Title

A simulated annealing approach for resolution guided homogeneous cryo-electron microscopy image selection

Journal

Quantitative Biology, 8(1)

ISSN

2095-4689

Authors

Shi, Jie
Zeng, Xiangrui
Jiang, Rui
et al.

Publication Date

2020-03-01

DOI

10.1007/s40484-019-0191-8

Peer reviewed

Published in final edited form as:

Quant Biol. 2020 March ; 8(1): 51–63. doi:10.1007/s40484-019-0191-8.

A simulated annealing approach for resolution guided homogeneous cryo-electron microscopy image selection

Jie Shi[1], **Xiangrui Zeng**[2], **Rui Jiang**[3], **Tao Jiang**[4], **Min Xu**[2,*]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China

[2]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[3]Department of Automation, Tsinghua University, Beijing 100084, China

[4]Department of Computer Science and Engineering, University of California-Riverside, Riverside, CA 92521, USA

Abstract

Background: Cryo-electron microscopy (Cryo-EM) and tomography (Cryo-ET) have emerged as important imaging techniques for studying structures of macromolecular complexes. In 3D reconstruction of large macromolecular complexes, many 2D projection images of macromolecular complex particles are usually acquired with low signal-to-noise ratio. Therefore, it is meaningful to select multiple images containing the same structure with identical orientation. The selected images are averaged to produce a higher-quality representation of the underlying structure with improved resolution. Existing approaches of selecting such images have limited accuracy and speed.

Methods: We propose a simulated annealing-based algorithm (SA) to pick the homogeneous image set with best average. Its performance is compared with two baseline methods based on both 2D and 3D datasets. When tested on simulated and experimental 3D Cryo-ET images of Ribosome complex, SA sometimes stopped at a local optimal solution. Restarting is applied to settle this difficulty and significantly improved the performance of SA on 3D datasets.

Results: Experimented on simulated and experimental 2D Cryo-EM images of Ribosome complex datasets respectively with SNR = 10 and SNR = 0.5, our method achieved better accuracy in terms of F-measure, resolution score, and time cost than two baseline methods. Additionally, SA shows its superiority when the proportion of homogeneous images decreases.

Conclusions: SA is introduced for homogeneous image selection to realize higher accuracy with faster processing speed. Experiments on both simulated and real 2D Cryo-EM and 3D Cryo-ET images demonstrated that SA achieved expressively better performance. This approach serves as an important step for improving the resolution of structural recovery of macromolecular complexes captured by Cryo-EM and Cryo-ET.

Author summary:

*Correspondence: mxu1@cs.cmu.edu.

The authors Jie Shi, Xiangrui Zeng, Rui Jiang, Tao Jiang and Min Xu declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

It is significant for improving quality of 3D reconstruction of large macromolecular complexes to select homogeneous images from all obtained projection images obtained by Cryo-electron microscopy and tomography and then gain higher resolution by image averaging. We aim to find a better working method for this job. A simulated annealing-based algorithm (SA) performs better than baseline methods on both accuracy and speed after testing on simulated and experimental datasets. In order to further enhance its performance on 3D images, restarting is introduced to prevent the algorithm from stopping at a local optimal solution.

Keywords

simulated annealing; image averaging; cryo-electron microscopy; cryo-electron tomography

INTRODUCTION

Cellular processes are arguably governed by macromolecular complexes. Knowing the structures of macromolecular complexes is important for understanding their function. Due to data acquisition limits, most of macromolecular structures are unknown. The recent revolutions in cryo electron microscopy (Cryo-EM) [1] and tomography (Cryo-ET) [2] have made powerful tools for the structural analysis of macromolecular complexes. Cryo-EM uses transmission electron microscope (TEM) to capture 2D projected images of numerous copies of a specific purified and frozen macromolecular complex. The projected images are then used for reconstructing the 3D structure of the macromolecular complex. By contrast, Cryo-ET often utilizes TEM to capture frozen and intact cells. The cell samples are rotated to capture projected images from different tilt angles. Then a tomogram is reconstructed from the projected images. The tomogram is a 3D gray scale image that represents a 3D electron density map of the cell sample at submolecular resolution and in close to native condition, which contains macromolecules of heterogeneous structures.

