Chapter 4
Quantitative Analysis in Iterative Classification Schemes for Cryo-EM Application

Bingxin Shen, Bo Chen, Hstau Liao, and Joachim Frank

Abstract Over the past three decades, cryogenic electron microscopy (cryo-EM) and single-particle reconstruction (SPR) techniques have evolved into a powerful toolbox for determining biological macromolecular structures. In its original form, the SPR requires a homogeneous sample, i.e., all the projection images represent identical copies of the macromolecules (Frank, Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state, Oxford University Press, Oxford, 2006). Recent developments in computational classification methods have made it possible to determine multiple conformations/structures of the macromolecules from cryo-EM data obtained from a single biological sample (Agirrezabala et al., Proc Natl Acad Sci 109:6094–6099, 2012; Fischer et al., Nature 466:329–333, 2010; Scheres, J Struct Biol 180:519–530, 2012). However, the existing classification methods involve different amounts of arbitrary decisions, which may lead to ambiguities of the classification results. In this work, we propose a quantitative way of analyzing the results obtained with iterative classification of cryo-EM data. Based on the logs of iterative particle classification, this analysis can provide quantitative criteria for determining the iteration of convergence and the number of distinguishable conformations/structures in a heterogeneous cryo-EM data set. To show its applicability, we tailored this analysis to the classification results of the program RELION (Scheres, Methods Enzymol 482:295–320, 2010; Scheres, J Mol Biol 415:406–418, 2011) using both benchmark and experimental data sets of ribosomes.

B. Shen • H. Liao • J. Frank (✉)
Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University, New York, NY 10032, USA
e-mail: bs2733@columbia.edu; hl2485@columbia.edu; jf2192@columbia.edu

B. Chen • J. Frank
Department of Biological Sciences, Columbia University, New York, NY 10027, USA
e-mail: bc2357@columbia.edu

G.T. Herman and J. Frank (eds.), *Computational Methods for Three-Dimensional Microscopy Reconstruction: Applied and Numerical Harmonic Analysis*,
DOI 10.1007/978-1-4614-9521-5_4, © Springer Science+Business Media New York 2014

4.1 Introduction

In cryogenic electron microscopy (cryo-EM), micrographs of frozen-hydrated macromolecular complexes are collected using the transmission electron microscope. The biological macromolecules, embedded in vitreous ice and free from intermolecular interactions, are called single particles. The micrographs of the single particles are interpreted as two-dimensional (2D) projections of a three-dimensional (3D) object. The single-particle reconstruction (SPR) method recovers the 3D object from a large number of these cryo-EM particles showing the object in different orientations. The SPR method includes automated particle selection [10] and 2D alignment steps (see below) and an iterative process of 3D projection matching and 3D reconstruction.

In its original form, the cryo-EM and SPR technique requires a homogeneous biological sample. Because the electron dose has to be very low to keep the biological sample from being damaged by the beam, individual particles contain a high level of background noise. The signal-to-noise ratio can be increased by averaging multiple particles representing the same view of the macromolecule. Therefore, the SPR method requires sample homogeneity: that all particles represent structurally and conformationally identical copies of the macromolecule [5]. Sample homogeneity is usually achieved by introducing chemical interventions or mutations to the macromolecules [6]. Without these interventions, the macromolecules can thermodynamically assume different conformations and/or contain different components at the point of freezing. This so-called sample heterogeneity problem had been a limiting factor to the applicability of the cryo-EM and SPR technique.

Recent developments in computational classification methods have made it possible to resolve multiple conformations/structures of the macromolecules from cryo-EM data obtained in the same biological sample [1, 2, 4, 17]. Classification methods can be divided into two categories, supervised and unsupervised methods [5]. Supervised classification utilizes two or more 3D density maps as references and separates the particles based on their similarities to these references. Unsupervised classification groups the particles based on their mutual relationships without such guidance from references. Although a low-resolution 3D map may be needed for initial 2D alignment and 3D projection matching, unsupervised classification methods are largely immune to the reference bias problem that limits the application of supervised classification methods.

Unsupervised classification methods for cryo-EM data usually employ statistical approaches, such as *maximum likelihood* (ML) and *maximum a posteriori* (MAP) estimation. The ML estimation has been applied to cryo-EM data classification in the past few years [14, 19]. The ML method estimates the underlying model (i.e., the structure of the 3D object) by optimizing a likelihood function which indicates how likely the model is correct given the observed data (i.e., particles). Theoretically, the ML estimation is asymptotically unbiased and efficient, i.e., when the data size tends to infinity, the ML estimator becomes as good as, or better than, any other asymptotically unbiased estimator of the true model [19]. However, in practice, the

cryo-EM data sets are noisy, finite in size, and lack the projection angle information for the particles. The ML estimator therefore may be prone to over-fitting, i.e., erroneously treating noise as signal.

The MAP estimation, on the other hand, provides a Bayesian approach to avoid over-fitting. The MAP estimator considers not only the likelihood function but also the prior knowledge of the underlying model. The prior knowledge is expressed in the form of a prior probability distribution of the model parameters. The MAP estimation optimizes the posterior probability, which is proportional to the product of the likelihood function and the prior distribution. Thus, the MAP estimation can be considered as a regularized ML estimation, in light of all the available information—the observed data and the prior knowledge. Very recently, the MAP estimation has been successfully implemented for cryo-EM 3D reconstruction and classification by Scheres, in an open-source program named REgularized Likelihood OptimizatioN (RELION) [15, 16]. RELION utilizes the smoothness of the 3D reconstruction as prior knowledge. The smoothness stipulation derives from the fact that the scattering potential detected by electrons varies smoothly in space [15].

Although existing SPR and classification methods have demonstrated their applicability, they all involve different amounts of heuristics, i.e., arbitrary decisions made by human experts. The heuristics, if properly exercised, can make the methods effective and efficient. However, they can also lead to over-fitting and limit the use of these methods by non-experts. Some heuristics include tuning free parameters, such as the shape of a low-pass filter and choice of effective resolution to impose smoothness on the 3D reconstruction [5]. There has yet to be an objective way of examining the classification and 3D reconstruction results.

RELION, based on a statistical framework, sets a good example in reducing the amount of heuristics in SPR and classification. RELION adopts an iterative *expectation-maximization* (MAP-EM, to avoid confusion with cryo-EM) scheme and limits the heuristics to choosing a numerical factor for the presumed degree of statistical dependence of signal components in Fourier space [15]. Nonetheless, users still need to choose an initial low-resolution reference map, the number of classes to start with, and the number of iterations before at which convergence is expected.

In this work, we develop and demonstrate a quantitative analysis of iterative classification applied to cryo-EM data. Based on the statistics of all the particles, this analysis can provide quantitative criteria both for determining the iteration of convergence and the number of distinguishable conformations/structures in a heterogeneous cryo-EM data set. We tailored this analysis to the classification results of the RELION program and demonstrate its applicability by using both benchmark and experimental data sets of ribosomes.

Specifically, the first step in this proposed method of quantitative analysis utilizes the change in the likelihood function and other quantitative measures to identify the iteration of convergence, i.e., the iteration at which the 3D reconstructions become stable. This step can substantially reduce the human effort that usually goes into evaluating the 3D reconstructions in each class obtained in each iteration.

Next, this method monitors the change in class assignments of all the particles after the iteration of convergence. This second step can provide a quantitative indication that certain groups of output classes may contain particles representing the same conformation/structure of the macromolecule. After examining the 3D reconstructions, the particles representing the same conformation/structure of the macromolecule can then be combined for further 3D reconstruction and refinement.

The rest of this chapter is organized as follows. The problem formulation is stated in Sect. 4.2.1. Algorithmic and mathematical details are presented in Sect. 4.2.2. The convergence and jumper analysis of RELION are discussed in Sect. 4.2.3. We demonstrate the implementation of the proposed methods in several examples using experimental data sets in Sect. 4.3. We conclude with some discussion on future work and final thoughts in Sect. 4.4.

4.2 Methods

The 3D reconstruction of cryo-EM data is by itself a challenging problem. The difficulties include: (1) Biological macromolecules, composed of mainly proteins and/or nucleic acids, are similar in electron density to the surrounding water molecules. Therefore, the cryo-EM micrographs of the macromolecules usually have low amplitude contrast. (2) The contrast transfer function (CTF) of the transmission electron microscope, the equivalent of the optical transfer function in light microscopy, results in phase inversions and loss of information at certain spatial frequencies. The CTF needs to be corrected in cryo-EM particles to obtain high-resolution structures [5]. In recent years, many procedures have been implemented to reconstruct the underlying structure from the noisy cryo-EM particles [9, 11, 13, 20].

Almost all the existing 3D reconstruction methods [9, 11, 13, 20] are based on the weak-phase-object approximation (WPOA), which leads to a linear model of particle formation in Fourier space [22]. In this section, we make use of the same linear model and express the 3D reconstruction problem in the case where multiple different structures exist in the same cryo-EM data set.

4.2.1 Problem Model

Assume that we have K structures in the sample and we collect N cryo-EM particles (also called particles for short). Each particle $\boldsymbol{\chi}_i$ is a noisy projection of a 3D volume $\mathbf{v}_{k_i}$ observed at orientation ϕ_i, where i is the index of the particle, with $i = 1, 2, \ldots, N$, and $\mathbf{v}_{k_i}$ is from one of the K structures, with $k_i \in [1, 2, \ldots, K]$. Usually $\boldsymbol{\chi}_i$ is a $D \times D$ array and each element represents the corresponding pixel value of the particle. The volume $\mathbf{v}_{k_i}$ is a $D \times D \times D$ array, and each element contains the corresponding voxel value of the structure. Let $\mathbf{X}_i$ be the 2D Fourier transform

of $\boldsymbol{\chi}_i$, with the same size $D \times D$; let $\mathbf{V}_{k_i}$ be the 3D Fourier transform of the 3D array representing the molecule $\mathbf{v}_{k_i}$, with dimension $D \times D \times D$. According to the WPOA, in Fourier space the particle $\mathbf{X}_i$ is formulated as

$$X_{ij} = CTF_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi_i} V_{k_i l} + N_{ij}, \tag{4.1}$$

where X_{ij} is the jth component of $\mathbf{X}_i$, with $j = 1, 2, \ldots, J$ and $J = D^2$; $V_{k_i l}$ is the lth component of $\mathbf{V}_{k_i}$, with $l = 1, 2, \ldots, L$ and $L = D^3$. CTF_{ij} is the jth component of the CTF for this particle. The term $\sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi_i} V_{k_i l}$ for $j = 1, 2, \ldots, J$ forms a $D \times D$ array, which is a central slice at orientation ϕ_i of $\mathbf{V}_{k_i}$. Such a slice in Fourier space is equivalent to a projection in real space in the same orientation ϕ_i. At the end, N_{ij} is complex noise in Fourier space. Note that the variables X_{ij}, $\sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi_i} V_{k_i l}$, and N_{ij} in Eq. (4.1) are complex.

