Unsupervised Cryo-EM Data Clustering through Adaptively Constrained K-Means Algorithm

Yaofang Xu[1], Jiayi Wu[2], Chang-Cheng Yin[1]*, Youdong Mao[2,3]*

1 Department of Biophysics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China, 2 State Key Laboratory of Artificial Microstructure and Mesoscopic Physics, Institute of Condensed Matter Physics, School of Physics, Center for Quantitative Biology, Peking University, Beijing, China, 3 Intel Parallel Computing Center for Structural Biology, Dana-Farber Cancer Institute, Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, United States of America

* youdong_mao@dfci.harvard.edu (YM); ccyin@hsc.pku.edu.cn (CY)

Abstract

In single-particle cryo-electron microscopy (cryo-EM), K-means clustering algorithm is widely used in unsupervised 2D classification of projection images of biological macromolecules. 3D *ab initio* reconstruction requires accurate unsupervised classification in order to separate molecular projections of distinct orientations. Due to background noise in single-particle images and uncertainty of molecular orientations, traditional K-means clustering algorithm may classify images into wrong classes and produce classes with a large variation in membership. Overcoming these limitations requires further development on clustering algorithms for cryo-EM data analysis. We propose a novel unsupervised data clustering method building upon the traditional K-means algorithm. By introducing an adaptive constraint term in the objective function, our algorithm not only avoids a large variation in class sizes but also produces more accurate data clustering. Applications of this approach to both simulated and experimental cryo-EM data demonstrate that our algorithm is a significantly improved alterative to the traditional K-means algorithm in single-particle cryo-EM analysis.

Introduction

Single-particle reconstruction in cryo-electron microscopy (cryo-EM) is a powerful technology to determine three-dimensional structures of biological macromolecular complexes in their native states [1]. Recent advances in direct electron detector and high-performance computing enabled 3D structural determination of biological macromolecular complexes at a near-atomic resolution [2–4]. The goal of single-particle reconstruction is to recover the 3D structure of a macromolecule from a large number of 2D transmission images, in which the macromolecules assume random, unknown orientations.

Due to high sensitivity of biological samples to radiation damage by electron beam, cryo-EM data are often acquired with very limited electron doses (10–50 electron/$Å^2$), which makes the cryo-EM images extremely noisy. To determine the relative orientations of molecular projections, a crucial step is to classify 2D projection images in an unsupervised fashion such that

the images in the same class come from similar orientations of projection [5, 6]. For each class, images are aligned, centered and averaged to produce class averages with enhanced signal-to-noise ratio (SNR). Generating 2D class averages is important for both common-line based 3D *ab initio* reconstruction [7–14] and some modern methods [15, 16]. Unsupervised classification is also useful for a quick evaluation on structural heterogeneity and quality of samples before entering time-consuming 3D refinement steps [17, 18].

If ignoring conformational dynamics of imaged macromolecules, the intrinsic difference among projection images mainly comes from two sources: projection direction and in-plane rotation. Prior to classification, single-particle images must be aligned to minimize the differences in their translation and in-plane rotation. There are two popular approaches for initial classification of 2D projection images, namely, multi-reference alignment (MRA) [19] and reference-free alignment (RFA) [20]. In MRA, a 2D image alignment step and a data-clustering step are performed iteratively until convergence. In the 2D image alignment step, each image is rotated and shifted incrementally with respect to each reference. All possible correlations between a rotated, translated image and a reference are computed. The distance between an image and a reference is defined as the minimum of all correlation values between them. Based on these distances, in the data-clustering step, traditional K-means clustering is used to classify all images into many classes. An implementation of the MRA strategy can be found in SPARX [21]. In RFA, all images are first aligned globally, which attempts to find rotations and translations for all images that minimize the sum of squared deviation from their mean. These aligned images are used as the input for data-clustering algorithms. This strategy was implemented in SPIDER [22].

Moreover, upon the suggestion of Jean-Pierre Bretaudière, multivariate statistical analysis (MSA) was introduced into cryo-EM [6, 23, 24]. MSA reduces the dimensionality of images by projecting them into a subspace spanned by several eigenvectors, which are also called features. Reducing dimensionality not only accelerates computing but also denoises projection images. The resulting features can also be used as the references for image alignment. For example, EMAN2 combines MSA with MRA (MSA/MRA) (see its script e2refine2d.py) [25]. It first generates translational and rotational invariants for initial classification. Then, a MSA step is iterated with a MRA step, in which images are aligned to those features and classified by the K-means algorithm, until a pre-defined number of iterations is reached.

Another approach to eliminate in-plane rotation is the use of rotationally invariant transformation (RIT), which transforms two images into the same invariant map if they only differ with respect to in-plane rotation. Then the transformed images are classified and averaged as there is no more in-plane rotation. However, finding such a transform is not trivial. Several attempts have been made [26]. The disadvantage is that the invariant transformation only preserves amplitude information, which may cause misclassification in the later steps. Recently, a new RIT approach was proposed [27] to preserve both amplitude and phase information. After RIT, the difference between each pair of invariant maps is used to measure the difference between the corresponding images. Then each image is denoised by averaging all the images in its neighbor. Thus, the n input images are transformed into n denoised images that are used to compute an *ab initio* model. Moreover, a Covariance Weiner Filter was recently developed to do CTF correction and denoising on individual images directly [28], making it possible to examine particles without 2D classification. However, this method does not negate the necessity of unsupervised data clustering.

Despite the aforementioned progress in data science, the traditional K-means algorithm remains one of the most popular data clustering approaches for single-particle cryo-EM, not only in 2D classification but also in the modern 3D analysis method [29]. However, the traditional K-means algorithm has certain limitations. A class average with a higher SNR correlates

preferably with noise in the high-frequency domain, resulting in attraction of more images into high-SNR classes. Therefore, when used with MRA, traditional K-means clustering tends to misclassify single-particle images to classes with more members [30]. Moreover, some classes may be depleted during iterations. Reseeding empty classes may tentatively remedy this problem. However, it can also break the balance of class sizes among distinct classes, resulting in the coexistence of both oversized and undersized classes. The same issue was also found in multi-reference maximum-likelihood classification, implemented in XMIPP and RELION [31–33].

To avoid class size getting improperly large, several approaches were proposed [22, 30, 31, 34]. First, a modified traditional K-means was implemented in SPIDER for data clustering, where the objective function is multiplied with a factor $\sqrt{s/(s \pm 1)}$. Here, s is class size. '–' is adopted when an image is compared with its own class average; and '+' is adopted when the image is compared to other class averages. For very large classes, this factor is almost 1, whereas it is well below one for small classes. Therefore, it tends to classify more images to small classes. This modification does avoid generating empty classes, but it does not exclude over-sized classes or ones containing only one image (We refer this modified traditional K-means in SPIDER as the traditional K-means in the later context, when there is no ambiguity). Second, an algorithm called equally-sized group K-means (EQK-means) was developed to control the sizes of classes [34], which was implemented in SPARX. In each iteration of EQK-means, every class is forced to have the same number of image members, which avoids the attraction of images to high-SNR classes. Therefore, the resulting class averages may achieve comparable SNRs. However, given that the experimental projection directions can be hardly evenly distributed, it is problematic to produce equally sized classes. Classes in denser angular areas should be assigned with more image members than classes in sparser areas. Third, a modified MRA approach using the CL2D algorithm [30] was implemented in XMIPP [33], which classifies images hierarchically. At each hierarchical level, images are classified with a control of class size by dividing large classes into many smaller classes. The hierarchical approach conducts classification at each level and may require more CPU time than non-hierarchical approaches.

In addition, outside of the cryo-EM field, there were several studies on K-means clustering with size constraints. Banerjee *et al.* proposed a method with three steps [35]. The algorithm is started by clustering a sampled subset of an original dataset. Then the remaining data are populated to the initial clusters satisfying the cluster size constraint motivated by stable marriage problem. Last, the clustering is refined by moving one point or a pair of data points that keep the size constraint. An equivalent implementation of this algorithm in SPA is EQK-means [34], where it starts clustering by randomly selecting k centroids and populates the remaining data by using distance matrix. Zhu *et al.* also proposed a size-constrained algorithm [36]. In their algorithm, the data are first clustered by existing clustering algorithms. The size constraint is then enforced when minimizing the difference between the old partition matrix and the new partition matrix. To find a more balanced clustering results, Malinen *et al.* proposed a method solving the assignment problem by Hungarian algorithm [37]. With the size constraint, it can find a good solution with a cost of $O(n^3)$ complexity. All the above algorithms were developed for direct marketing, document clustering, energy aware sensor networks, etc., with predefined cluster sizes from known information. However, in SPA one expects the clustering to be approximately uniform so that the SNRs of different classes are comparable. In fact, we do not know the exact number of particles that a class should have.

In this study, we introduce a novel data clustering approach, named adaptively constrained K-means algorithm (ACK-means), for unsupervised cryo-EM image classification. Different from EQK-means that enforces an equal size on all classes or other clustering algorithms with predefined class sizes, ACK-means controls the class size with an adaptive balance between

class size and classification accuracy. Thus, ACK-means can in principle avoid an excessive growth of class size while producing a more accurate angular assignment. Our study suggests that ACK-means is a significantly improved alternative to the traditional K-means for cryo-EM data clustering.