The macromolecule images captured by both Cryo-EM and Cryo-ET techniques often have very low signal-to-noise ratio (SNR), which is the main limiting factor to the resolution of recovered structures. When processing the single particle by Cryo-EM, in order to limit radiation damage, it should record the projection images of the particle at low election dose. However, it also causes the projections with much noise and low contrast. Typically, its SNR is less than 1 [3]. To improve resolution through increasing SNR, a key step is to average large numbers of images containing identical structure with identical orientation. Currently, given large number of heterogeneous macromolecule images available, how to efficiently and accurately select such images (see Fig. 1A) is an open research topic, especially for Cryo-ET data.

Given a set of images, we aim to find a subset S of images that produces the best average in terms of its resolution score. The resolution score is defined by fourier ring correlation (FRC) for 2D Cryo-EM images and fourier shell correlation (FSC) for 3D Cryo-ET images [4], derived from spatial SNR (SSNR). A simulated annealing (SA) [5] approach is proposed for finding such an optimal image subset S, with constraint on its size range. We compare our method with two baseline methods, namely image matching algorithm (MA) and genetic

algorithm (GA) [6]. Our experiments on both Cryo-EM and Cryo-ET images of known macromolecular complex structures demonstrate that our SA achieves significantly better performance in terms of image average resolution, selection accuracy, and convergence rate.

RESULTS

Construction of simulated images

In our experiments, 2D Cryo-EM and 3D Cryo-ET images of the ribosome complex are simulated in a similar way as in [7], and we add certain amount of noise controlled by the SNR parameter. The simulated images constitute three 2D image datasets with different SNR (see Fig. 1B) and a 3D dataset. Each 2D dataset consists of 100 homogeneous images and 100 heterogeneous images generated by randomly rotating the macromolecule. In the simulation 2D dataset, we set the standard deviation of rotational angle interval for heterogeneous images to 5° and the standard deviation of translation to 2 pixels (voxels for 3D dataset). For the generated homogeneous images, the average rotational alignment error is 0.003, which is calculated in the same way as in [7]. In each dataset, these methods aim to select the set containing all homogeneous images.

Performance measures

Performance of these algorithms is measured by calculating final resolution score (see Section of “Resolution based objective functions”), time costs, precision, and recall. *F*-measure is applied to summarize precision and recall.

$$F_\beta = \left(1+\beta^2\right)\cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}\ \cdot\beta^2+\text{Recll}}\ . \tag{1}$$

Since precision is expected to have a larger influence on the score, we utilize a particular form of *F*-measure with bias, F_β and set $\beta = 0.5$ to weight precision more [8].

Resolution based objective functions

Cryo-EM 2D images—Cryo-EM captures a collection of 2D images, each is the projected image of a particular macro-molecule. Given a subset S of aligned 2D images represented in Fourier space, we let a_{ij} be the Fourier coeffcient of image i at location j. Then the average coefficient of a subset S of images at j is $\overline{a_j} = \frac{1}{|S|}\sum_{i\in S} a_{ij}$. Let J be the set of locations that correspond to a particular frequency in Fourier space, and let $\mathcal{X}$ be the collection of such sets. Then the corresponding SSNR is calculated as [4]:

$$s(J) = \frac{\sum_{j\in J}\left|\sum_{i\in S} a_{ij}\right|^2}{\frac{|S|}{|S|-1}\sum_{j\in J}\sum_{i\in S}\left|a_{ij}-\overline{a_j}\right|^2}\ . \tag{2}$$

Similar to [6], such SSNR can be directly calculated from several quantities that have additive property in the following form $f(S_1 \cup S_2) = f(S_1) + f(S_2), \forall S_1 \cap S_2 = \Phi$ for all disjoint image subsets of S. Such property makes it computationally efficient for updating

the corresponding SSNR after adding or removing a single image from an image set, which significantly increases the speed of SA. The corresponding FRC is then calculated as $c(\mathrm{J}) = \frac{s(J)}{2 + s(J)}$ [4]. Then the resolution score is calculated as $r = \sum_{J \in \chi} c(J)$, which is the objective function to be maximized.