The 3D reconstruction and classification problem for cryo-EM data is to find a solution for the 3D electron density distributions with parameter sets $\boldsymbol{\Theta}$ based on the observed data $\mathbf{X}$. The parameter set $\boldsymbol{\Theta} = \{\mathbf{V}_k\}$ comprises the underlying structures in Fourier space with $k = 1, 2, \ldots, K$. The available data $\mathbf{X} = \{\mathbf{X}_i\}$ are the particles as represented in Fourier space. Besides the particles, one may also have prior knowledge of the structures $p(\boldsymbol{\Theta})$. An important feature of cryo-EM data is the smoothness of the density distribution of the macromolecule, which directly implies that the $p(\boldsymbol{\Theta})$ have limited power in the high-frequency part of Fourier space. The MAP estimation is optimal in estimation theory as it finds the best model in the light of all the available information, namely the observed data along with the prior knowledge of the unknowns. The general MAP estimation and its usual implementation by MAP-EM are described in Sect. 4.2.2. After that the convergence and jumper analysis as applied to RELION are studied in Sect. 4.2.3

4.2.2 General Solution

In the Bayesian framework, we are interested in obtaining the MAP estimate of the structure parameter sets $\boldsymbol{\Theta}$, given a set of observations $\mathbf{X}$. The MAP estimate maximizes the posterior distribution:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\mathbf{X}), \tag{4.2}$$

but the posterior distribution does not allow any known closed-form expression in such high-dimensional applications as encountered in cryo-EM. Using the formula for Bayes' law, we have

$$p(\boldsymbol{\Theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{X})}, \tag{4.3}$$

where $p(\mathbf{X}|\boldsymbol{\Theta})$ is the likelihood of observing the data set $\mathbf{X}$ given the parameter set $\boldsymbol{\Theta}$; $p(\boldsymbol{\Theta})$ is the prior distribution of the parameters; and $p(\mathbf{X})$ is the evidence that the data sets $\mathbf{X}$ are observed. Note that the evidence $p(\mathbf{X})$ is constant for a known data set and the posterior may then be expressed as proportional to the numerator in Eq. (4.3):

$$p(\boldsymbol{\Theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}), \tag{4.4}$$

Instead of maximizing the posterior, we can equally maximize the regularized likelihood:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}), \tag{4.5}$$

One may notice that the estimate of $\boldsymbol{\Theta}$ which maximizes the likelihood $p(\mathbf{X}|\boldsymbol{\Theta})$ is the solution of the ML algorithm discussed in [14, 17]. In the limit of infinite data size, the ML estimate approaches the MAP estimate, where both give the best estimate of the underlying parameter sets.

With the observed particles $\mathbf{X}$ only, it is still mathematically infeasible and computationally too demanding to solve the maximization problem in Eq. (4.5). This is because a complete data set includes not only the particles $\mathbf{X}$ but also the class identity, i.e., which one of the structures the particle represents, and the orientation of the particle, i.e., from which projection angle it was obtained. The ML/MAP estimation can be relatively simple with a complete data set. However, due to the way that sample preparation and data collection are done in cryo-EM, the information of class assignment and orientation is missing. We therefore employ the MAP-EM algorithm to solve this maximization problem, which provides a framework to alternately and iteratively estimate the missing data and the unknown parameters of interest. The MAP-EM algorithm is popular for performing typical high-dimensional ML/MAP estimation [12].

To implement the MAP-EM algorithm in our case, we introduce two sets of missing data (also called hidden variables), which are not observed directly: let $\mathbf{k} = (k_1, k_2, \ldots, k_N)$ denote the class assignment of the particles, where $k_i = k$ if $\mathbf{X}_i$ comes from the kth structure; and let $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)$ represent the projection angles at which $\mathbf{X}_i$ is observed. The class assignment k_i follows a discrete distribution, with probability c_{ik} for taking the value k , where $k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} c_{ik} = 1$. The orientation angle ϕ_i theoretically follows a continuous distribution with all possible values on a sphere. Yet practically, this continuous distribution is approximated by a discrete distribution based on a discrete angular sampling grid. Note that $\boldsymbol{k}$ and $\boldsymbol{\phi}$, along with the observed particles $\mathbf{X}$, form the complete data set.

We usually assume the probability of observing each particle is independent and therefore the likelihood in Eq. (4.5) can be written as

$$
\begin{aligned}
p(\mathbf{X}|\boldsymbol{\Theta}) &= \prod_{i=1}^{N} p(\mathbf{X}_i|\boldsymbol{\Theta}) \\
&= \prod_{i=1}^{N} \sum_{k=1}^{K} \int_{\phi} p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta})\, p(k,\phi|\boldsymbol{\Theta}) \mathrm{d}\phi.
\end{aligned}
\tag{4.6}
$$

For simplicity, one can treat $p(k,\phi|\boldsymbol{\Theta})$ as a uniform distribution, i.e., the k and ϕ are evenly distributed among all possible values for class and orientation assignments. Furthermore, $p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta})$ can be obtained as follows. We assume the real and imaginary parts of the complex-valued noise as independent and Gaussian distributed

$$
Re(N_{ij}) \sim \mathcal{N}(0,\sigma_{ij}^2),\ Im(N_{ij}) \sim \mathcal{N}(0,\sigma_{ij}^2),
\tag{4.7}
$$

and then N_{ij} has zero mean and variance $2\sigma_{ij}^2$. According to Eq. (4.1), X_{ij} is Gaussian distributed with mean of $CTF_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}$ and variance $2\sigma_{ij}^2$, and therefore

$$
\begin{aligned}
p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta}) &= \prod_{j=1}^{J} p(X_{ij}|k,\phi,\boldsymbol{\Theta}) \\
&= \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} \exp\left(\frac{\left| X_{ij} - CTF_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl} \right|^2}{-2\sigma_{ij}^2} \right).
\end{aligned}
\tag{4.8}
$$

For a given algebraic form of the likelihood, the different forms of the prior $p(\boldsymbol{\Theta})$ pose different levels of complexity for maximizing the posterior. For a given likelihood function, a prior $p(\boldsymbol{\Theta})$ is called a *conjugate prior* if the prior and the likelihood have the same algebraic form, i.e., Gaussian form in our case. Therefore, we assume that each element in the Fourier transform of the 3D density map has zero-mean Gaussian distribution:

$$
Re(V_{kl}) \sim \mathcal{N}(0,\tau_{kl}^2),\ Im(V_{kl}) \sim \mathcal{N}(0,\tau_{kl}^2),
\tag{4.9}
$$

and the prior can be expressed as

$$
\begin{aligned}
p(\boldsymbol{\Theta}) &= \prod_{k=1}^{K} p(\mathbf{V}_k) \\
&= \prod_{k=1}^{K} \prod_{l=1}^{L} p(V_{kl}) \\
&= \prod_{k=1}^{K} \prod_{l=1}^{L} \frac{1}{2\pi\tau_{kl}^2} \exp\left(\frac{|V_{kl}|^2}{-2\tau_{kl}^2} \right).
\end{aligned}
\tag{4.10}
$$

The MAP-EM algorithm is an iterative method, alternating between an expectation step (E-step) and a maximization step (M-step). The algorithm can be summarized as follows: let n denote the iteration number, $n = 1, 2, \ldots$, and let i represent the index of the particles, $i = 1, 2, \ldots, N$.

`Step 1.` When $n = 1$, we initialize the parameter set $\boldsymbol{\Theta} = \{V_{kl}, \sigma_{ij}, \tau_{kl}\}$ [recall Equations (4.7) and (4.9)], with $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L$, $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, J$. We usually initialize

$$V_{kl}^{(1)} = V_{0l},$$

where $\mathbf{V}_0 = \{V_{0l}\}$ is the Fourier transform of a low-resolution known density map, called the reference.

We initialize

$$\sigma_i^{2\,(1)} = \frac{1}{M} \sum_{\mathrm{i}' \in \mathrm{A_i}} |\mathbf{X}_{i'}|^2 - |\frac{1}{M} \sum_{\mathrm{i}' \in \mathrm{A_i}} \mathbf{X}_{i'}|^2,$$

where $A_i = \{i' \mid \mathbf{X}_{i'}$ *are from the micrograph where* $\mathbf{X}_i$ *is from*$\}$ and M is the number of particles in that micrograph; $|\mathbf{X}_{i'}|^2$ is the power spectrum of an individual particle, and $|\frac{1}{M} \sum_{\mathrm{i}' \in \mathrm{A_i}} \mathbf{X}_{i'}|^2$ is the power spectrum of the averaged, unaligned particles in that micrograph. The subtraction removes strong low-frequency power from the averaged power spectrum, which is true signal rather than noise. Note that $\sigma_i^{2\,(1)}$ is a $D \times D$ array and $\sigma_{ij}^{2\,(1)}$ is the jth element in it.

Lastly, we initialize $\tau_{kl}^{2\,(1)} = \frac{1}{2} \left| V_{kl}^{(1)} \right|^2$.