Methods

A brief review on the traditional K-means algorithm

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent a set of projection images to be classified. Each class is represented by a centroid μ_j, for $j = 1, 2, \ldots, k$. The goal is to partition X into k classes so that the following objective function is minimized:

$$J_k = \sum_{j=1}^{k}\sum_{p_i=j} dissim(x_i, \mu_j),$$

where $dissim(\cdot,\cdot)$ is a function measuring the dissimilarity between images x_i and μ_j. The partitioning is denoted by an assignment vector $p = (p_1, p_2, \ldots, p_n)$, which assigns image x_i to the p_i-th class. The commonly used dissimilarity measure is the Euclidian distance [38]. In most MRA approaches, the dissimilarity is defined as a minimum over all possible relative rotations and translations of an image with respect to another.

To solve this minimization problem globally is NP-hard [39]. As a local minimum solution, the traditional K-means algorithm was first developed by MacQueen [40]. He gave the name "K-means" to the algorithm that assigns each image to the class of the nearest centroid. This can be formulated in the following:

1. **Initialization step**: Determine k initial centroids (seed points) $\{\mu_j\}_{j=1}^{k}$ by randomly selecting k images from X;

2. **Assignment step**: For each image, assign it to the class specified by the most similar centroid;

$$p_i = \arg\min_{j} dissim(x_i, \mu_j)$$

3. **Update step**: Recalculate the centroid by the image mean in each class:

$$\mu_j = mean(\{x_i : p_i = j\})$$

4. Repeat (2) and (3) until there are no more changes of membership.

Adaptive constraint

To introduce an adaptive constraint to K-means clustering, we add an additional term to the objective function as shown in the following expression:

$$J = \sum_{j=1}^{k}\sum_{p_i=j} dissim(x_i, \mu_j) + \lambda ss^{\mathrm{T}}. \tag{1}$$

The first term is the sum of dissimilarity between image $\boldsymbol{x}_i$ and centroid $\boldsymbol{\mu}_j$. In the second term, $\boldsymbol{s}$ is a vector, whose element s_i denotes the number of images belonging to the i-th class; and λ is a non-negative parameter.

Note that the sum of all the elements of s is the total number of images, which is a constant n. According to the Cauchy-Schwarz inequality, the second term is minimized only when $s_1 = s_2 = \ldots = s_k$. Therefore, by introducing the second term, we establish a competition between the dissimilarity and the balance of class sizes. If we set $\lambda = 0$, no constraint is exercised on class sizes and minimizing expression (1) has the same effect as that is achieved by the traditional K-means algorithm. If $\lambda = +\infty$, all classes would have the same size. As λ is changed from 0 to $+\infty$, more weight is given on the balance of class sizes. It allows us to tune the class sizes adaptively by regulating the strength of the constraint, as opposed to the EQK-means algorithm that enforces equally sized classes [34]. For this reason, we call the second term an adaptive constraint. Thus, the proposed algorithm allows more images to update their memberships in a large class than in a small class during the optimization of the objective function.

For a given set of centroids, n images are assigned to k classes one by one through minimizing the objective function (1). Suppose that at the end of the previous iteration, image $\boldsymbol{x}_i$ is assigned to class p_i and that the class size vector is $\boldsymbol{s} = (s_1, s_2, \ldots, s_k)$. In the current iteration, the class size vector is first recalculated as $\boldsymbol{s}'$ by omitting $\boldsymbol{x}_i$. Almost all the elements of $\boldsymbol{s}'$ is the same as $\boldsymbol{s}$, except $s'_{p_i} = s_{p_i} - 1$. Then image $\boldsymbol{x}_i$ is reassigned to class p_i by solving the following minimization:

$$p_i = \arg\min_j \{ dissim(\boldsymbol{x}_i, \boldsymbol{\mu}_j) + \lambda \sum_{h=1}^{k} [s'_h + \delta_{hj}]^2 \}$$

$$= \arg\min_j \{ dissim(\boldsymbol{x}_i, \boldsymbol{\mu}_j) + 2\lambda s'_j + \lambda + \lambda \sum_{h=1}^{k} s'^2_h \}$$

$$= \arg\min_j \{ dissim(\boldsymbol{x}_i, \boldsymbol{\mu}_j) + 2\lambda s'_j \},$$

where δ_{hj} is the Kronecker delta function. When all the images are reassigned, we end up with a new assignment vector $\boldsymbol{p}$ describing the updated partition in the current iteration. To further minimize the objective function (1), we update the k centroids by averaging images in the same class. This process is iterated until there are no more changes in membership.

Characteristic dissimilarity

Due to background noise and variation in molecular projections, the scale of pixel intensities in single-particle images is expected to vary from case to case. To keep the competition between the two terms of Eq (1) at the same magnitude, one needs a larger λ for an image dataset with large dissimilarities than that with small dissimilarities. Therefore, we developed a strategy to tune the value of λ that is applicable to varying scales. One quantity reflecting data scaling is the maximum value of dissimilarity between any pairs of images in given dataset. However, computing all the dissimilarities between any image pair is extremely time-consuming and practically prohibited. Instead, because we already computed dissimilarities between images and centroids during image assignment to different classes, we can construct a quantity called "characteristic dissimilarity" from these values. In each iteration, we randomly select 10 images. For each one of them, we find its smallest and largest dissimilarities among the k dissimilarities with k centroids. The difference between the largest and smallest dissimilarities is calculated for each image, and is then averaged together to make characteristic dissimilarity d_c.

Hence, we rewrite 2λ as:

$$2\lambda = \beta \frac{d_c}{\lfloor n/k \rfloor},$$

where β is a free parameter in the range from 0 to $+\infty$. So we decompose 2λ into two parameters. The first parameter d_c describes the change of pixel intensity scaling, whereas the second parameter β decides the weight on the class size whose value is independent of data scaling. The constant $\lfloor n/k \rfloor$ is the class size if all images are partitioned equally. Hence, if the partition of all classes is ideally balanced, $\Delta s'_j / \lfloor n/k \rfloor$ represents the fraction of membership-changed images in the j-th class. Given $0 < \beta < +\infty$, images are partitioned in accordance to dissimilarity while the class sizes are monitored by the adaptive constraint. Note that the only parameter to be considered during the application of ACK-means is β. The smaller β is, the less balance of class sizes. Our experiments show that ACK-means can generate satisfying results with $\beta = 0.5$ (see below). In this case, if d_c is the diameter of the area occupied by the data, then βd_c is the radius.

Implementation algorithm

The algorithm of adaptively constrained K-means is implemented in the following pseudo code.

```
Algorithm: ACK-means
Input:
{x_i}_{i=1}^n : Set of data points.
k: Number of classes.
σ_0: The minimum fraction of data points that are changed membership.
β: The weight on the adaptive constraint term.
1: Initialize centroids {μ_j}_{j=1}^k by randomly selecting k data points from input
data.
2: Compute p according to p_i = arg min_j dissim(x_i, μ_j).
3: Update {μ_j}_{i=1}^k by averaging data points in the same class.
4: while σ > σ_0 do
5:   Randomly select 10 images {r_m}_{m=1}^{10} and compute t_m for m = 1:10 according to:
```

$$t_m = \max_j \; dissim(r_m, \mu_j) - \min_j \; dissim(x_m, \mu_j)$$

```
6:   Compute d_c according to d_c = \sum_{m=1}^{10} t_m / 10
7:   Compute λ according to λ = β \frac{d_c/2}{\lfloor n/k \rfloor}.
8:   Save the old assignment vector according to p_old = p
9:   for i = 1: n do
10:      Compute s' according to s'_l = \sum_{m \neq i} δ_{p_m, l}
11:      Update p_i according to p_i = arg min_j dissim(x_i, μ_j) + 2λs'_j
12:   end for
13:   Update {μ_j}_{i=1}^k by averaging data points in the same class.
14:   Compute σ according to σ = 1 - \frac{1}{n} \sum_{i=1}^{n} δ_{p_i^{old}, p_i}
15: end while
Return: assignment vector p.
```

Note that we exercise no assumption regarding classification and initialize the algorithm as what the traditional K-means does. To ensure an unsupervised nature of the classification, k images are randomly selected from the dataset as the initial centroids. This guarantees that

there is at least one member in each class. Then, each of the rest images is assigned to the class whose centroid is the nearest to the image. To devise a termination criterion for the algorithm, we set a threshold parameter σ_0 (usually set as 0.01) here. If the fraction of data points changing their membership in the current iteration decreases to this threshold, the algorithm is terminated.

Since adding the adaptive term costs computational time linear to the number of images, the complexity of ACK-means is at the same level as the traditional K-means, namely, $O(knt_{dis} + 10k^2)$. However, EQK-means sorts all the distances after they are computed. The complexity of EQK-means is $O(knt_{dis} + k^2n^2)$, where t_{dis} is the time for computing the dissimilarity between a pair of images. In the MRA approach, images are aligned before their distances are computed. The alignment process may cost considerable time depending on the dimension of images. Thus, in MRA knt_{dis} contributes substantially to the complexity. In the MSA/MRA or RFA approach, images are well aligned before entering clustering algorithm, the calculation of t_{dis} is faster. Generally, ACK-means shares the complexity of the traditional K-means.