Cryo-ET 3D images—Cryo-ET captures a collection of 3D images (a.k.a. subtomograms), each a 3D electron density map of a particular macro-molecule. Given a subset S of aligned 3D images represented in Fourier space, we let a_{ij} be a Fourier coeffcient of image i at location j, and m_{ij} be an indicator function that equals to 1 when a_{ij} is observable, and 0 when a_{ij} is missing. Then the SSNR that take into account of missing values is calculated as [6]:

$$s(J) = \frac{\sum_{j \in J} m_j |\overline{a_j}|^2}{\sum_{j \in J} \frac{1}{m_j - 1} \sum_{i \in S} m_{ij} |a_{ij} - \overline{a_j}|^2}, \tag{3}$$

where $m_j = \Sigma_{i \in S}\, m_{ij}$ and $\overline{a_j} = \frac{1}{m_j} \sum_{i \in S} m_{ij} a_{ij}$. Similar to the 2D case, such SSNR can also be directly calculated from several quantities that have additive property [6]. Given the SSNR, the corresponding FSC and resolution score is calculated in the similar way as in Section of "Cryo-EM 2D images".

Difference of processing 2D and 3D images—The major difference between 2D and 3D case is that 3D images contain missing values in their Fourier representation. For 3D images, the projection data are usually incomplete since not all 3D information of the original structure can be gathered. It is mainly limited by current technology of acquiring the 3D images. For instance, some objects may be not inside the radiation field [9].

In 3D case, particles are cut out as small 3D images (subtomograms) from a large 3D tomogram image. Our target is to find a subset of subtomograms whose average achieves max resolution that would be very useful for mining meaningful structural patterns inside these subtomograms. A subtomogram is a small cubic region of a 3D tomogram image that potentially contains one macromolecular complex. It is a small 3D image. It is also convenient to represent such image in Fourier space because in such case its missing value can be simply represented.

Experiments

Ratio adjustment—For studying the applicability of algorithms in more realistic conditions, we set different proportion of homogeneous images in the data set. Ratios in {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} are tested respectively, with

$$ratio := \frac{|homogeneous\ images|}{|heteroeneous\ images|}. \tag{4}$$

After 10 iterations, SA can gain higher resolution score than other two methods, and the final resolution score it achieves is higher with a larger proportion of homogeneous images in the dataset.

From the results (see Time (s) in Table 1), we can also find that following the increase of iteration number, GA has a much more time cost than SA and MA, but its performance is not better than SA.

Moreover, MA achieves much worse resolution score than the other two methods (see Table 1) especially when the proportion of homogeneous images is very small. It comes under observation that performance of MA in this problem is too dependent on the initial image selection. Since the initial images are chosen randomly (see Section of “Image matching”), for MA, if a good image pair (homogeneous) is chosen at the beginning, it may have a relatively good performance eventually. Once it starts from a heterogeneous image pair, it would cause a disastrous consequence because the algorithm cannot recover this error.

In addition, the results of experiments demonstrate that “resolution” performs not very well as the measurement in this project. The pair of images in S_0 achieved highest resolution score is not always homogeneous image pair, so sometimes resolution increases but precision decreases. In MA, it unremittingly seeks the image which can be averaged to realize the highest resolution. In this extreme case, it is easier for the algorithm to be misguided by “resolution”. For SA, even if it starts from a bad initial situation, it can remove the images that are added wrongly in the following procession (see Section of “Contraction mechanism of SA”).

By testing, when the proportion of homogeneous images is larger in the whole dataset, performances of three algorithms all have an improvement to some extent. SA also performs better with a larger ratio. However, in reality, the proportion of homogeneous images is usually small, that is also why selection of homogeneous images is a very hard work but significant for better 3D refinement.