`Step 2.` At iteration n, in the E-step, the distributions of missing data k_i and ϕ_i are estimated based on the current estimate of the parameter set $\boldsymbol{\Theta}^{(n)}$:

$$(k_i, \phi_i) \sim p(k, \phi | \mathbf{X}_i, \boldsymbol{\Theta}^{(n)}), \tag{4.11}$$

where $p(k, \phi | \mathbf{X}_i, \boldsymbol{\Theta}^{(n)})$, which will be denoted as $\Gamma_{i,k,\phi}^{(n)}$ for the rest of the chapter, is the posterior probability of class and orientation assignment for the ith particle, given the observation $\mathbf{X}_i$ and the current estimate of parameters $\boldsymbol{\Theta}^{(n)}$, which can be written as

$$\begin{aligned} \Gamma_{i,k,\phi}^{(n)} &= p(k, \phi | \mathbf{X}_i, \boldsymbol{\Theta}^{(n)}) \\ &= \frac{p(\mathbf{X}_i | k, \phi, \boldsymbol{\Theta}^{(n)}) \, p(k, \phi | \boldsymbol{\Theta}^{(n)})}{p(\mathbf{X}_i | \boldsymbol{\Theta}^{(n)})} \\ &= \frac{p(\mathbf{X}_i | k, \phi, \boldsymbol{\Theta}^{(n)}) \, p(k, \phi | \boldsymbol{\Theta}^{(n)})}{\sum_{k'=1}^{K} \int_{\phi'} p(\mathbf{X}_i | k', \phi', \boldsymbol{\Theta}^{(n)}) \, p(k', \phi' | \boldsymbol{\Theta}^{(n)}) \mathrm{d}\phi'}, \end{aligned} \tag{4.12}$$

where $p(k,\phi|\boldsymbol{\Theta}^{(n)})$ is a uniform distribution and

$$p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta}^{(n)}) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^{2\,(n)}} \exp\left(\frac{\left| X_{ij} - CTF_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}^{(n)} \right|^2}{-2\sigma_{ij}^{2\,(n)}} \right). \quad (4.13)$$

The Q function is built as follows:

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \int_{\phi} p(k,\phi|\mathbf{X}_i,\boldsymbol{\Theta}^{(n)}) \log p(k,\phi,\mathbf{X}_i|\boldsymbol{\Theta}) \mathrm{d}\phi + \log p(\boldsymbol{\Theta}). \quad (4.14)$$

`Step 3.` At iteration n, in the M-step, $\boldsymbol{\Theta}$ are estimated by maximizing the Q function:

$$\begin{aligned}
(\boldsymbol{\Theta}^{(n+1)}) &= \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{k=1}^{K} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \log p(k,\phi,\mathbf{X}_i|\boldsymbol{\Theta}) \mathrm{d}\phi + \log p(\boldsymbol{\Theta}) \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{k=1}^{K} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \log[p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta}) p(k,\phi|\boldsymbol{\Theta})] \mathrm{d}\phi + \log p(\boldsymbol{\Theta}) \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{k=1}^{K} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \log p(\mathbf{X}_i|k,\phi,\boldsymbol{\Theta}) \mathrm{d}\phi + \log p(\boldsymbol{\Theta}).
\end{aligned} \quad (4.15)$$

Note that $p(k,\phi|\boldsymbol{\Theta})$ was dropped in Eq. (4.15) as it is a uniform distribution.

The solution of the maximization problem for $\boldsymbol{\Theta} = \{V_{kl},\ \sigma_{ij},\ \tau_{kl}\}$ in Eq. (4.15) is [15]

$$V_{kl}^{(n+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \sum_{j=1}^{J} \mathbf{P}_{lj}^{\phi\ T} \frac{CTF_{ij} X_{ij}}{\sigma_{ij}^{2\,(n)}} \mathrm{d}\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \sum_{j=1}^{J} \mathbf{P}_{lj}^{\phi\ T} \frac{CTF_{ij}^2}{\sigma_{ij}^{2\,(n)}} \mathrm{d}\phi + \frac{1}{\tau_{kl}^{2\,(n)}}}. \quad (4.16)$$

Then the updated noise variance $2\sigma_{ij}^2$ is derived as

$$2\sigma_{ij}^{2\,(n+1)} = \sum_{k=1}^{K} \int_{\phi} \Gamma_{i,k,\phi}^{(n)} \left| X_{ij} - CTF_{ij} \sum_{l=1}^{L} \mathbf{P}_{jl}^{\phi} V_{kl}^{(n+1)} \right|^2 \mathrm{d}\phi, \quad (4.17)$$

and the updated variance $2\tau_{kl}^2$ is obtained as

$$2\tau_{kl}^{2\,(n+1)} = \left| V_{kl}^{(n+1)} \right|^2. \quad (4.18)$$

The M-step generates estimates for $\boldsymbol{\Theta}$ by combining all the available information, i.e., the complete data set and the prior information.

`Step 4.` If more iterations are needed, set $n = n + 1$, and repeat `step 2` and `step 3`.

The MAP-EM algorithm makes it possible to iteratively refine the estimates of the distribution of hidden variables, and the parameter sets for the underlying unknown structures. Thus this algorithm is an attractive choice because it provides a feasible framework to ensure the convergence of the posterior function toward a stationary point under fairly general conditions [23].

The whole optimization process is done to estimate the Fourier transforms of the density maps, $\{\boldsymbol{V}_k^{(n)}\}$ with $k = 1, 2, \ldots, K$. Therefore, it is straightforward to monitor the changes of the $\{\boldsymbol{V}_k^{(n)}\}$ generated in the M-step to determine when to stop the optimization. The MAP-EM steps have the goal of maximizing the posterior. Practically, when the point is reached where the posterior no longer increases consistently, but only fluctuates due to the noise, we can safely stop the process at such iteration n^*, thus defining the iteration of convergence. Moreover, some outputs in RELION in the E-step, e.g., distribution of class assignment k_i and orientation ϕ_i, and the maximum probability $\max(\Gamma_{i,k,\phi}^{(n)})$ can be used as well to determine the convergence of the MAP-EM procedure quantitatively. Because when the distributions of k_i and ϕ_i stabilize for most particles, and the corresponding peak value $\max(\Gamma_{i,k,\phi})^{(n)}$ of the above probability distributions show little change, the $V_{kl}^{(n)}$ will stabilize as well, as suggested by Eq. (4.16). In the following section, we discuss how to utilize the statistical distribution of all these variables.

4.2.3 Convergence and Jumper Analysis

In practice, one can employ RELION to apply the MAP estimation to a noisy cryo-EM data set. RELION maximizes the posterior in Fourier space; refer to Eqs. (4.2) and (4.4). However, there has yet to be an objective way of examining the classification and 3D reconstruction results. We demonstrate a quantitative analysis of the statistics (i.e., the distribution of k_i and ϕ_i, and $\Gamma_{i,k,\phi}^{(n)}$) of each particle, which can help determine the iteration of convergence and provide clues for the number of distinguishable conformations obtained by the RELION classification.

If we process a data set using RELION with $\tilde{K}$ classes and N iterations, we will get $\tilde{K} \times N$ density maps as output. For instance, in the first example in Sect. 4.3.1, we have $\tilde{K} = 4$ and $N = 25$, in which case RELION provides 100 density maps for the users to check. We aim to determine the iteration of convergence n^* by quantitative analysis to lessen the manual examination by users. We assume that from the iteration of convergence on, all the estimated density maps have converged to a local maximum/stationary point, i.e.,

$$\boldsymbol{V}_k^{(n)} \approx \boldsymbol{V}_k^{(n^*)} \quad \textit{for } n \geq n^* \textit{ and } k = 1, 2, \ldots, \tilde{K}. \tag{4.19}$$

In other words, $\boldsymbol{V}_k^{(n)}$ for $n \geq n^*$ stay on average close to the local optimal solution. However, one should notice that the solution obtained may be suboptimal, i.e., it may represent a local maximum, depending on the suitability of the initial references. Furthermore, the fluctuations of $\boldsymbol{V}_k^{(n)}$ after the iteration of convergence $n^*$are mainly caused by noise.

At the E-step, we focus on available statistics in RELION outputs, the maximum probability $\max(\Gamma_{i,k,\phi}^{(n)})$, and the associated class estimates $\hat{k}_i^{(n)}$, orientation $\hat{\phi}_i^{(n)}$, in which class and at which orientation particle $\mathbf{X}_i$ contributes the most at iteration n. Note that $\hat{k}_i^{(n)}$ and $\hat{\phi}_i^{(n)}$ are actually the MAP estimates of class and orientation,

$$(\hat{k}_i^{(n)}, \hat{\phi}_i^{(n)}) = \arg\max_{k,\phi} \Gamma_{i,k,\phi}^{(n)},$$

and, for simplicity, we will just call them *class and orientation assignment* for the rest of the chapter.

In practice, if the selected number of classes $\tilde{K} > K$, there may be at least two classes in which the particles represent the macromolecules of the same conformation, that is, not distinguishable at the given resolution. In this case, when the whole MAP-EM process has converged with $\boldsymbol{V}_k^{(n)} \approx \boldsymbol{V}_k^{(n^*)}$ $for\ n \geq n^*$,

$$\exists s, t \in \{1, 2, \ldots, \tilde{K}\}, s \neq t$$

such that

$$\boldsymbol{V}_s^{(n)} \approx \boldsymbol{V}_s^{(n^*)} \approx \boldsymbol{V}_t^{(n^*)} \approx \boldsymbol{V}_t^{(n)} \quad for\ n \geq n^*.$$

If particle $\mathbf{X}_i$ represents a conformation with Fourier transform $\boldsymbol{V}_s \approx \boldsymbol{V}_t$, then

$$\max_{\phi} \Gamma_{i,s,\phi}^{(n)} \approx \max_{\phi} \Gamma_{i,t,\phi}^{(n)} > \max_{\phi} \Gamma_{i,r,\phi}^{(n)}, \quad r \in \{1, 2, \ldots, \tilde{K}\},\ r \neq s,\ r \neq t,$$

and

$$\max_{k,\phi} \Gamma_{i,k,\phi}^{(n)} = \max\{\max_{\phi} \Gamma_{i,s,\phi}^{(n)},\ \max_{\phi} \Gamma_{i,t,\phi}^{(n)}\}$$

By the definition, it is guaranteed

$$\Gamma_{i,\hat{k}_i^{(n)},\hat{\phi}_i^{(n)}}^{(n)} \geq \Gamma_{i,k,\phi}^{(n)}.$$

We will therefore have

$$\hat{k}_i^{(n)} \in \{s, t\}, \quad for\ n \geq n^*.$$

For an extreme case free of noise, where

$$\boldsymbol{V}_s^{(n)} = \boldsymbol{V}_s^{(n^*)} = \boldsymbol{V}_t^{(n^*)} = \boldsymbol{V}_t^{(n)} \quad for\ n \geq n^*,$$

the particle $\mathbf{X}_i$ may be contributed to class s or t with equal probability. Statistically, the $\hat{k}_i$ of half of such particles which were assigned to class s will be assigned to, or jump to, class t in the following iteration, and *vice versa*. A particle that changes its class assignment between two or among multiple classes, we call it a *jumper* particle. In contrast, a particle that remains in the same class in two consecutive iterations, we call it a *non-jumper*. Jumper particles indicate the similarities among output density maps of these classes. The particles from these similar classes may be merged to yield a density map with higher resolution. Note that jumping will not only occur when two or more classes have the same or similar density maps, but may also occur when the different density maps happen to have similar views from certain projection angles, e.g., the top views of 50S subunit and 70S ribosome shown in Fig. 4.9.