Benchmark with simulated data

The density map of Escherichia coli 70S ribosome [41] was used to generate 10,000 simulated projection images (S1 Fig). Most protein structures are of lower symmetry or asymmetric. Therefore, some orientations are expected to appear more frequently than others in vitreous ice. To emulate this phenomenon, we uniformly chose 100 orientations covering half a sphere. Each orientation is regarded as a Gaussian center, around which 100 projections were generated with a Gaussian distribution. Due to electron lens aberrations and defocusing, we further modified the projection images with the contrast transfer function (CTF). The projections were then additively contaminated with Gaussian noise at different SNR = 1/3, 1/10, 1/30 (S2 Fig), which allowed us to investigate the proposed algorithm at different noise level. The input projections to all experiments were CTF-corrected by phase flipping [42].

To examine the performance of our algorithm, we compared the results of classifying the 10,000 simulated images into 100 classes by using ACK-means with those from other existing approaches. For the standard MRA, we compared ACK-means against traditional K-means and EQK-means algorithms implemented in SPARX. The script isac.py in SPARX is part of a method called ISAC (Iterative Stable Alignment and Clustering) proposed in [34], consisting of the standard MRA part, followed by analysis within classes. The within-class analysis traces the change of membership of images and selects stably classified images that do not change their membership in each iteration. To focus on the effect of different data clustering algorithms, only the MRA part is used in our test. To see the influence of ACK-means in MRA with MSA (MRA/MSA), we replaced the traditional K-means with ACK-means of e2refine.py in EMAN2 and compared their performance. For RFA, we followed the protocol of SPIDER [41] and compared the classification results of the traditional K-means with those of ACK-means.

Results

Simulated data

Since all the original angles of the input projections are known, the angular difference between any pair of projections in each class, also termed "angular distances", can be computed. The statistical behavior of the angular distances can be used to measure the quality of the corresponding class [30]. A class assigned with n image members has $\binom{n}{2}$ pairs of angular

distances. We used two plots to compare the results from different algorithms. The first plot is the histogram distribution of the angular distances from all classes [30]. A better algorithm is expected to exhibit a distribution curve with a sharper, higher peak at lower angular distances. The second plot ranks the sizes of all classes to show the balance of classification. As shown in Fig 1, we compared the unsupervised classification results by ACK-means on the simulated data with a SNR of 0.1 with those obtained by several existing K-means implementations: the traditional K-means and EQK-means in the standard MRA approach implemented in SPARX [21, 34], the traditional K-means in the MRA/MSA approach implemented in EMAN2 [25], and the one in the RFA approach implemented in SPIDER [22].

In the MRA approach, although EQK-means avoids the attraction of dissimilar images by delivering equally sized classes, it exhibits reduced angular accuracy of classification (Fig 1A and 1B). By contrast, ACK-means makes little compromise on the balance of class size, yet improves the classification accuracy. In both MRA and MRA/MSA approaches, ACK-means gives rise to a prominent improvement in both the classification accuracy and the balance of class sizes (Fig 1A–1D). However, in the RFA approach, although the improvement of classification accuracy is not obvious (Fig 1E), ACK-means still generates more balanced class sizes (Fig 1F).

The three experiments behave differently as SNR is changed (S3 and S4 Figs). Since MRA has the strongest effect of attraction of dissimilar particles over other approaches, the attraction effect becomes much stronger with decreasing SNR. By contrast, the performance of ACK-means in controlling the class size adaptively does not degrade with decreasing SNR (Fig 1A and 1B; S3A, S3B, S4A and S4B Figs). In the MRA/MSA approach, ACK-means outperforms the traditional K-means at moderately low SNR level (Fig 1C and S3C Fig). However, at lower SNR (0.033), their difference in the histogram disappears (S4AC Fig). This is likely because alignment errors introduced in the early step cannot be eliminated by ACK-means in the later step. In the RFA approach, ACK-means generates similar classification accuracy with the traditional K-means at all noise levels (Fig 1E, S3E and S4E Figs). This result confirms that the classification accuracy is bound by the alignment error. In all cases, ACK-means gives rise to a well-balanced classification (S4D Fig).

We also compared the classification results of ACK-means under different approaches with those of the maximum-likelihood-based 2D classification in RELION [31] (S5 Fig). ACK-means in both MRA and RFA outperforms the results by RELION at all noise levels tested (S5A, S5B, S5E and S5F Fig). However, the MRA/MSA approach does not improve over RELION at a higher noise level (S5C and S5D Fig). In all the cases tested, RELION generates many classes with no or few particles, which might prevent the resulting class averages from supporting initial reconstructions with sufficient quality.

To examine the convergence behavior of the ACK-means algorithm, we recorded the fraction of images changing membership in each iteration relative to the previous iteration. The curve generated from the above experiments in the MRA approach is plotted in Fig 2. Other approaches have similar results. All three algorithms converge exponentially in a similar way. The ACK-means and the traditional K-means uses less iterations to achieve the same degree of convergence as does EQK-means. We also examined the total run time for the above experiments with simulated data to show the time cost of theses algorithms in different approaches (Table 1). For each iteration, the ACK-means algorithm used slightly more time than others, which is nonetheless affordable.

Experimental cryo-EM data

Three real experimental datasets were used to examine our ACK-means algorithm in this study. We compared the results of our algorithm against the traditional K-means
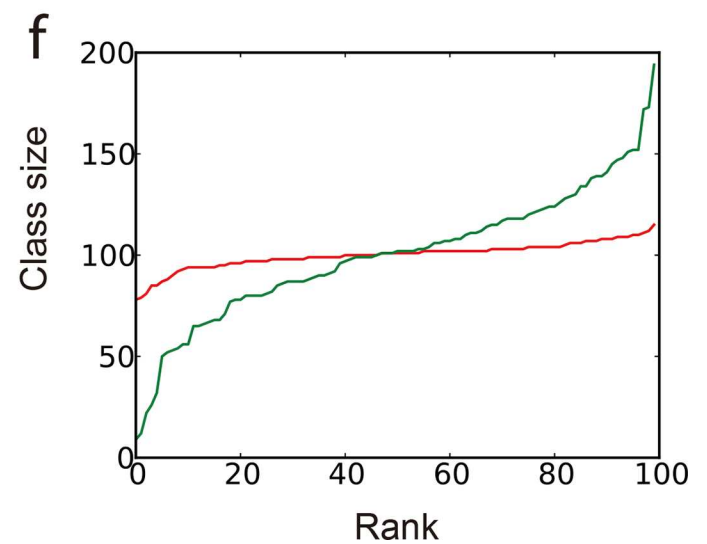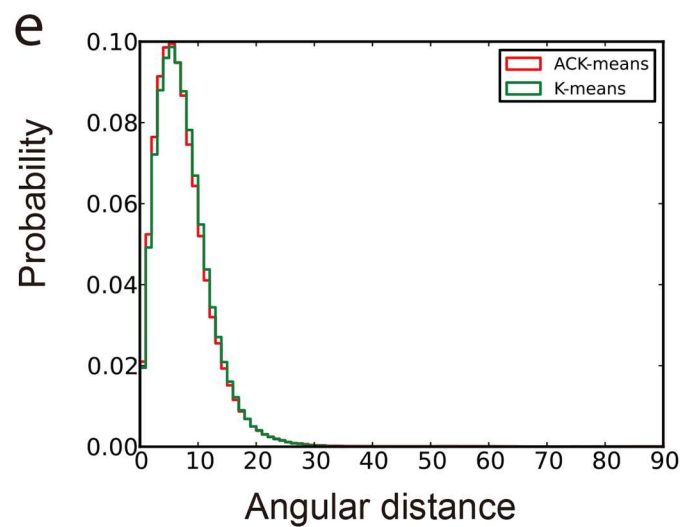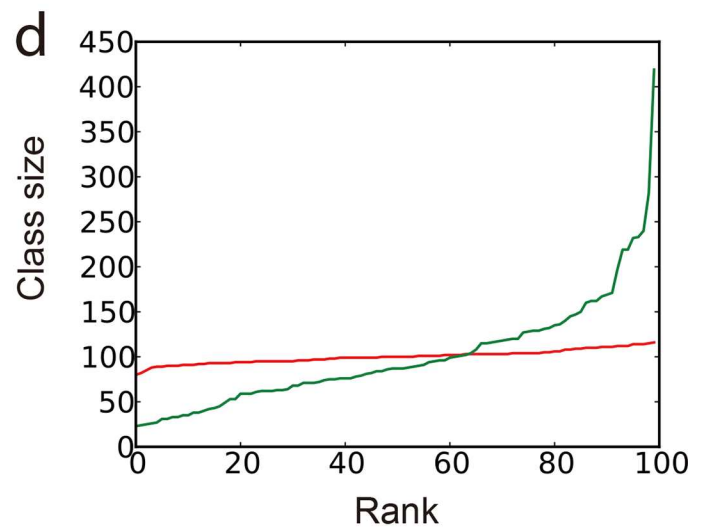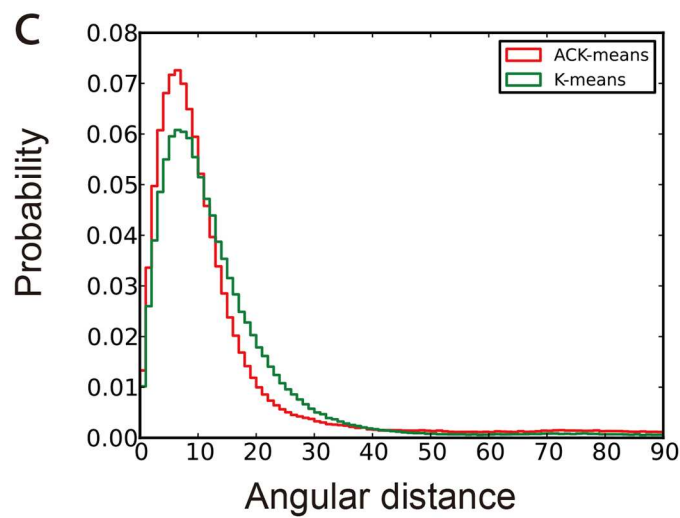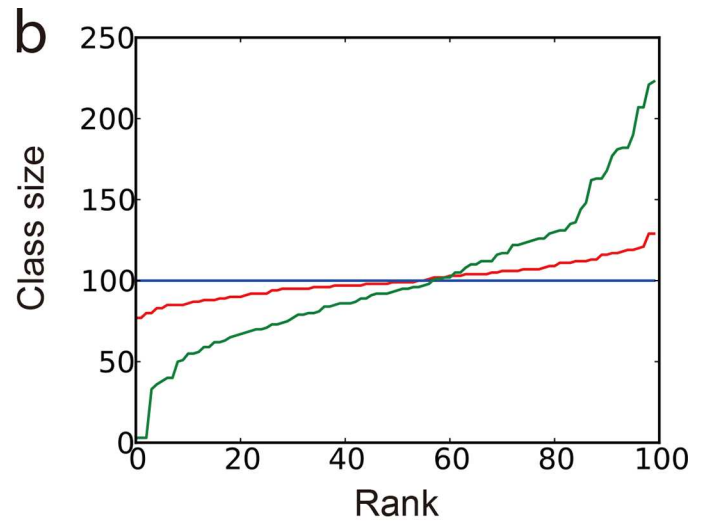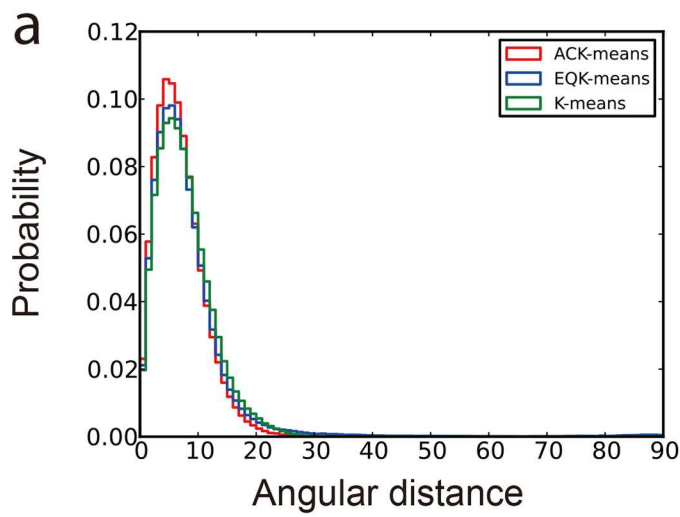
Fig 1. Comparison of classification results of simulated data with SNR = 1/10. The First column (panels a, c and e) is the normalized histogram of angular distances. More accurate classification produces curve with higher peak concentrated at lower angular distance. The second column (panels b, d and f) shows the class sizes arranged in an ascend order. The most balanced classification has a horizontal line in this plot. (a) and (b) are from experiments using different clustering algorithms in MRA approach under SPARX. (c) and (d) are from experiments using different clustering algorithms in MRA/MSA approach under EMAN2. (e) and (f) are from experiments using different clustering algorithms in RFA approach under SPIDER. In all graphs, red curves present the results from the ACK-means algorithm.