Tests on 2D experimental data—For demonstrating the potential of this method, these three methods are evaluated respectively not only on 2D simulated data but experimental data obtained from [10]. The image set is manually divided into two separate parts, 100 homogeneous images and 100 heterogeneous images (see Fig. 2A). The dataset for testing consists of images chosen randomly from homogeneous and heterogeneous image sets by a certain proportion. As Table 2 shows, SA algorithm can realize higher resolution and higher $F_{0.5}$ than two baseline methods with lower time complexity.

The experimental results on the actual cryo-EM image set also prove that SA performs better than two baseline methods. It has relatively stable performance for classifying homogeneous and heterogeneous images no matter what ratio is and spends much less time than GA. Furthermore, it can realize a higher resolution score finally.

In order to further prove the validity of this method, we choose another image set with 100 homogeneous images and 100 heterogeneous images which are projection images of the particle from another view direction (see Fig. 2B). For testing different ratios, heterogeneous images are also selected at random with different number. Time costs of three algorithms are similar on the datasets with projection images from different view directions.

Based on the above results, performance of SA on picking homogeneous projection images of macromolecular complex particle is obviously better than GA and MA. It can gain a higher resolution score after 10 iterations using much less time.

Tests and improvement on 3D data—SA and GA are also tested on 3D cryo-ET simulated data with SNR = 0.01 and 3D cryo-ET experimental data respectively. The 3D experimental dataset includes 859 3D projected images of ribosome where 200 images contain identical structures with identical orientation and thus homogeneous ($ratio \approx 0.3$) [11].

SA algorithm usually converges after 10 iteration on 3D data, but partly due to the missing value of 3D images, it is more difficult for the algorithm to pick homogeneous images in 3D dataset than in 2D dataset. The preliminary results (see Fig. 3B) display that SA algorithm often results in a much small solution set S. Though the convergence rate of SA is much faster than GA, SA still stops at a local optimal solution untimely rather than continues to pursue a better solution.

From experiments, it is harder for SA algorithm to select a good initial condition on 3D data than 2D data. In the 3D data, the image pair which is chosen initially with high resolution score is not a homogeneous pair with high probability. The defect of resolution as the measurement is more obvious on 3D data due to the lower SNR and the missing value. In order to overcome it, SA is adapted for 3D data in the process of restarting, named “SA+”. After a certain number of iterations, SA+ will select the two non-selected images as the new initial condition and start selecting again on all non-selected images. A new solution set S' will be generated. We choose the part of S' which has no intersection with previous solution sets S, and combine it with S, $S \leftarrow S \cup S'$. Then the combination will continue to be input in the contraction mechanism of SA.

The results shown in Tables 4 and 5 demonstrate the superior performance of SA+ as compared with GA and SA. Distinctly, SA and SA+ use far less time than GA on this work but gain higher $F_{0.5}$ and resolution. From Fig. 4, SA+ has slightly higher time complexity than SA. However, SA+ will keep searching for a better solution rather than stop at a locally optimal solution after certain iterations like SA.

Comparison of convergence rate—Some reconstructions of asymmetric particles such as ribosomes use 10^6 images of particles, so speed is crucial in this project. Time consumed in each iteration is recorded to compare the speed and convergence of SA and GA. From records, time complexity of GA is obviously higher than SA (Tables 2, 3 and 5).

Analysis of different performances of SA and GA—As the experimental results show, SA performs better than GA aimed at this problem, whether for 2D or 3D data. Both of these two algorithms guide the optimization procedure based on the idea of evolution. GA concentrates on the evolution of next population and realizes optimization by selection. Different with it, SA is bent on individual optimization.