The orientation assignment also gives hints on the iteration of convergence. In RELION, the orientation assignment ϕ_i is specified as a projection angle on the HEALPix (Hierarchical Equal Area isoLatitude Pixelization) grid [8] φ_i, θ_i, and in-plane rotation angle ψ_i. The changes in orientation assignment are used as important convergence indicators in iterative SPR methods. For instance, the angular refinement process implemented in SPIDER [18] is deemed converged when the orientation assignments have reached the point where they fluctuate little as the iterations proceed. In RELION classification, it is not straightforward to use orientation assignment to determine convergence, because only the ϕ_i associated with the assigned class k_i are output, but the relative orientations of all the density maps are missing. If there are jumpers among classes, it is still difficult to quantify the changes in the particle orientation, because $\mathbf{V}_k^{(n)}$ can have different angular offsets for different classes. However, one can analyze the changes of the orientation for the particles that are consecutively assigned to the same class, assuming the angular offset of each class is near zero in two consecutive iterations.

The value of $\max(\Gamma_{i,k,\phi}) \in [0, 1]$ for each particle at each iteration is an indication for the confidence in the current estimated class and orientation assignment being the correct ones compared to the other possible values. For simplicity, $\max(\Gamma_{i,k,\phi})$ will be defined as the maximum probability, *MaxProb*. A value of MaxProb close to 1 indicates high confidence for the class and orientation assignments; a value close to 0 indicates high uncertainty in these assignments. Typically, these values increase as the iterations proceed, because the reconstructions will have higher resolution.

Another issue we wish to discuss here is how to choose $\tilde{K}$ without knowing the number of truly distinguishable classes (true K). It may be chosen based on user experience, for instance, the number of conformations that are expected based on prior biochemical knowledge and sample preparation conditions [2,4,24]. However, it is infeasible to determine the exact true K in most scientific experiments. Applying classification in a hierarchical manner may avoid this problem. As we

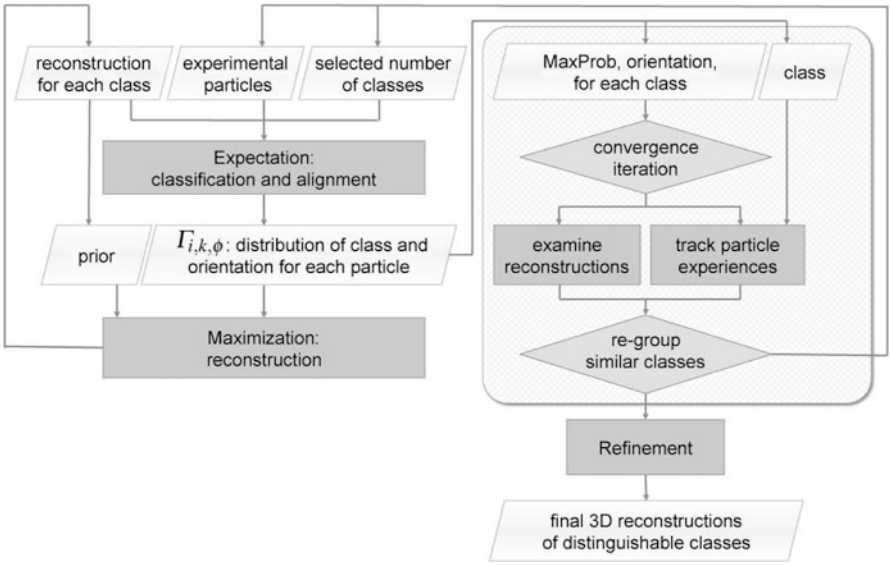

Fig. 4.1 Flowchart for the MAP estimation in RELION and the proposed quantitative analysis. RELION workflow is illustrated outside the *rounded box*. It employs the MAP-EM algorithm to alternately and iteratively implement the MAP estimation. The class assignment, orientation assignment, and maximum probability (MaxProb) for each particle at each iteration are used to quantify the general performance. The quantitative analysis, which is implemented in the rounded box using MATLAB, helps to determine the iteration of convergence and whether to confidently merge some of the classes

mentioned above, when the selected number of classes $\tilde{K} > K$, we can use jumper analysis to help determine the classes producing similar reconstructions and merge them for further processing or refinement. When the selected number of classes $\tilde{K} < K$, there will be at least one class which contains heterogeneous particles, i.e., particles representing macromolecules of different conformations or structures. The reconstruction of such a class may contain local densities that are blurred or averaged out or may have a lower resolution than a class with similar number of homogeneous particles. In such case, one should restart the classification with a greater number of $\tilde{K}$.

To sum up, the flow chart of quantitative analysis we discussed in this section is illustrated in Fig. 4.1. We determine the iteration of convergence n^* based on the distribution of maximum probability $\max(\Gamma_{i,k,\phi})^{(n)}$ and orientation assignment ϕ_i for each of the non-jumpers. The jumper analysis is carried out after the iteration of convergence to provide clues for determining the classes with similar reconstructions. Note that the users are still responsible for confirming the iteration of convergence and the similar classes before further processing. This whole procedure can be implemented in a hierarchical manner, and examples with details are presented in the following section.

4.3 Results and Discussions

The proposed quantitative analysis tools are tested in two examples with both benchmark and experimental data sets using RELION v1.0 in this section. The following examples illustrate the workflow of the proposed convergence and jumper analysis. The performance for the different number of classes selected $\tilde{K}$ is also studied and results are compared. Additionally, the computational cost, i.e., memory and time usages, is discussed at the end.

4.3.1 Benchmark Data Set

This benchmark data set for 3D classification algorithm comprises 10,000 *Escherichia coli* ribosome particles [3]. Supervised classification suggests that half of the particles are from "rotated" 70S ribosomes bearing elongation factor G (EF-G) and one tRNA; the other half are from "nonrotated" 70S bearing three tRNAs. Here "rotation" refers to the ratchet-like intersubunit rotation of the ribosome during translation [7].

We use a density map of a 70S ribosome filtered to 70 Å resolution as the initial reference. The number of iterations was set as 60, and the number of classes $\tilde{K}$ was selected as 4.

Histograms of MaxProb for class 1 are illustrated in Fig. 4.2. Most particles have lower MaxProb values at earlier iterations, for instance, iterations 4, 5, and 6 in the figure, which indicates low confidence in their class and orientation assignments regarding the current reconstructed density maps. A great number of particles have close to 1 MaxProb in later iterations, which show higher certainty of their class and orientation assignments.

One can use the mean, median, or mode values of the histograms in Fig. 4.2 to represent the feature of the distributions of the MaxProb. The trend of the MaxProb for each class along iterations is shown in Fig. 4.3, which can be interpreted as a measure of average confidence of particles regarding their class and orientation assignments. In this case, particles in classes 1, 3, and 4 have higher probability of having the correct class and orientation assignments, whereas those in class 2 have lower probability. There can be a number of reasons for the uncertainty of assignment to class 2, including, but not limited to, low-resolution reconstruction quality due to multiple conformations being entangled, small number of particles, or too many particles with low contrast caused by thick ice.

The mean values of MaxProb increases at earlier steps and then only fluctuates after a certain iteration, which indicates the iteration of convergence. We monitor the changes of these mean values along iterations in terms of $+/-$ percentage, as shown in Fig. 4.4. The mean values of MaxProb converge within 5 % fluctuation for at least five consecutive iterations after iteration 18, which is determined as the iteration of convergence for this run. The fluctuation range is usually related to the noise level in the data set.

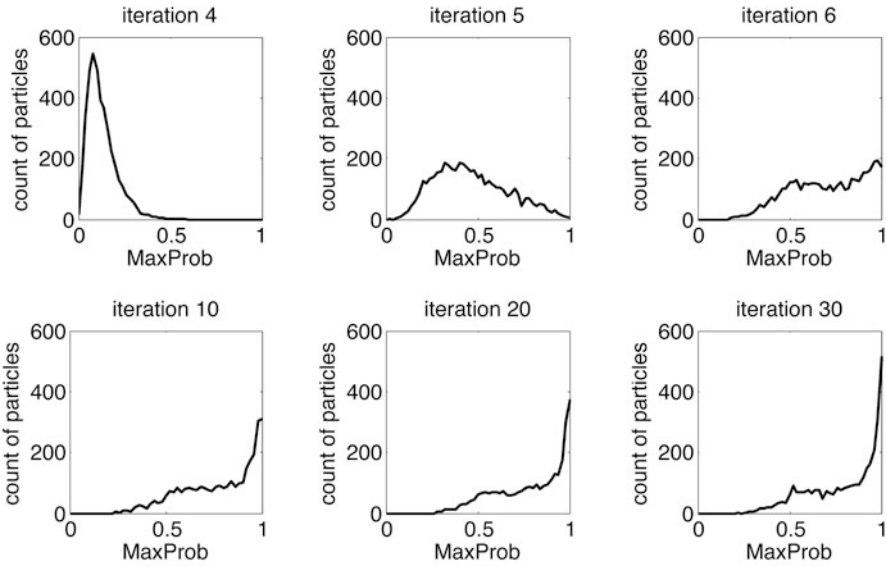

Fig. 4.2 Histogram of MaxProb for class 1 of the benchmark data set. The value of MaxProb increases along iterations, as the majority of particles gradually step toward better estimates of class and orientation assignments