implementations in SPARX, EMAN2 and SPIDER, as well as EQK-means in SPARX. Since it is impossible to know the true projection angles of individual single particles, we evaluate the classification results by inspecting the quality of 2D class averages visually.

Case 1: GroEL. The first dataset consists of 5,000 particles selected from a GroEL dataset of 26 micrographs, whose pixel size is 2.10 Å/pixel. The size of particles is 140 × 140 pixels [43]. These particles were first phase-flipped and then classified into 25 classes by different algorithms. As shown in Fig 3, a set of 2D class averages were computed with the traditional K-means (Fig 3A), EQK-means (Fig 3B) and ACK-means (Fig 3C) through the MRA protocol in SPARX. The traditional K-means did not control class sizes and produced more blurred classes than other two algorithms (Fig 3A). Although EQK-means and ACK-means both produced balanced results, EQK-means generated two worst class averages among all class averages (Fig 3B). For the MRA/MSA approach in EMAN2, class averages from the traditional K-means (Fig 4A) and ACK-means (Fig 4B) were compared. ACK-means generated clearer class averages with balanced sizes. Furthermore, comparison between the traditional K-means (Fig
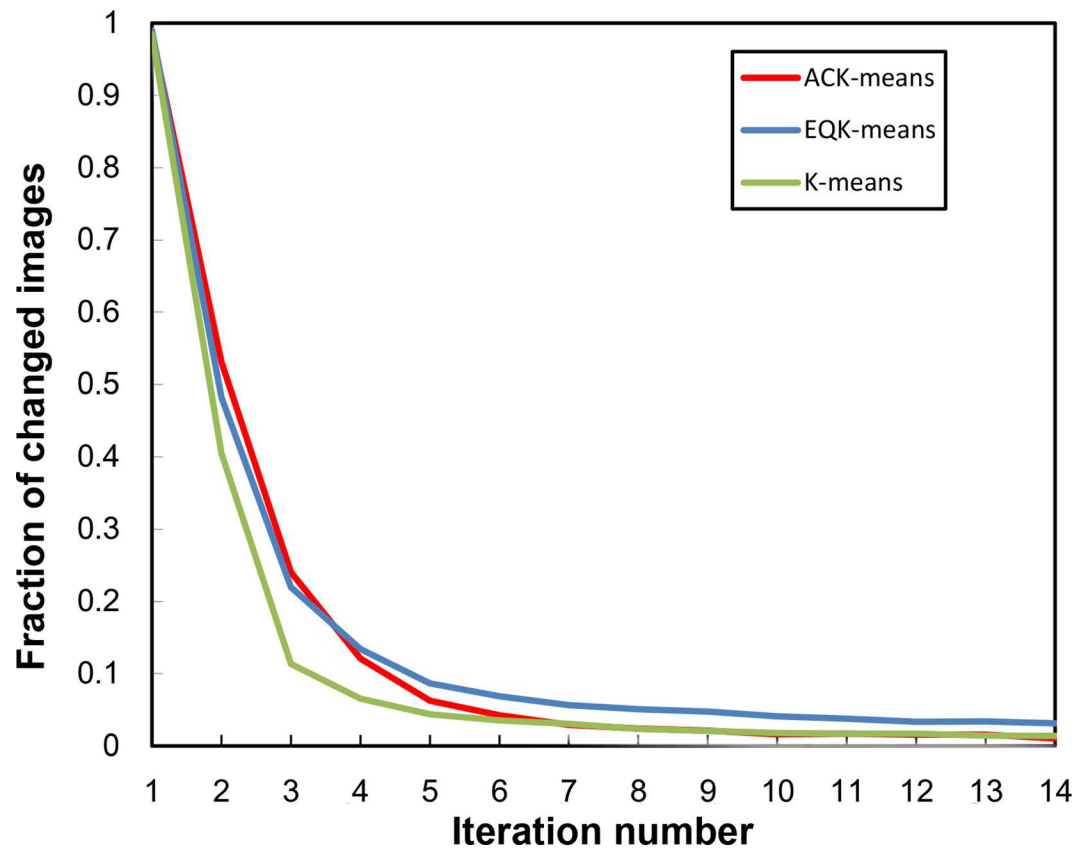

Fig 2. Convergence of K-means, EQK-means and ACK-means in MRA. The three algorithms behave similarly as iteration increases, converging very fast at the first several iterations.

Table 1. The running time of different algorithms in different approaches.

Approach	MRA, SPARX (48 cores)			MSA/MRA, EMAN2 (1 core)		RFA, SPIDER (1 core)	
Algorithm	EQK-means	K-means	ACK-means	K-means	ACK-means	K-means	ACK-means
Time (in hour)	0.3h	0.25h	0.45h	7h	12h	0.8h	1.3h

doi:10.1371/journal.pone.0167765.t001

5A) and ACK-means (Fig 5B) were made with the RFA approach in SPIDER. Both the traditional K-means and ACK-means generated class averages of comparable quality, but the latter substantially improved the balance of class sizes, avoiding both oversized and empty classes.

