4 map is different from Class 1 map (Fig. 3h). The Class 4 map lacks EF-G, but contains additional A- and P-site tRNAs, and its ribosomal L1 stalk and small subunit head are in different conformations compared to the Class 1 map. Therefore, we conclude that the jumper analysis results are consistent with examining the maps visually, and that it can facilitate the identification of classes that have similar reconstructions in the iterative classification of cryo-EM data.

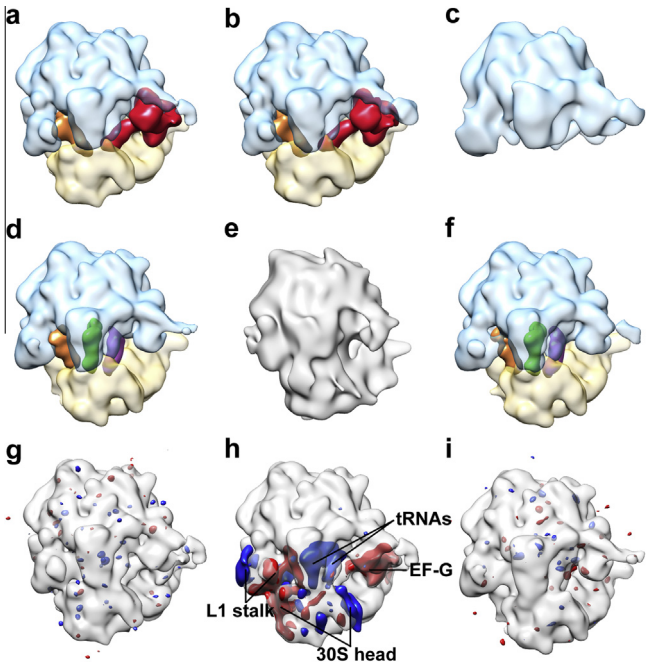

Fig.3. Cryo-EM map of each class and comparison between maps using difference map. (a–f) Cryo-EM maps of Class 1–6, respectively. Ribosome large and small subunit, transparent blue and transparent yellow, respectively; elongation factor G (EF-G), red; A-, P-, and E-site tRNAs, purple, green, and orange, respectively. The resolutions of Class 1-6 maps are: 18.3 Å, 19.3 Å, 30.5 Å, 19.3 Å, 30.5 Å, 19.3 Å, respectively. (g–i) The difference maps of (Class 1–Class 2), (Class 1–Class 4), and (Class 4–Class 6), respectively, shown at the threshold of $\pm 5\times$ standard deviation of each difference map. The maps being subtracted from are shown in transparent gray for viewing aid. Red mass represents positive difference; blue mass represents negative difference.

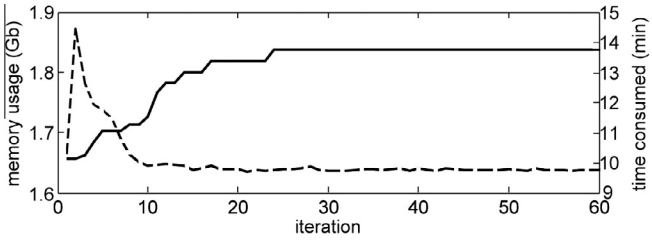

Fig.4. Memory usage and computation time to calculate expectation as a function of iteration number. The memory usage (solid line) reaches a plateau after iteration 24, the iteration of convergence. The computation time to calculate expectation (dashed line) peaks before iteration 5, and remains low after iteration 10.

3.6. Computational cost of the jumper analysis

The jumper analysis requires moderate amount of extra computational cost of RELION 3D classification. Fig. 4 shows the memory usage and computational time for the benchmark dataset containing 10,000 particles with 130-pixel window size, on a 104 CPU $\times$ 2GB/CPU cluster. The first 24 iterations consumed $\sim 40\%$ of the total computational time from iteration 1 to 60, indicating that the jumper analysis, which requires extra iterations of classification after the iteration of convergence, adds $\sim 1\times$ of computational cost. The reason for shorter computational time to calculate expectation in the later iterations of RELION 3D classification is mainly due to the "spiky" angular search: in the early iterations, the probability distribution among projection angles for each particle is flat, thus many projection angles need to be considered when calculating the particle statistics; in later iterations, the probability distribution among projection angles becomes sharper, approaching the delta function, therefore only a few projection angles need to be considered.

3.7. Toward fully automated classification

In this work, we have demonstrated the usefulness of the jumper analysis in an iterative classification of cryo-EM data for determining the iteration of convergence and the number of distinguishable classes. The jumper analysis method has also been used to facilitate the classification of datasets of other ribosomal complexes, i.e., the 70S-EttA complex (Chen et al., 2014) and the eukaryotic ribosome pre-termination complex (des Georges et al., 2014); however, the jumper analysis was not explicitly referenced in these two publications as this paper was still in preparation. Our method has great potential for further development into a fully automated classification process, which is an essential step toward a fully automated pipeline for cryo-EM and single-particle reconstruction – from data collection, particle picking, to 3D reconstruction and classification. Such a fully automated pipeline will greatly facilitate processing high-quality large datasets collected from direct detection devices, and will make cryo-EM more friendly to less-experienced users.

Acknowledgments

This work is supported by the Howard Hughes Medical Institute and the National Institute of Health Grant *R*01 *GM*55440 to J.F.