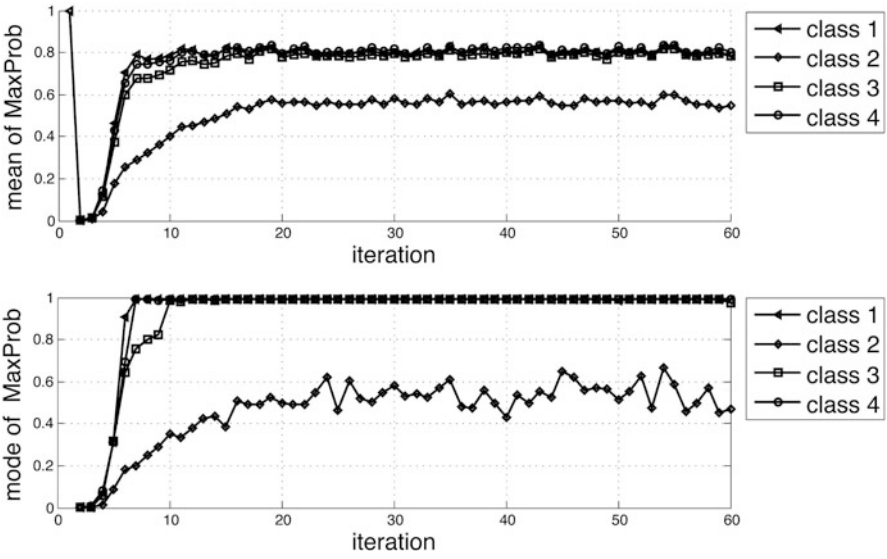

Fig. 4.3 Mean/mode values of MaxProb for each class along iterations. Particles in classes 1, 3, and 4 have higher confidence in their class assignments and orientation assignments, while particles from class 2 do not

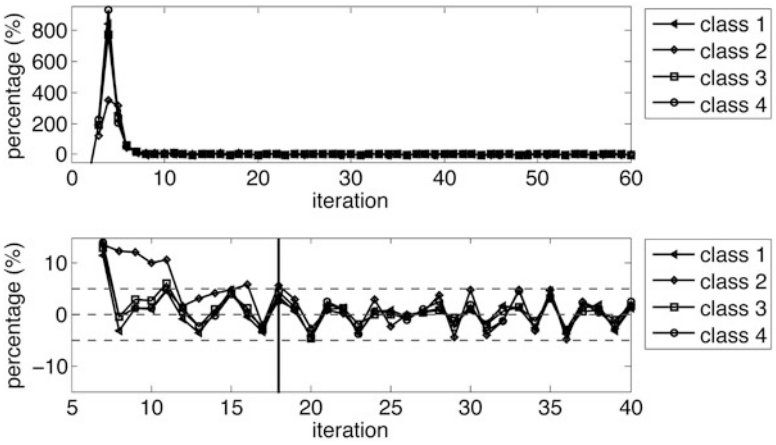

Fig. 4.4 Determine the iteration of convergence by monitoring the fluctuations of the mean values of MaxProb. The *upper box* shows all the fluctuations, which can be as great as 800 %, and the *lower one* focuses on the fluctuations within 5 %. This run converges after iteration 18 within 5 % fluctuation for at least five consecutive iterations

Recall Eq. (4.12) that particle orientation is assumed to possess a continuous distribution. However, we have to pixelize the sphere, i.e., to use a discrete sampling grid, to meet the real computational limit. HEALPix [8] is employed as the sampling grid, providing a mathematical structure which supports a suitable discretization of a sphere at sufficiently high resolution. In this run of the benchmark data set, we use an angular sampling interval of 7.5° which yields 768 discrete orientations. In order to explore as much of the sphere as possible, RELION randomizes the discrete sampling grid. Therefore, even after iteration of convergence, there are still fluctuations of the orientation assignments, not only due to the noise in the data set but also the changing sampling grids. Around 80 % of non-jumpers had orientation changes of less than 7.5°, and almost all of them had orientation changes of less than 15°, except class 2, as shown in Fig. 4.5. Note that particles in class 2 have a consistently worse performance than those in the other classes.

Particle number for each class versus iteration is shown in Fig. 4.6. Class 2 has fewer particles than the rest, less than 800. With this small data subset, we cannot expect a high-resolution density map from class 2. Therefore, the particles in this class have difficulties in alignment to the low-resolution reference, resulting in lower MaxProb values and more unstable orientation assignments, which explains the worse performance of class 2 in Figs. 4.3 and 4.5.

As we see from Fig. 4.6, the number of particles for each class does not stabilize after the iteration of convergence. More precisely, the counts of classes 1 and 3 vary frequently and they are actually anticorrelated with coefficient $\rho = -0.9$. If the particles interchange consistently between these two classes, we may expect that classes 1 and 3 represent the same or very similar structures. However, further analysis is required before we can draw such a conclusion.

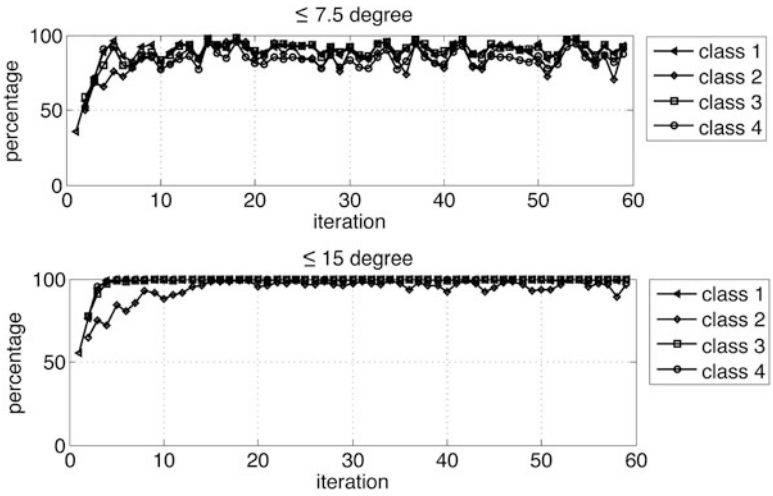

Fig. 4.5 Orientation changes for non-jumpers with sampling interval of 7.5°. *Above*: percentage of non-jumpers with orientation changes less than 7.5°. *Below*: percentage of non-jumpers with orientation changes less than 15°

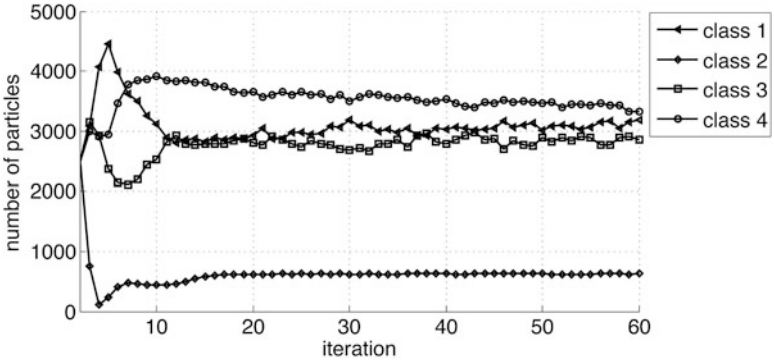

Fig. 4.6 Number of particles for each class versus iteration. The counts of classes 1 and 3 are anticorrelated with $\rho = -0.9$

As we discussed in the last section, when $\tilde{K} > K$, some classes might represent equivalent or very similar structures. After the iteration of convergence, particles belonging to such classes may jump frequently among them. Instead of monitoring the change of particle class assignments between two consecutive iterations, we can track particles jumping across multiple iterations, for example, from iteration 18 to 60. The following discussion will consider the interval between iteration 18 and 60. For any particle, the average number of visits to different classes is called the *average experience*. The average experience of particles after iteration 18 is shown in Fig. 4.7. As seen in the figure, classes 1 and 3 exchange their particles

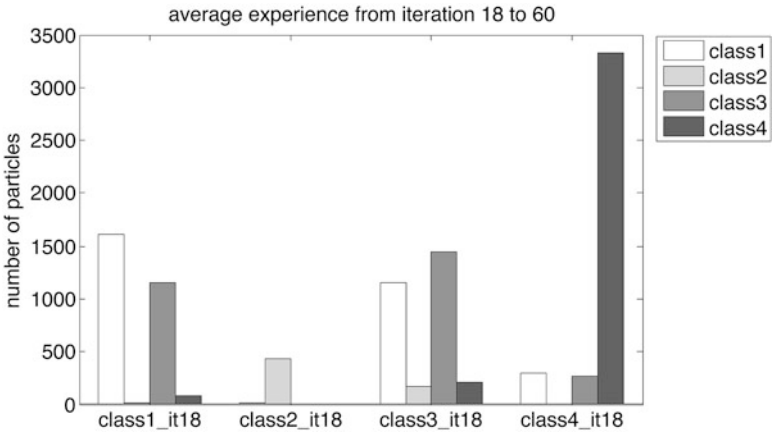

Fig. 4.7 Particle average experience from iteration 18 to iteration 60. The *first set of four bars* denote the particles which were assigned to class 1 at iteration 18. These *four bars* represent the average number of particles that were assigned to classes 1, 2, 3, and 4, respectively, after iteration 18. Classes 1 and 3 are suggested to generate the same or very similar density maps, as they exchange particles almost exclusively

exclusively, indicating that they represented the same underlying structure. Class 4 has some degree of similarity with classes 1 and 3 as they exchange a small portion of particles. Class 2 is distinct from the rest as it has isolated average experience.

The particle average experience can also be normalized and plotted as the average particle class transition, as illustrated in Fig. 4.8. Each square represents the normalized average particle transition after the iteration of convergence based on their average experience within the past 42 iterations. This map will help the user to quickly group similar classes when $\tilde{K}$ is relatively large (see the example with $\tilde{K} = 6$ in Fig. 4.17). In this example, classes 1 and 3 are naturally grouped together based on the associated particle average transitions.