Case 2: Inflammasome.   We used 281 particles of inflammasome to benchmark our algorithm [44]. The data were collected with a pixel size of 1.72 Å/pixel and an acceleration voltage of 200 kV. The particles have a size of 160×160 pixels. These particles were pre-selected such that only side views with different lengths, corresponding to different oligomeric states of inflammasome, were included in the dataset. After phase-flipped, the dataset was classified into 20 classes using different algorithms. In the MRA approach with SPARX, the class averages were generated by the traditional K-means (Fig 6A), EQK-means (Fig 6B) and ACK-means (Fig 6C). Without the constraint on class sizes, many classes in the traditional K-means were not assigned with enough particles, producing blurred class averages (Fig 6A). Although many class averages of EQK-means and ACK-means are similar, some classes of EQK-means present misaligned features, indicating the failure of classifying different particles. Similarly, when compared in the MRA/MSA approach with EMAN2, ACK-means produced generally improved classification results than did the traditional K-means (Fig 7A and 7B). We further compared our approach with the traditional K-means in the RFA approach implemented in SPIDER. The traditional K-means generated many classes with only one particle (Fig 8A). By contrast, this was well avoided in the results from unsupervised classification by our ACK-means algorithm (Fig 8B).

Case 3: Proteasome.   Finally, we used a dataset containing proteasomal RP (regulatory particle) and RP-CP (regulatory particle associated with core particle) subcomplex [45]. The total number of particle is 3,960, with pixel size 2.00 Å/pixel and particle size 160×160 pixels. All particles were pre-processed by phase-flipping and classified into 40 classes. For the MRA approach in SPARX, the class averages generated by the traditional K-means, EQK-means and ACK-means are shown in Fig 9A–9C, respectively. The traditional K-means (Fig 9A) generated many blurred classes because of no constraint on class sizes. ACK-means (Fig 9C) and EQK-means (Fig 9B) yielded comparable results. For the MRA/MSA approach in EMAN2 and RFA approach in SPIDER, we compared the class averages of traditional K-means (Figs 10A and 11A) and ACK-means (Figs 10B and 11B). Class averages of ACK-means show more classes with clear details. Additionally, the traditional K-means in SPIDER generated 3 classes with only one particle.

Discussion

In this study, we propose a new data-clustering algorithm, which generates adaptively balanced, unsupervised classification, preventing the attraction of dissimilar particles into classes of large sizes or higher SNRs. This allows significant improvement in unsupervised image classification over the traditional K-means algorithm. Meanwhile, by controlling class sizes adaptively, our approach also improves angular accuracy of image clustering as compared to EQK-means, allowing more particles to be assigned to a class if the operation can improve classification accuracy.
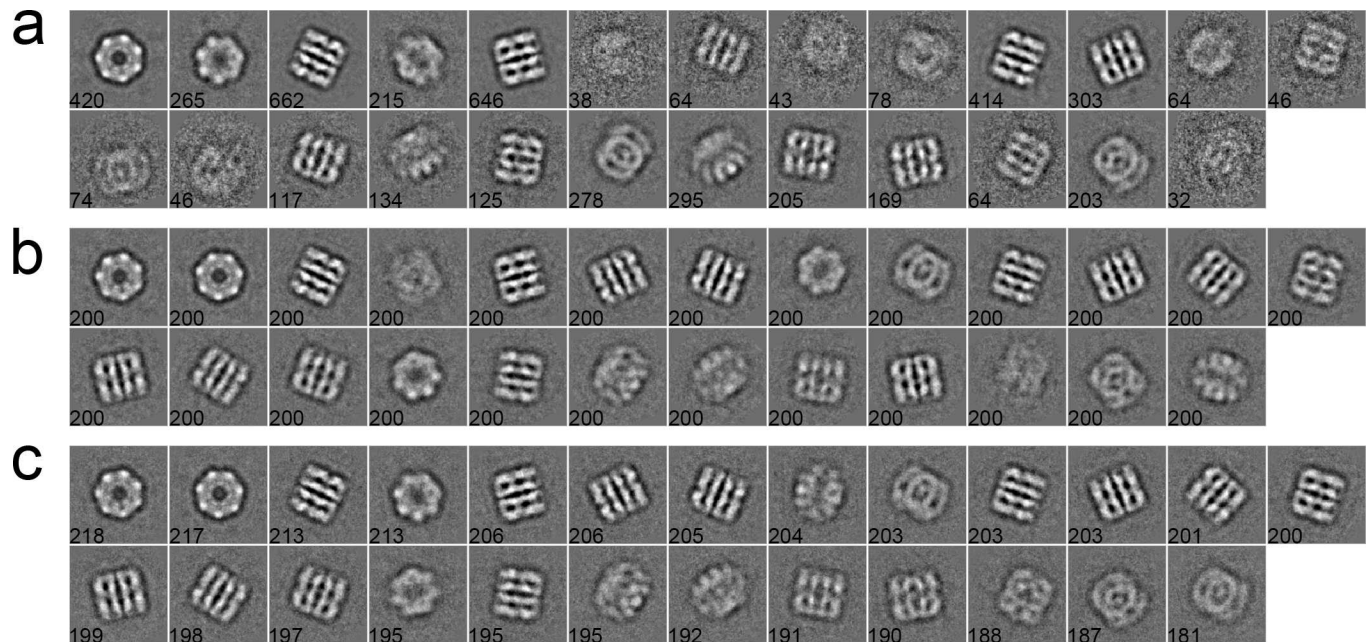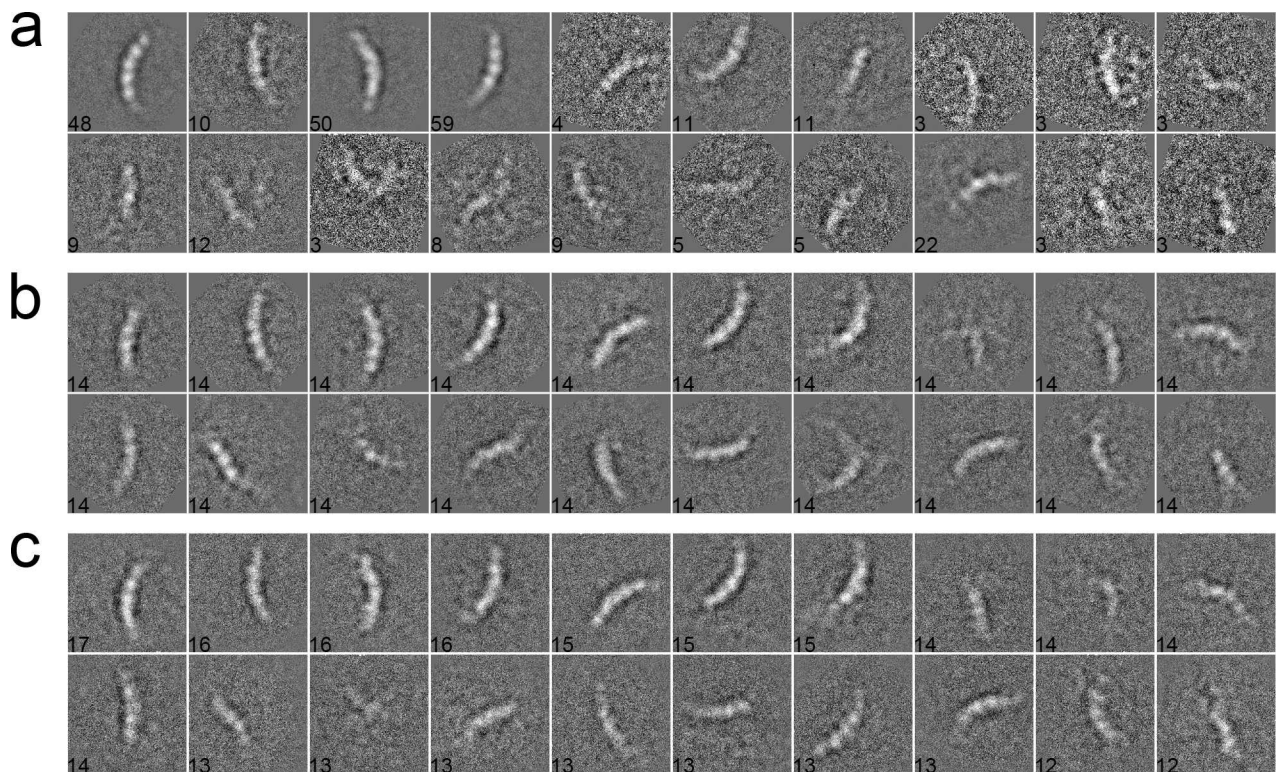
Fig 3. 2D class averages of GroEL using the traditional K-means (a), EQK-means (b) and ACK-means (c) in MRA approach from SPARX. Class size is shown at the left bottom of each class average. ACK-means (b) is the best by having the most number of clear classes.