Appendix A. MATLAB Functions

The MATLAB functions for running the jumper analysis are accessible at (http://franklab.cpmc.columbia.edu/franklab/wp-content/uploads/2014/08/JumperAnalysis.zip).

References

Agirrezabala, X., Liao, H., Schreiner, E., Fu, J., Ortiz-Meoz, R., Schulten, K., Green, R., Frank, J., 2012. Structural characterization of mRNA–tRNA translocation intermediates. Proc. Natl. Acad. Sci. 109 (16), 6094–6099.
Amestoy, P.R., Davis, T.A., Duff, I.S., 1996. An approximate minimum degree ordering algorithm. SIAM J. Matrix Anal. Appl. 17 (4), 886–905.
Amestoy, P.R., Davis, T.A., Duff, I.S., 2004. Algorithm 837: AMD, an approximate minimum degree ordering algorithm. ACM Trans. Math. Softw. 30 (3), 381–388.
Bammes, B.E., Rochat, R.H., Jakana, J., Chen, D.-H., Chiu, W., 2012. Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency. J. Struct. Biol. 177 (3), 589–601.
Baxter, W.T., Grassucci, R.A., Gao, H., Frank, J., 2009. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. J. Struct. Biol. 166 (2), 126–132.
Casella, G., Berger, R.L., 2001. Statistical Inference. Duxbury Press.
Chen, B., Boel, G., Hashem, Y., Ning, W., Fei, J., Wang, C., Gonzalez Jr, R.L., Hunt, J.F., Frank, J., 2014. EttA regulates translation by binding the ribosomal E site and restricting ribosome-tRNA dynamics. Nat. Struct. Mol. Biol. 21, 152–159.
des Georges, A., Hashem, Y., Unbehaun, A., Grassucci, R.A., Taylor, D., Hellen, C.U., Pestova, T.V., Frank, J., 2014. Structure of the mammalian ribosomal pre-termination complex associated with eRF1.eRF3.GDPNP. Nucleic Acids Res. 42, 3409–3418.
Fischer, N., Konevega, A., Wintermeyer, W., Rodnina, M., Stark, H., 2010. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. Nature 466 (7304), 329–333.
Frank, J., 2006. Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford University Press, USA.
Frank, J., 2010. The ribosome comes alive. Isr. J. Chem. 50 (1), 95–98.
Gorski, K.M., Hivon, E., Banday, A., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M., 2005. HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. Astrophys. J. 622 (2), 759.
Lander, G.C., Stagg, S.M., Voss, N.R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., Lyumkis, D., Potter, C.S., Carragher, B., 2009. Appion: an integrated, database-driven pipeline to facilitate EM image processing. J. Struct. Biol. 166 (1), 95–102.
Langlois, R., Pallesen, J., Ash, J.T., Ho, D.N., Rubinstein, J.L., Frank, J., 2014. Automated particle picking for low-contrast macromolecules in cryo-electron microscopy. J. Struct. Biol. 186 (1), 1–7.

Lyumkis, D., Brilot, A.F., Theobald, D.L., Grigorieff, N., 2013. Likelihood-based classification of cryo-EM images using FREALIGN. J. Struct. Biol. 183 (3), 377–388.

Mastronarde, D. et al., 2005. Automated electron microscope tomography using robust prediction of specimen movements. J. Struct. Biol. 152 (1), 36–51.

Milazzo, A.-C., Cheng, A., Moeller, A., Lyumkis, D., Jacovetty, E., Polukas, J., Ellisman, M.H., Xuong, N.-H., Carragher, B., Potter, C.S., 2011. Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy. J. Struct. Biol. 176 (3), 404–408.

Scheres, S., 2012a. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180 (3), 519–530.

Scheres, S.H., 2012b. A Bayesian view on cryo-EM structure determination. J. Mol. Biol. 415 (2), 406–418.

Scheres, S.H., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P., Frank, J., Carazo, J.-M., 2007. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat. Methods 4 (1), 27–29.

Scheres, S.H., Valle, M., Grob, P., Nogales, E., Carazo, J.-M., 2009. Maximum likelihood refinement of electron microscopy data with normalization errors. J. Struct. Biol. 166 (2), 234–240.

Scheres, S.H., Valle, M., Nuñez, R., Sorzano, C.O., Marabini, R., Herman, G.T., Carazo, J.-M., 2005. Maximum-likelihood multi-reference refinement for electron microscopy images. J. Mol. Biol. 348 (1), 139–149.

Shaikh, T.R., Trujillo, R., LeBarron, J.S., Baxter, W.T., Frank, J., 2008. Particle verification for single-particle, reference-based reconstruction using multivariate data analysis and classification. J. Struct. Biol. 164 (1), 41.

Shen, B., Chen, B., Liao, H., Frank, J., 2014. Quantitative analysis in iterative classification schemes for cryo-EM application. In: Herman, G.T., Frank, J. (Eds.), Computational Methods for Three-Dimensional Microscopy Reconstruction. Applied and Numerical Harmonic Analysis. Springer, New York, pp. 67–95.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., Carragher, B., 2005. Automated molecular microscopy: the new Leginon system. J. Struct. Biol. 151 (1), 41–60.

Yin, Z., Zheng, Y., Doerschuk, P.C., Natarajan, P., Johnson, J.E., 2003. A statistical approach to computer processing of cryo-electron microscope images: virion classification and 3-D reconstruction. J. Struct. Biol. 144 (1–2), 24.