The validity of grouping classes 1 and 3 together is confirmed by inspection of the 3D density maps at iteration 18, shown in Fig. 4.9. Classes 1 and 3 have visually identical density maps showing the 70S ribosome bound bearing an E-site tRNA and EF-G. Class 4 also represents a 70S ribosome but bearing three tRNAs and no EF-G. In sharp contrast, class 2 contains only the 50S subunit.

We test the consistency of our proposed analysis method on the benchmark data set by repeating the process several times. The iteration of convergence for each run is determined based on MaxProb distributions. Classes are regrouped based on the particle average experience after iteration of convergence. The groups and average count of particles for each class are summarized in Table 4.1.

The coincidence of particle assignments from different iterations for run #1 after convergence is shown in Table 4.2. Iterations 18, 40, and 60 are illustrated in the table. Elements on the diagonal of the table are the numbers of particles in the corresponding group at the specific iteration. For instance, group 1 has 5,725, 5,827,

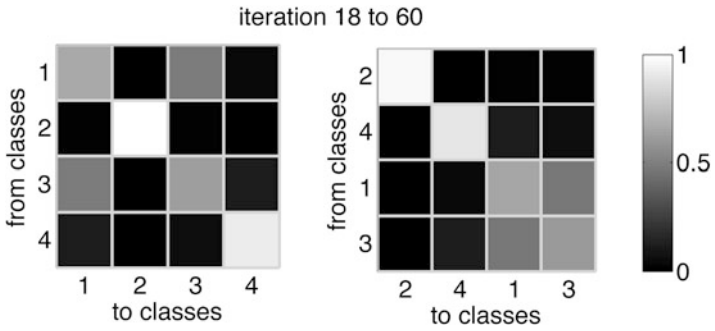

Fig. 4.8 Particle average transition map based on average experience going from iteration 18 to iteration 60. Each *square* represents a transition rate, between 0 (no transition) and 1 (largest number of transition). For example, the *first row* in the *left plot* denotes the particles that were assigned to class 1 at iteration 18. The *brightness of the four squares* represents the average proportions of particles that are assigned to classes 1, 2, 3, and 4, respectively, after iteration 18. The classes are reordered in the *right plot* according to how distinct they are from the rest. Classes 1 and 3 have similar brightness levels of non-jumpers and jumpers between them and therefore are grouped together

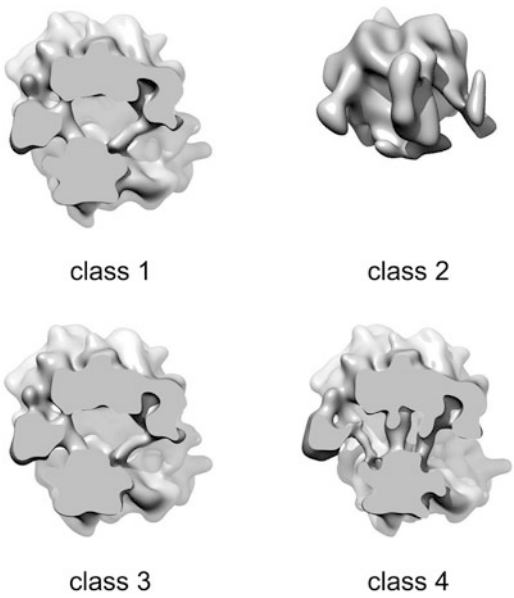

Fig. 4.9 Density maps at iteration 18. Classes 1 and 3 represent the 70S ribosome bound with E-site tRNA and EF-G. Class 4 also represents for a 70S ribosome but with three tRNAs and no EF-G. Class 2 contains only the 50S subunit

and 6,057 particles at iteration 18, 40, and 60, respectively. Moreover, the upper left 3×3 matrix tells us the coincidence of particles which are assigned as group 1 at iterations 18, 40, and 60. Among these particles, there were 5,483 particles that are in group 1 for both iterations 18 and 40; 5,260 particles for both iterations 18 and 60; and 5,245 particles for both iterations 40 and 60. Almost the same subset of particles

Table 4.1 Repeated runs for the benchmark data set

Run	Iteration of convergence	group1	group2	group3
#1	18	Classes 1, 3	Class 4	Class 2
	Average #	5,858	3,517	625
#2	18	Classes 1, 3	Class 4	Class 2
	Average #	5,864	3,512	624
#3	18	Classes 1, 3	Class 4	Class 2
	Average #	5,865	3,511	624
Notes		70S with EF-G and 1 tRNA	70S with 3 tRNAs	50S

Table 4.2 The coincidence of particle class assignments at different iterations from run #1

	iter18 group1	iter40 group1	iter60 group1	iter18 group2	iter40 group2	iter60 group2	iter18 group3	iter40 group3	iter60 group3
iter18 group1	5,725	5,483	5,260	0	210	428	0	32	37
iter40 group1	5,483	5,827	5,245	327	0	555	17	0	27
iter60 group1	5,260	5,245	6,057	768	779	0	29	33	0
iter18 group2	0	327	768	3,657	3,328	2,884	0	2	5
iter40 group2	210	0	779	3,328	3,539	2,755	1	0	5
iter60 group2	428	555	0	2,884	2,755	3,316	4	6	0
iter18 group3	0	17	29	0	1	4	618	600	585
iter40 group3	32	0	33	2	0	6	600	634	595
iter60 group3	37	27	0	5	5	0	585	595	627

is assigned in the same group and is used to reconstruct the density maps. In another example, the first column labeled as `iter18 group1` show that there were 5,725 particles in group 1 at iteration 18. At iteration 40, 5,483 of them (95.8 %) stayed in group 1, while 210 particles (3.7 %) went to group 2, and 32 particles (0.5 %) went to group 3. As we see from the density maps, groups 1 and 2 were both 70S ribosomes but bound with a different combination of factor and tRNAs. Since they had small local compositional differences, they still shared a small portion of particles even after the iteration of convergence. However, as seen from the table, the majority of particles had stable group assignments after the iteration of convergence.

The coincidence of particle assignments from different runs is shown in Table 4.3. For instance, the first row labeled as `run#1 group2` shows that there were 6,057 particles in group 1 in run #1. 5,996 of them (98.99 %) were found in group 1 in run #2, while 5,997 of them (99.01 %) were found in group 1 in run #3. Different runs gave very similar results once the particles are grouped properly according to their average experiences. As seen from Table 4.3, the majority of particles had consistent group assignments for different runs.

Table 4.3 Coincidence of particle class assignments from different runs at iteration 60

	run#1 group1	run#2 group1	run#3 group1	run#1 group2	run#2 group2	run#3 group2	run#1 group3	run#2 group3	run#3 group3
run#1 group1	6,057	5,996	5,997	0	46	53	0	15	7
run#2 group1	5,996	6,057	5,988	44	0	60	17	0	9
run#3 group1	5,997	5,988	6,086	73	85	0	16	13	0
run#1 group2	0	44	73	3,316	3,271	3,242	0	1	1
run#2 group2	46	0	85	3,271	3,322	3,236	5	0	1
run#3 group2	53	60	0	3,242	3,236	3,296	1	0	0
run#1 group3	0	17	16	0	5	1	627	605	610
run#2 group3	15	0	13	1	0	0	605	621	608
run#3 group3	7	9	0	1	1	0	610	608	618

4.3.2 Yjjk Data Set

The data set presented in this section is a subset of cryo-EM data of the 70S ribosome bound with a novel translation factor, called Yjjk, and two tRNAs. The data were collected in the low-dose mode on the FEI Tecnai F20 at 200kV extraction voltage using the Leginon program [21]. Micrographs were recorded on a Gatan UltraScanTM 4000 CCD camera binned by 2× with effective magnification on the CCD of $110,637\times$ and pixel size of 2.71 Å on the object scale. Only a subset of 21,182 particles of the total 108,691 particles are used here to explain our proposed convergence and jumper analysis. This subset contains the 50S subunit and the 70S ribosome bound with the factor $+/-$ tRNA.

This subset of 21,182 particles was assigned to one class with a 70S density map in the original RELION classification results, but with a relative low mean value of MaxProb 0.7 compared to 0.85 of the other classes, which indicates high uncertainties in class and orientation assignments. As discussed earlier, there can be a number of reasons for a low value of MaxProb, including, but not limited to, a low-resolution reconstruction due to multiple coexisting conformations, small number of particles, or too many particles with low contrast caused by thick ice. In this section we are showing how to further process such a subset.

We run RELION again only on this subset, using a 70S ribosome map filtered to 70 Å resolution as the initial reference. The number of iterations was set as 50, and the number of classes $\tilde{K}$ was selected as 4.

As discussed in the previous section, the mean value of MaxProb across iterations can be used to determine the iteration of convergence. This is shown in Fig. 4.10. Particles in classes 1, 3, and 4 have higher confidence in their class and orientation assignments, whereas those in class 2 had lower confidence.

The mean values of MaxProb increased in earlier iterations but then only fluctuated within 5 % after iteration 27, which was determined as the iteration of convergence for this run (Fig. 4.11).