We tested our algorithm with both simulated and real experimental datasets. The projection orientations of simulated data were generated with both dense and sparse angular areas to imitate realistic situation. We found that our ACK-means algorithm consistently outperforms the traditional K-means in all cases. In MRA, traditional K-means suffers from attracting dissimilar particles into classes with more particles. It does not control class sizes and often generates many classes of very few image members, resulting in blurred class averages. ACK-means and EQK-means both generate balanced class averages, whereas ACK-means gives rise to improved classification accuracy, allowing more details to be recovered in class averages. In contrast to EQK-means that avoids the growing of class sizes by forcing each class to have the
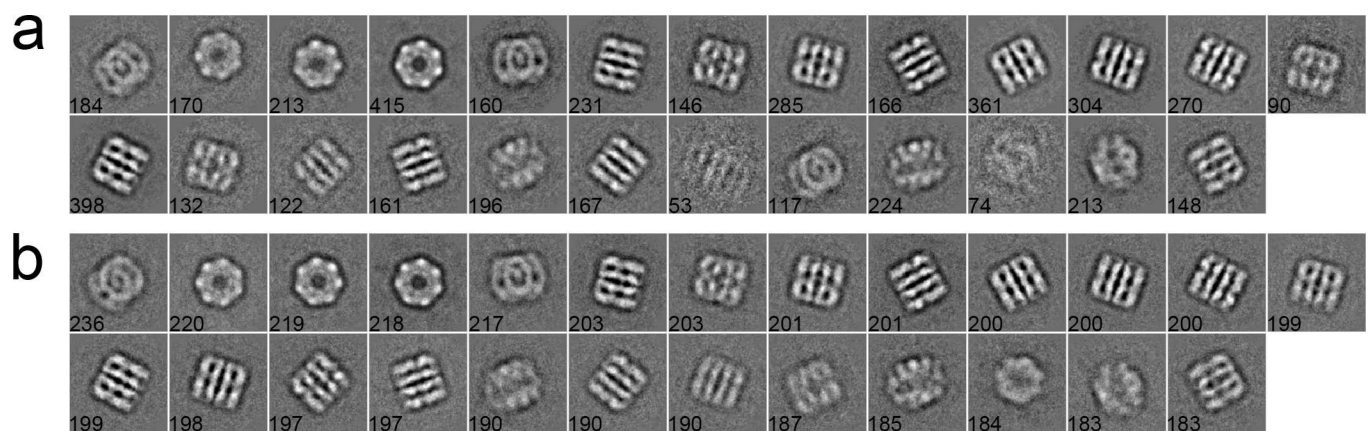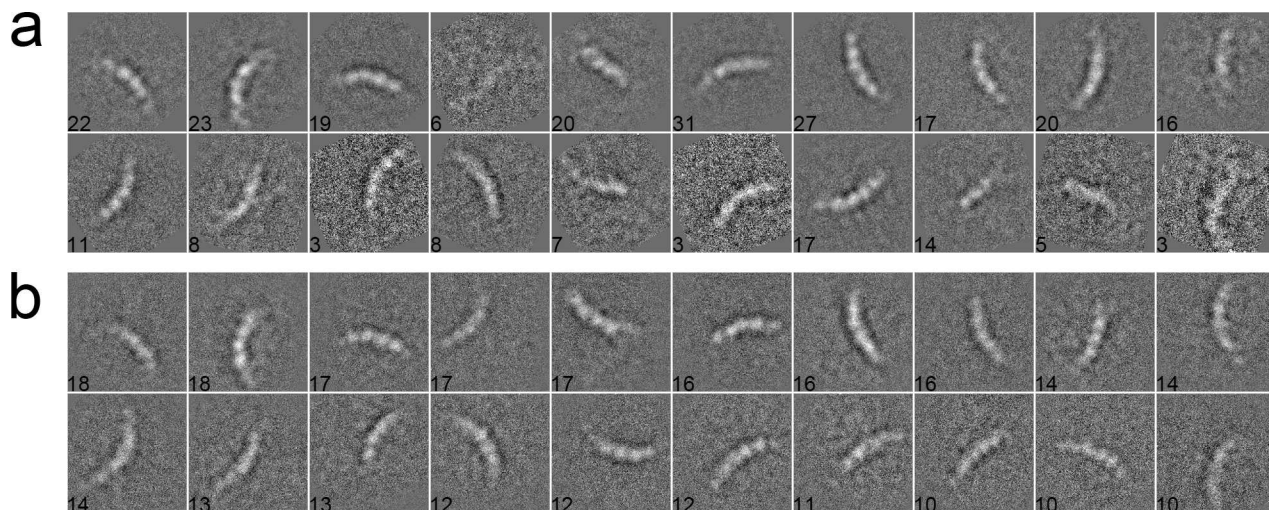

Fig 4. 2D class averages of GroEL using the traditional K-means (a) and ACK-means (b) in MRA/MSA from EMAN2. Class size is shown at the left bottom of each class average. Their performance is similar, but ACK-means (b) has the more number of clear classes.
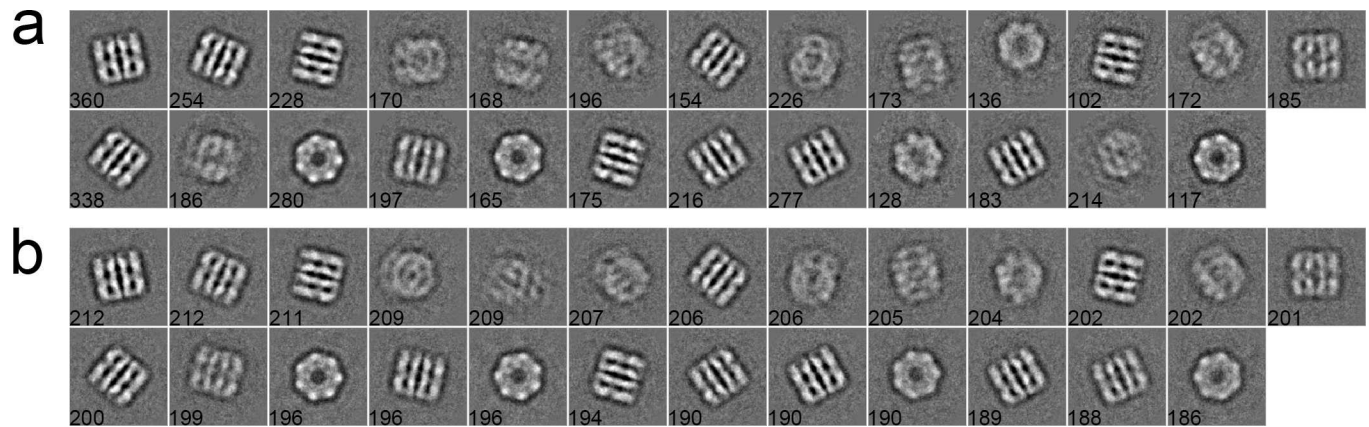
Fig 5. 2D class averages of GroEL using the traditional K-means (a) and ACK-means (b) in RFA from SPIDER. Class size is shown at the left bottom of each class average. The quality of class averages from both algorithms is comparable, but ACK-means (b) substantially improved the balance of class sizes.

doi:10.1371/journal.pone.0167765.g005

same size, ACK-means monitors class sizes adaptively when determining the class assignment of particles.

In the tests with the simulated and experimental GroEL dataset, we found little improvement on the accuracy of classification by ACK-means against the traditional K-means in the RFA approach in SPIDER. However, in the tests with the experimental inflammasome and RP

Fig 6. 2D class averages of Inflammasome using the traditional K-means (a), EQK-means (b) and ACK-means (c) in MRA from SPARX. Class size is shown at the left bottom of each class average. The traditional K-means generated many blurred class averages and EQK-means produced some class averages with misaligned features.
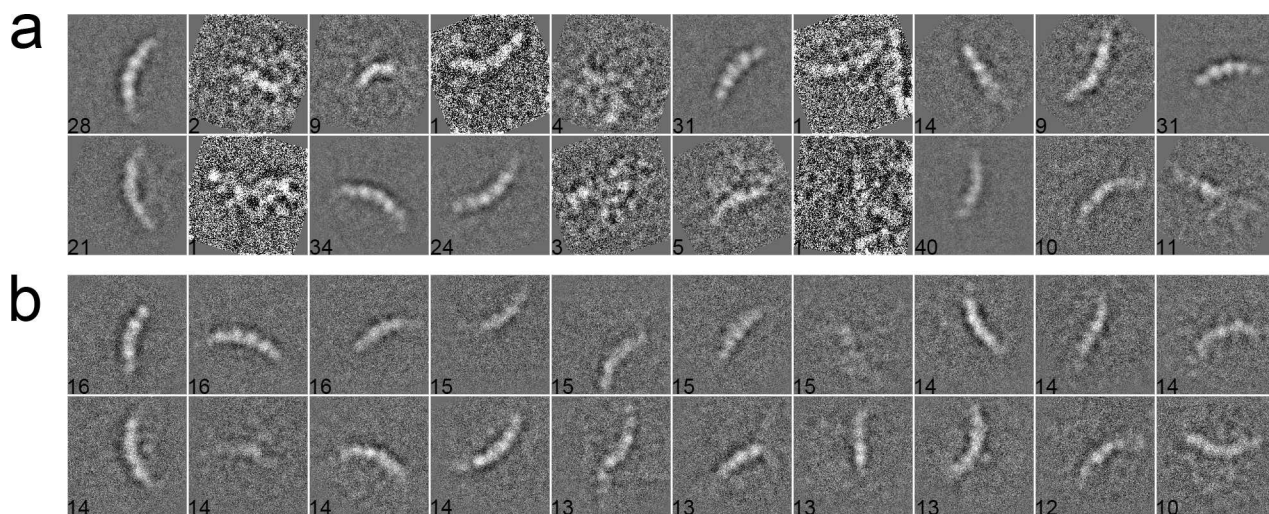
doi:10.1371/journal.pone.0167765.g006

Fig 7. 2D class averages of Inflammasome using the traditional K-means (a), and ACK-means (b) MRA/MSA from EMAN2. Class size is shown at the left bottom of each class average. ACK-means generated improved results as compared to the traditional K-means.

doi:10.1371/journal.pone.0167765.g007

datasets, we observed prominent differences between the two algorithms. Although the traditional K-means was modified by a factor in SPIDER, it still generated many classes with only one image. By contrast, ACK-means produced balanced classes with more informational class averages. Interestingly, we further combined ACK-means with our recently proposed statistical manifold learning algorithm [45] and found a significant improvement in the RFA approach (data not shown). It bodes well for the future development of improved data clustering protocols that integrate both ACK-means and manifold learning approaches. In summary, the ACK-means algorithm takes into account both the classification accuracy and the balance of class sizes. It presents a significantly improved alternative to the traditional K-means as a data-clustering algorithm for cryo-EM data analysis.

Fig 8. 2D class averages of inflammasome using the traditional K-means (a), and ACK-means (b) in RFA from SPIDER. Class size is shown at the left bottom of each class average. There are many classes in (a) with only one particle. Traditional K-means generated many classes with only one particle, which is avoided in the results from ACK-means.

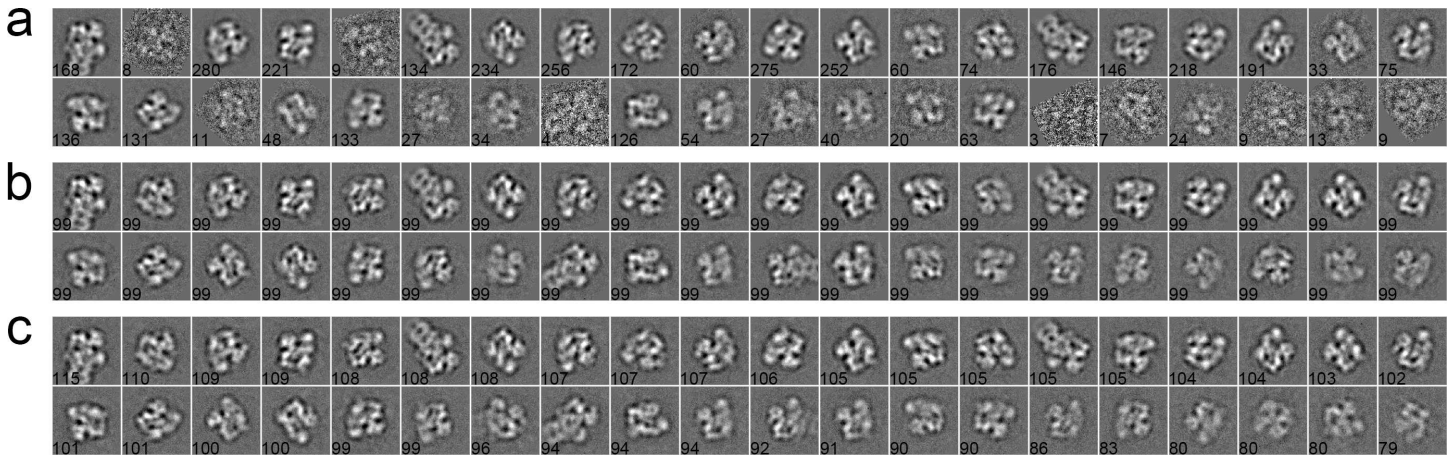doi:10.1371/journal.pone.0167765.g008

Fig 9. 2D class averages of RP using the traditional K-means (a) and ACK-means (b) in MRA from SPARX. Class size is shown at the left bottom of each class average. There are many blurred classes in (a) generated by the traditional K-means.
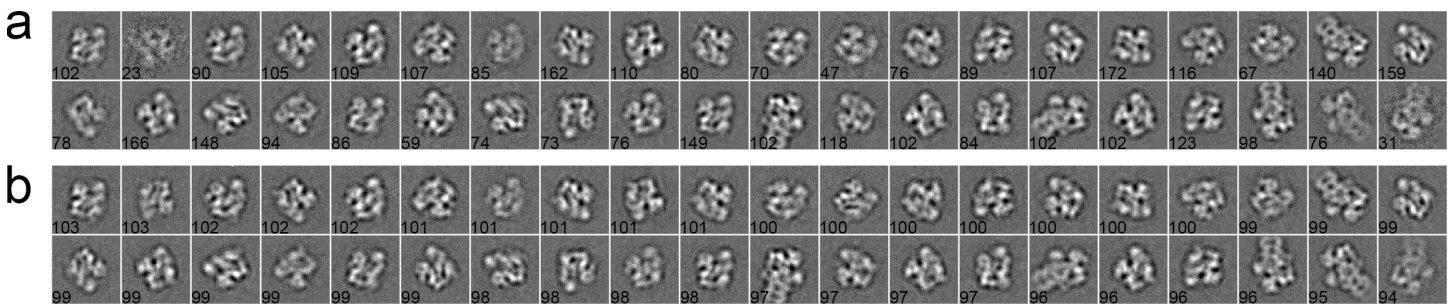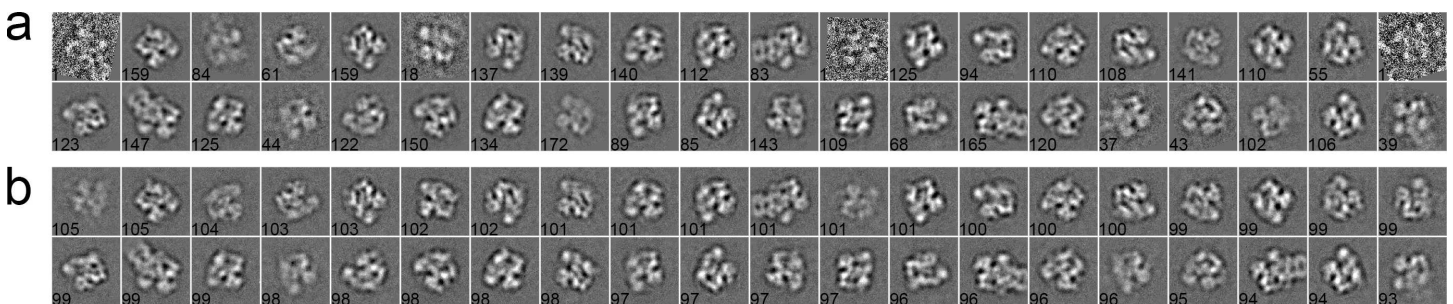
doi:10.1371/journal.pone.0167765.g009

Fig 10. 2D class averages of RP using the traditional K-means (a) and ACK-means (b) in MRA/MSA from EMAN2. Class size is shown at the left bottom of each class average. Classes generated by ACK-means (a) are clearer than those by the traditional K-means (a).

doi:10.1371/journal.pone.0167765.g010

Fig 11. 2D class averages of RP using the traditional K-means (a) and ACK-means (b) in RFA from SPIDER. Class size is shown at the left bottom of each class average. The traditional K-means (a) generated some poor classes with only one particle. The performance of ACK-means (b) is better than the traditional K-means.

doi:10.1371/journal.pone.0167765.g011

Supporting Information

S1 Fig. Simulated projections of *Escherichia coli* 70S ribosome. For each noise level, six orientations are shown. (a) SNR = 1/3. (b) SNR = 1/10. (c) SNR = 1/30.
(TIF)

S2 Fig. Comparison of classification results of ACK-means on simulated data with different β and at different noise levels. The first column (panels a, c and e) is the normalized histogram of angular distances. The second column (panels b, d and f) is the size of classes which is arranged in ascend order. The most balanced classification has a horizontal line in this plot. The experiments are conducted by the MRA approach in SPARX. (a) and (b) SNR = 1/3. (c) and (d) SNR = 1/10. (e) and (f) SNR = 1/30.
(TIF)

S3 Fig. Comparison of classification results of simulated data with SNR = 1/3. The first column (panels a, c and e) is the normalized histogram of angular distances. The second column (panels b, d and f) is size of classes which is arranged in ascend order. The most balanced classification has a horizontal line in this plot. (a) and (b) are for experiments using different clustering algorithms in the MRA approach under SPARX. (c) and (d) are for experiments using different clustering algorithms in the MRA/MSA approach under EMAN2. (e) and (f) are for experiments using different clustering algorithms in the RFA approach under SPIDER. In all graphs, red curves present the results from the ACK-means algorithm.
(TIF)

S4 Fig. Comparison of classification results of simulated data with SNR = 1/30. The first column (panels a, c and e) is the normalized histogram of angular distances. The second column (panels b, d and f) is size of classes which is arranged in ascend order. The most balanced classification has a horizontal line in this plot. (a) and (b) are for experiments using different clustering algorithms in the MRA approach under SPARX. (c) and (d) are for experiments using different clustering algorithms in the MRA/MSA approach under EMAN2. (e) and (f) are for experiments using different clustering algorithms in the RFA approach under SPIDER. In all graphs, red curves present the results from the ACK-means algorithm.
(TIF)

S5 Fig. Comparison of classification results of ACK-means in different approaches with RELION. The first column (panels a, c and e) is the normalized histogram of angular distances. The second column (panels b, d and f) is the size of classes which is arranged in ascend order. The most balanced classification has a horizontal line in this plot. (a) and (b) are from ACK-means in MRA and RELION. (c) and (d) are for ACK-means in MRA/MSA and RELION. (e) and (f) are for ACK-means in RFA and RELION. Many classes generated by RELION have no or few particles.
(TIF)

Acknowledgments

Author Contributions

Conceptualization: YX CCY YM.