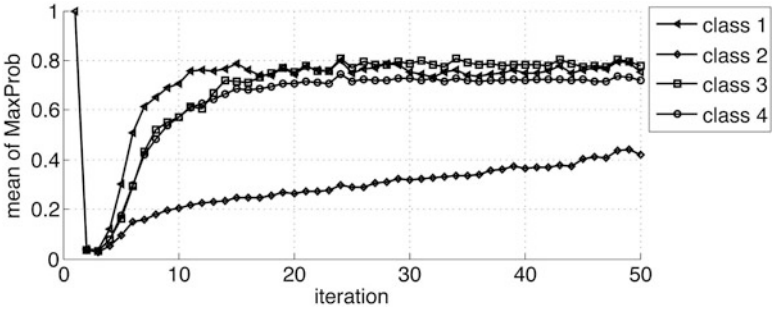

Fig. 4.10 Mean value of MaxProb for each class along iterations. Particles in classes 1, 3, and 4 have high confidence in their class and orientation assignment, while particles from class 2 have lower values

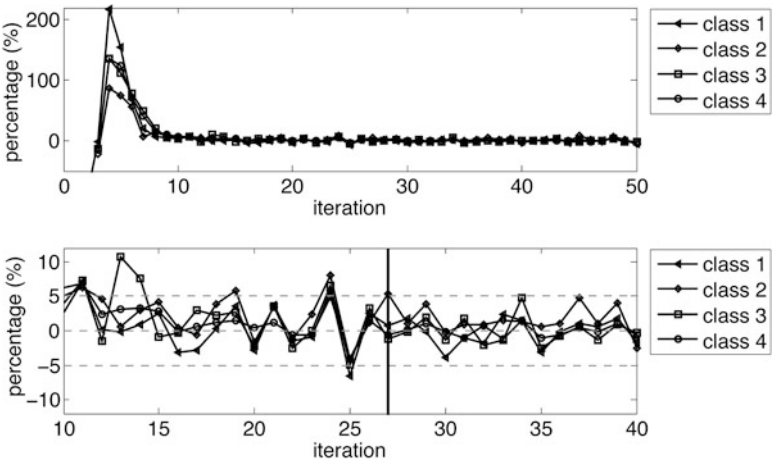

Fig. 4.11 Determination of the iteration of convergence by monitoring the changes of mean values of MaxProb. The *upper box* shows all the fluctuations, which can be as great as 200 %, and the *lower one* focuses on the fluctuations within 5 %. This run converged after iteration 27 within 5 % fluctuation for five consecutive iterations

The number of particles for each class versus iteration count is shown in Fig. 4.12. For class 4, this number stabilizes after iteration 10, suggesting that this class has distinct features compared to the rest. As shown in the figure, the number of particles varies substantially for classes 1 and 3 even after the iteration of convergence. More specifically, they are almost perfectly anticorrelated, with coefficient $\rho = -0.96$. Particles seem to jump exclusively between these two classes; indicating classes 1 and 3 may represent very similar structures. These hints require further confirmation by jumper analysis.

The average experiences of particles from iteration 27 onwards are shown in Figs. 4.13 and 4.14. Class 4 is distinct from the rest as its average experience is isolated. After the iteration 18, particles in classes 1 and 3 jumped frequently

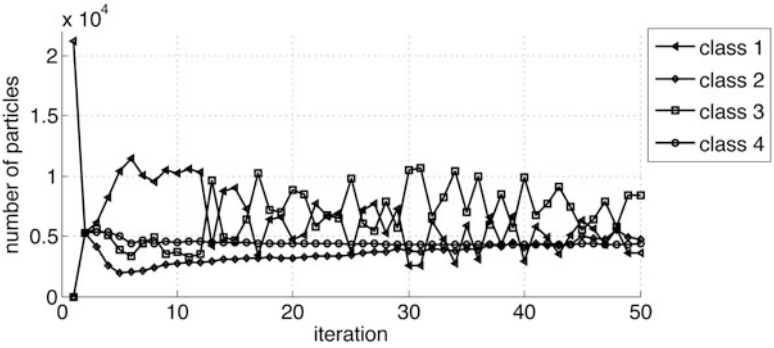

Fig. 4.12 Particle number for each class versus iteration. The counts of classes 1 and 3 are anticorrelated

Fig. 4.13 Particle average experience from iteration 27 to iteration 50. The *first set of four bars* denote the particles which were assigned to class 1 at iteration 27. These *four bars* represent the average number of particles that were assigned to classes 1, 2, 3, and 4 after iteration 27, respectively. Classes 1 and 3 were suggested to generate very similar density maps as they exchange a big portion of particle after iteration of convergence

between these two classes, indicating they might represent the same structure or very similar structures. As class 2 exchanges a small portion of particles with classes 1 and 3, it might have some degree of similarity with those two classes.

The above suggestions are confirmed upon visual inspection of the 3D density maps at iteration 29, an arbitrary iteration selected after convergence (Fig. 4.15). Class 4 contains only the 50S subunit, which is distinct from the other three classes. Classes 1 and 3 produce virtually identical density maps, showing the 70S ribosome bound with two tRNAs and the factor YjjK. The density map of class 2 is a 70S ribosome with less structural details and contains only Yjjk, with very weak tRNA densities. A small portion of particles assigned to class 2 visits the classes 1 and 3, as all three classes represent a 70S ribosome except for small local compositional differences.

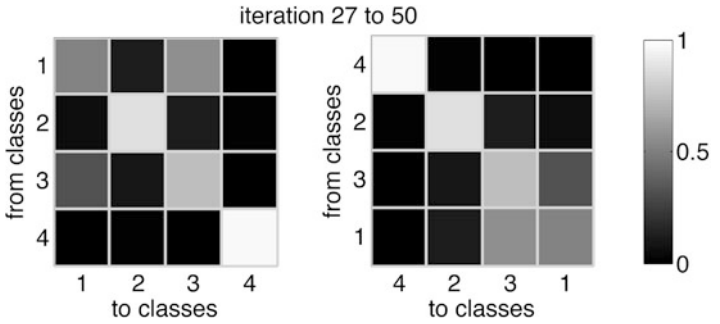

Fig. 4.14 Particle average transition map based on average experience going from iteration 27 to iteration 50. The transition rate is between 0 (no transition) and 1 (largest number of transitions). For example, the *first row* on the *left plot* denotes the particles which were assigned to class 1 at iteration 27. The *four squares* in the *first row* represent the average portion of particles that are assigned to classes 1, 2, 3, and 4 after iteration 27, respectively, in all iterations from 27 to 50. Classes are reordered on the *right* according to transition rates. Classes 1 and 3 have similar transition rates for both non-jumpers and jumpers between them and therefore are grouped together

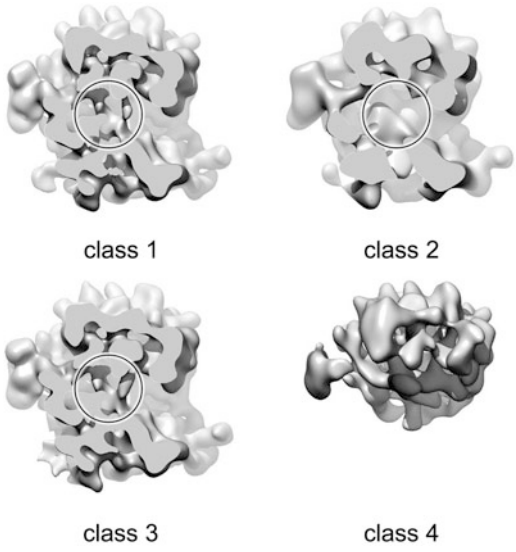

Fig. 4.15 3D density maps at iteration 29. Classes l and 3 represent a 70S ribosome with the translation factor Yjjk and two tRNAs bound, and they have virtually identical structure but with an angular offset of 10.1°. Class 2 represented a 70S ribosome with only factor YjjK bound. Class 4 generates the 50S subunit

We tested our proposed analysis method on this Yjjk data subset with different selected number of classes, $\tilde{K} = 4$ and 6. The general performance is summarized in Table 4.4. The iteration of convergence for each run was determined based on MaxProb distributions. Classes were grouped based on the associated particle average experience after convergence. The average numbers of particles for each group after convergence are also included in the table.

As shown, the iteration of convergence for runs #1, #2, and #3 are 27, 30, and 35, respectively. Run #2 is essentially a repeat of run #1, while $\tilde{K}$ is increased to

Table 4.4 Repeated runs for Yjjk data set

Run	$\tilde{K}$	Iteration of convergence	group1	group2	group3
#1	4	27	Classes 1, 3	Class 2	Class 4
		Average #	12,595	4,357	4,340
#2	4	30	Classes 1, 3	Class 2	Class 4
		Average #	13,712	3,029	4,441
#3	6	35	Classes 1, 2, 6	Class 3, 4	Class 5
		Average #	14,385	2,623	4,174
Notes			70S with Yjjk and 2 tRNAs	70S no tRNA	50S

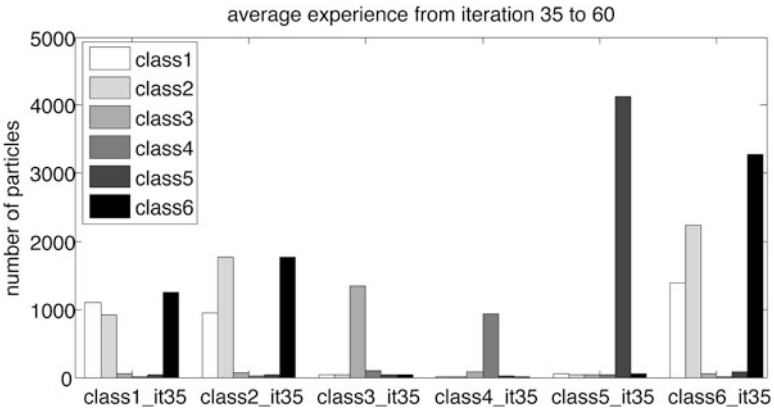

Fig. 4.16 Particle average experience from iteration 35 to iteration 60. The *first set of six bars* denote the particles who were assigned as class 1 at iteration 35. These *six bars* represent the average number of particles who were assigned as classes 1, 2, 3, 4, 5, and 6, respectively, after iteration 35. Class 5 is fairly distinct from the rest. Classes 1, 2, and 6 interchanged a big portion of particles

6 for run #3. Particle average experience from 35 onwards is summarized as bar plots in Fig. 4.16 and the associated particle average transition map is shown in Fig. 4.17. Classes 1, 2, and 6 share a large portion of jumpers, shown in Fig. 4.17. The density maps of classes 3 and 4 both represent a 70S with no tRNAs. But the resolution of each map is low, around 30 Å, possibly because the number of particles in each class is small, only about 1,000. These particles in each class were distributed among 768 orientations on the HEALPix sampling grid, so there are on average only 1–2 particle per orientation. This may explain why there are only a few jumpers between classes 3 and 4, although they are likely to represent the same 70S ribosome complex. We therefore grouped classes 3 and 4 together to compare with the other runs in Tables 4.4 and 4.5. The coincidences of particle class assignments from different runs are shown in Table 4.5. All particle class assignments are from iteration 50, picked arbitrarily after convergence. For instance, the sixth column labeled with `run#3 group2` shows that there were 2,560 particles in group 2 in