Data curation: YX JW YM.

Formal analysis: YX CCY YM.

Funding acquisition: CCY YM.

Investigation: YX YM.

Methodology: YX YM.

Project administration: YX CCY YM.

Resources: YX JW CCY YM.

Software: YX JW CCY YM.

Supervision: CCY YM.

Validation: YX YM.

Visualization: YX.

Writing – original draft: YX YM.

Writing – review & editing: YX CCY YM.

References

1. Frank J. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state: Oxford University Press; 2006.

2. Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends in biochemical sciences. 2015; 40(1):49–57. doi: 10.1016/j.tibs.2014.10.005 PMID: 25544475

3. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, et al. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. Science. 2015; 348(6239):1147–51. doi: 10.1126/science.aab1576 PMID: 25953817

4. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, et al. Structure of the E. coli ribosome-EF-Tu complex at <3 A resolution by Cs-corrected cryo-EM. Nature. 2015; 520(7548):567–70. doi: 10.1038/nature14275 PMID: 25707802

5. Frank J, van Heel M. Correspondence analysis of aligned images of biological particles. Journal of molecular biology. 1982; 161(1):134–7. PMID: 7154073

6. van Heel M, Frank J. Use of multivariate statistics in analysing the images of biological macromolecules. Ultramicroscopy. 1981; 6(2):187–94. PMID: 7268930

7. Elmlund D, Elmlund H. SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles. J Struct Biol. 2012; 180(3):420–7. doi: 10.1016/j.jsb.2012.07.010 PMID: 22902564

8. Elmlund H, Elmlund D, Bengio S. PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. Structure. 2013; 21(8):1299–306. doi: 10.1016/j.str.2013.07.002 PMID: 23931142

9. Elmlund D, Davis R, Elmlund H. Ab initio structure determination from electron microscopic images of single molecules coexisting in different functional states. Structure. 2010; 18(7):777–86. doi: 10.1016/j.str.2010.06.001 PMID: 20637414

10. Crowther RA, DeRosier DJ, Klug A. The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 1970; 317(1530):319–40.

11. Van Heel M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy. 1987; 21(2):111–23. PMID: 12425301

12. Goncharov A, Gelfand M. Determination of mutual orientation of identical particles from their projections by the moments method. Ultramicroscopy. 1988; 25(4):317–27.

13. Penczek PA, Zhu J, Frank J. A common-lines based method for determining orientations for N> 3 particle projections simultaneously. Ultramicroscopy. 1996; 63(3):205–18.

14. Van Heel M, Orlova E, Harauz G, Stark H, Dube P, Zemlin F, et al. Angular reconstitution in three-dimensional electron microscopy: historical and theoretical aspects. Scanning Microscopy. 1997; 11:195–210.

15. Vargas J, Alvarez-Cabrera AL, Marabini R, Carazo JM, Sorzano CO. Efficient initial volume determination from electron microscopy images of single particles. Bioinformatics. 2014; 30(20):2891–8. doi: 10.1093/bioinformatics/btu404 PMID: 24974203

16. Jaitly N, Brubaker MA, Rubinstein JL, Lilien RH. A Bayesian method for 3D macromolecular structure inference using class average images from single particle electron microscopy. Bioinformatics. 2010; 26(19):2406–15. doi: 10.1093/bioinformatics/btq456 PMID: 20702397

17. Elad N, Clare DK, Saibil HR, Orlova EV. Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J Struct Biol. 2008; 162(1):108–20. doi: 10.1016/j.jsb.2007.11.007 PMID: 18166488

18. Zhang W, Kimmel M, Spahn CM, Penczek PA. Heterogeneity of large macromolecular complexes revealed by 3D cryo-EM variance analysis. Structure. 2008; 16(12):1770–6. doi: 10.1016/j.str.2008.10.011 PMID: 19081053

19. van Heel M, Marina S-M. Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. The EMBO Journal. 1985; 4(9):2389–95. PMID: 3908096

20. Penczek P, Radermacher M, Frank J. Three-dimensional reconstruction of single particles embedded in ice. Ultramicroscopy. 1992; 40(1):33–53. PMID: 1580010

21. Hohn M, Tang G, Goodyear G, Baldwin P, Huang Z, Penczek PA, et al. SPARX, a new environment for Cryo-EM image processing. Journal of structural biology. 2007; 157(1):47–55. doi: 10.1016/j.jsb.2006.07.003 PMID: 16931051

22. Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, et al. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. Journal of structural biology. 1996; 116(1):190–9. doi: 10.1006/jsbi.1996.0030 PMID: 8742743

23. Van Heel M, Frank J. Classification of particles in noisy electron micrographs using correspondence analysis. Pattern Recognition in Practice I Gelsema ES, Kanal L (eds) North-Holland Publishing, Amsterdam. 1980:235–43.

24. Van Heel M. Multivariate statistical classification of noisy images (randomly oriented biological macro-molecules). Ultramicroscopy. 1984; 13(1–2):165–83. PMID: 6382731

25. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, et al. EMAN2: an extensible image processing suite for electron microscopy. Journal of structural biology. 2007; 157(1):38–46. doi: 10.1016/j.jsb.2006.05.009 PMID: 16859925

26. Schatz M, Van Heel M. Invariant classification of molecular views in electron micrographs. Ultramicroscopy. 1990; 32(3):255–64. PMID: 2382314

27. Zhao Z, Singer A. Rotationally invariant image representation for viewing direction classification in cryo-EM. J Struct Biol. 2014; 186(1):153–66. PubMed Central PMCID: PMC4014198. doi: 10.1016/j.jsb.2014.03.003 PMID: 24631969

28. Bhamre T, Zhang T, Singer A. Denoising and covariance estimation of single particle cryo-EM images. J Struct Biol. 2016; 195(1):72–81. doi: 10.1016/j.jsb.2016.04.013 PMID: 27129418

29. Penczek PA, Kimmel M, Spahn CM. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. Structure. 2011; 19(11):1582–90. PubMed Central PMCID: PMC3255080. doi: 10.1016/j.str.2011.10.003 PMID: 22078558

30. Sorzano C, Bilbao-Castro J, Shkolnisky Y, Alcorlo M, Melero R, Caffarena-Fernández G, et al. A clustering approach to multireference alignment of single-particle projections in electron microscopy. Journal of structural biology. 2010; 171(2):197–206. doi: 10.1016/j.jsb.2010.03.011 PMID: 20362059

31. Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. Journal of structural biology. 2012; 180(3):519–30. doi: 10.1016/j.jsb.2012.09.006 PMID: 23000701

32. Scheres SH, Valle M, Nunez R, Sorzano CO, Marabini R, Herman GT, et al. Maximum-likelihood multi-reference refinement for electron microscopy images. Journal of molecular biology. 2005; 348(1):139–49. doi: 10.1016/j.jmb.2005.02.031 PMID: 15808859

33. de la Rosa-Trevin JM, Oton J, Marabini R, Zaldivar A, Vargas J, Carazo JM, et al. Xmipp 3.0: an improved software suite for image processing in electron microscopy. J Struct Biol. 2013; 184(2):321–8. doi: 10.1016/j.jsb.2013.09.015 PMID: 24075951

34. Yang Z, Fang J, Chittuluru J, Asturias FJ, Penczek PA. Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure. 2012; 20(2):237–47. doi: 10.1016/j.str.2011.12.007 PMID: 22325773

35. Banerjee A, Ghosh J. Scalable Clustering Algorithms with Balancing Constraints. Data Mining and Knowledge Discovery. 2006; 13(3):365–95.

36. Zhu S, Wang D, Li T. Data clustering with size constraints. Knowledge-Based Systems. 2010; 23 (8):883–9.

37. Malinen MI, Fränti P. Balanced K-Means for Clustering. In: Fränti P, Brown G, Loog M, Escolano F, Pelillo M, editors. Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20–22, 2014 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 32–41.

38. Frank J. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. 2nd ed. Oxford; New York: Oxford University Press; 2006. xiv, 410 p. p.

39. Aloise D, Deshpande A, Hansen P, Popat P. NP-hardness of Euclidean sum-of-squares clustering. Machine Learning. 2009; 75(2):245–8.

40. MacQueen J, editor Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967: Oakland, CA, USA.

41. Shaikh TR, Gao H, Baxter WT, Asturias FJ, Boisset N, Leith A, et al. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nature protocols. 2008; 3(12):1941–74. doi: 10.1038/nprot.2008.156 PMID: 19180078

42. Penczek PA. Image Restoration in Cryo-Electron Microscopy. 2010. p. 35–72.

43. Ludtke SJ, Chen DH, Song JL, Chuang DT, Chiu W. Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy. Structure. 2004; 12(7):1129–36. doi: 10.1016/j.str.2004.05.006 PMID: 15242589

44. Zhang L, Chen S, Ruan J, Wu J, Tong AB, Yin Q, et al. Cryo-EM structure of the activated NAIP2-NLRC4 inflammasome reveals nucleated polymerization. Science. 2015; 350(6259):404–9. PubMed Central PMCID: PMC4640189. doi: 10.1126/science.aac5789 PMID: 26449474

45. Wu J, Ma YB, Condgon C, Brett B, Chen S, Ouyang Q, et al. Unsupervised single-particle deep classification via statistical manifold learning. arXiv. 2016:1604.04539 [physics.data-an].