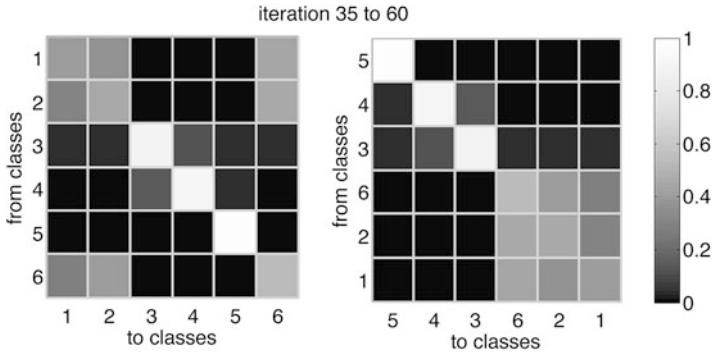

Fig. 4.17 Particle average transition map based on average experience going from iteration 35 to iteration 60. The transition rate is between 0 (no transition) and 1 (largest number of transitions). For example, the *first row* on the *left plot* denotes the particles which were assigned to class 1 at iteration 27. The *six squares* in the *first row* represent the average portion of particles that are assigned to classes 1, 2, 3, 4, 5, and 6 after iteration 35, respectively. Classes are reordered on the *right* according to transition rates. Classes 1, 2, and 6 were grouped by their average experience

Table 4.5 The coincidence of particle class assignments from different runs at iteration 50

	run#1 group1	run#2 group1	run#3 group1	run#1 group2	run#2 group2	run#3 group2	run#1 group3	run#2 group3	run#3 group3
run#1 group1	13,154	11,654	12,896	0	270	185	0	1,230	73
run#2 group1	11,654	13,123	12,569	1,240	0	397	229	0	157
run#3 group1	12,896	12,569	14,460	1,352	527	0	212	1,364	0
run#1 group2	0	1,240	1,352	3,675	2,200	2,134	0	235	189
run#2 group2	270	0	527	2,200	2,857	2,009	387	0	321
run#3 group2	185	397	0	2,134	2,009	2,560	241	154	0
run#1 group3	0	229	212	0	387	241	4,353	3,737	3,900
run#2 group3	1,230	0	1,364	235	0	154	3,737	5,202	3,684
run#3 group3	73	157	0	189	321	0	3,900	3,684	4,162

run #3. And 2,134 of them (83.36 %) were found in group 2 in run #1, while 2,009 of them (78.48 %) were found in group 2 in run #2. Only 185 and 397 particles (7.23 % and 15.51 %) were found in group 1 in runs #1 and #2, respectively. Different runs gave very similar results once we group the particle properly according to their average experiences. As we can see from the density maps in Fig. 4.15 from run #1, groups 1 and 2 both represent a 70S ribosome bound with or without Yjjk and tRNAs. Since they only had small local compositional differences, they still have a small portion of particles visiting each other even after the iteration of convergence. However, as we can see from the table, the majority of particles have stable group assignments after iteration 35.

4.3.3 Computational Cost

Finally, the computational cost of RELION 3D classification for the benchmark data set presented in Sect. 4.3.1 and the Yjjk data set presented in Sect. 4.3.2 needs to be addressed. The memory usage and time consumed during E-step, which is the major computational part, is illustrated in Fig. 4.18. The black plots correspond the benchmark data set of 10,000 particles. The gray plots are for the Yjjk subset of 21,182 particles. There was a total of 152 CPU $\times$ 2 Gb/CPU memory available for all runs. Both data sets used 76 parallel processes, with 4Gb memory available for each parallel job. The actual memory usage increased in the course of each run and stabilized around the iteration of convergence. We also observe that the first few iterations are the most computationally expensive in terms of the time consumed. In the case shown in the figure, the first 5 iterations took 75 % of computational time of the first 30 iteration. Both sets of plots follow a similar pattern; however, as the number of particles increases, the computational cost increases accordingly.

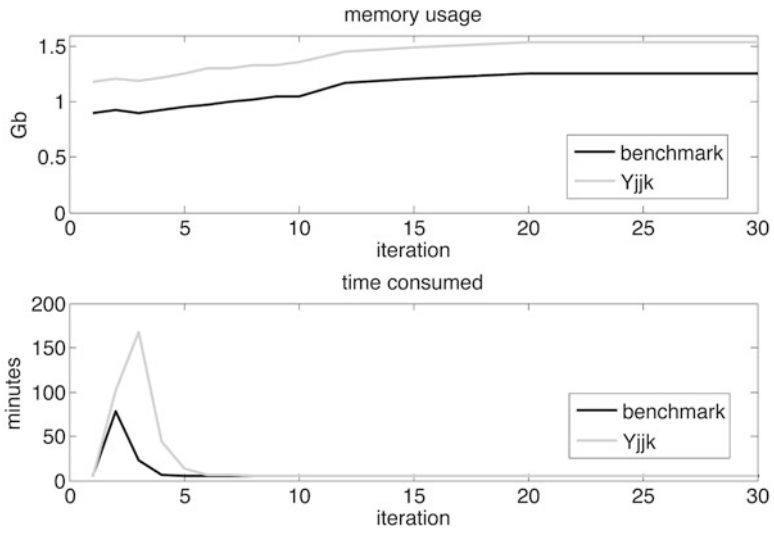

Fig. 4.18 *Black plot*: the benchmark data set of 10,000 particles presented in Sect. 4.3.1, using window size 130^2 pixel, $\tilde{K} = 4$, angular sampling interval of 7.5°, translation search step of 1 pixel, and translation search range of 6 pixel. *Gray plot*: the Yjjk subset of 21,182 particles presented in Sect. 4.3.2, using window size 134^2 pixel, $\tilde{K} = 4$, angular sampling interval of 7.5°, translation search step of 1 pixel, and translation search range of 5 pixel. Both data sets used 76 parallel processes with total 152 CPU

4.4 Conclusion

We analyzed the statistics of each particle as the primary criteria to determine the iteration of convergence, i.e., from which iteration onwards the 3D reconstructions become trustworthy and stable. The proposed quantitative analysis can also reveal the groupings of classes with very similar underlying structures, as validated by two examples using the benchmark and Yjjk data sets with different selected number of classes $\tilde{K}$. The convergence analysis was validated by comparison of particle class assignments, orientation assignments, and density maps from different iterations after convergence. The coincidences of particle class assignments were also studied across different groups at multiple iterations. In addition, the computational cost, i.e., memory and time usages, is discussed at the end. The proposed quantitative method lowers the amounts of arbitrary decisions by users, which may lead to ambiguities of the classification results.

Acknowledgements The authors are grateful to Sjors Scheres, Ming Sun, and Amy Jobe for valuable comments. The authors also would like to thank Nam Ho and Melissa Thomas for their help on figure illustrations and Bob Grassucci for aid with data collection. This work is supported by the Howard Hughes Medical Institute and the National Institute of Health Grant $R01\ GM55440$.

References

1. Agirrezabala X, Lei J, Brunelle JL, Ortiz-Meoz RF, Green R, Frank J (2008) Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. Mol Cell 32:190–197
2. Agirrezabala X, Liao HY, Schreiner E, Fu J, Ortiz-Meoz RF, Schulten K, Frank J (2012) Structural characterization of mRNA-tRNA translocation intermediates. Proc Natl Acad Sci 109:6094–6099
3. Baxter WT, Grassucci RA, 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 166:126–132
4. Fischer N, Konevega AL, Wintermeyer W, Rodnina MV, Stark H (2010) Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. Nature 466:329–333
5. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press, Oxford
6. Frank J (2010) The ribosome comes alive. Isr J Chem 50:95–98
7. Frank J, Agrawal RK (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature 406:318–322
8. Gorski KM, Hivon E, Banday AJ, Wandelt BD, Hansen FK, Reinecke M, Bartelmann M (2008) HEALPix: a framework for high-resolution discretization fast analysis of data distributed on the sphere. Astrophys J 622:759–771
9. Grigorieff N (1998) Three-dimensional structure of bovine NADH: ubiquinone oxidoreductase (complex I) at 22 $\mathring{A}$ in ice. J Mol Biol 277:1033–1046
10. Langlois R, Pallesen J, Frank J (2011) Reference-free particle selection enhanced with semi-supervised machine learning for cryo-electron microscopy. J Struct Biol 175:353–361
11. Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. J Struct Biol 128:82–97

12. McLachlan GJ, Krishnan T (1997) The EM algorithm extensions. Wiley Series in Probability Statistics, Hoboken
13. Penczek PA (2010) Image restoration in cryo-electron microscopy. Methods Enzymol 482: 35–72
14. Scheres SHW (2010) Classification of structural heterogeneity by maximum-likelihood methods. Methods Enzymol 482:295–320
15. Scheres SHW (2011) A Bayesian view on Cryo-EM structure determination. J Mol Biol 415:406–418
16. Scheres SHW (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol 180:519–530
17. Scheres SHW, Gao H, Valle M, Herman GT, Eggermont PPB, Frank J, Carazo JM (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat Methods 4:27–29
18. Shaikh TR, Gao H, Baxter WT, Asturias FJ, Boisset N, Leith A, Frank J (2008) SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat Protoc 3:1941–1974
19. Sigworth FJ, Doerschuk PC, Carazo JM, Scheres SH (2010) An introduction to maximum-likelihood methods in Cryo-EM. Methods Enzymol 482:263–294
20. Sorzano COS, Marabini R, Velazquez-Muriel J, Bilbao-Castro JR, Scheres SH, Carazo JM, Pascual-Montano A (2004) XMIPP: a new generation of an open-source image processing package for electron microscopy. J Struct Biol 148:194–204
21. Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, Stagg S, Potter CS, Carragher B (2005) Automated molecular microscopy: the new Leginon system. J Struct Biol 151:41–60
22. Wade RH (1992) A brief look at imaging contrast transfer. Ultramicroscopy 46:145–156
23. Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103
24. Yang Z, Fang J, Chittuluru J, Asturias FJ, Penczek PA (2012) Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure 20:237–247