GA is an adaptive algorithm for globally searching the optimal solution, which has a good performance of global searching [12]. Nevertheless, similar to natural world, GA knows nothing about the problem to solve and has no requirements for the searching space. Therefore, it is more appropriate for GA to solve those problems without an obvious direction of optimization. In other words, in some scenarios, we cannot define what kind of solution is better. For instance, when allocating a fund to multiple stocks and determining the proportion of investment in each stock, no clear indicator can tell what proportion of them must increase return of the whole portfolio. GA can generate more portfolios by crossover and mutation. It has a greater chance to find the optimal solution. Whereas, in this project, we have a specific target to find out homogeneous images and then improve resolution of their average. It is foreseeable that adding heterogeneous images would reduce the resolution of the averaged image. Nonetheless, when processing original image set by GA, it may generate many obviously wrong combinations unavoidably as a result of random mutation and crossover. It cannot guarantee the quality of new generated solutions. It means time to select the optimal solution from such many solutions. SA prevents this kind of work, which can search for a new solution with clear direction and achieve local optimum quickly. Owning to accepting a worse solution in a probability, it can also avoid to be trapped in the locally optimal solutionn. Upon the results of experiments, it can be seen that this approach dose save much time and achieve satisfactory resolution at the same time. GA is also believed to gain a satisfactory solution eventually, but it needs much more time especially when the amount of data is very large.

In this paper, we aim at finding out a homogeneous image set from all images obtained by Cryo-EM or Cryo-ET. Through it cannot be promised to gain the global optimum, SA's solution is with realistic significance. It is helpful for more accurate reconstruction of real particle structure to enhance resolution of projection images in each direction [13].

DISCUSSION

Cryo-EM dates back to 1968, which is a significant approach for 3D reconstruction of biological macromolecular complexes [1]. Aiming at single-particle Cryo-EM [14], some software has been developed, such as EMAN2 [15], Relion [16]. Such software applies an integrated system to implement the entire process of Cryo-EM single particle reconstruction including image data acquisition, CTF estimation correction, particle selection, image matching, image classification, model reconstruction optimization, density image optimization and final model output.

However, due to the bottleneck of this technique, it can usually produce a projection image set with many noisy images which are acquired from different view directions. For each direction, more than one projected images are obtained with different resolution and different orientations. If all these images are utilized for construction, accurate structure of the particle cannot be obtained. Thus, more focus has been put towards image processing algorithm for high resolution. Filtering the projection dataset for picking homogeneous images is an extremely meaningful step before 3D refinement [17].

In traditional cryo-EM, unsupervised classification is widely applied to find out the homogeneous projection image sets, typically clustering homogeneous images by *K*-Means [18]. Nonetheless, This unsupervised method has some insurmountable difficulties. At first, it is hard to evaluate the homogeneity of each cluster [18]. Most of relevant work evaluates performance of the algorithm by observing the output image. Beyond that, its performance relies much on the chosen of cluster number *K*. A few supervised methods of classification are proposed like MPA classification [19] which is based on density maps and correlation-based projection-matching. This methods focus on molecular dynamics of protein complexes, which is extremely different with the target of our paper. Moreover, it is only aimed at cryo-EM images.

Our algorithm hopes to take out heterogeneous images from the whole image set so that high-resolution projected images are left, which is a supervised algorithm for a better refinement of 3D macromolecule complex particle. This method can assess the homogeneity of selected image subset well and truly by *F*-measure. Goals of our experiments is to test performance of conventional optimization algorithms on selection of homogeneous projection images of macromolecule complex particles. It is proved that SA can achieve a satisfactory resolution eventually with less time costs. In the meanwhile, by improving SA, it can also perform well on cryo-ET datasets to implement the selection of homogeneous 3D images.

Cryo-ET is another approach of 3D reconstruction using electron microscopy [1]. It is applied to study the 3D structure of proteins, viruses, organelles, and complexes with structural heterogeneity. It retrieves 3D structure of object of study through acquiring projected images from multiple perspectives. Although now resolution of structures gained by Cryo-ET cannot be compared with that of Cryo-EM single particle analysis, it plays an irreplaceable role on studying the structure and function of non-fixed form, asymmetric, heterogeneous biological samples [20,21]. Similarly, in Cryo-ET, registration of projected images is a key step, which is the foundation of high-quality reconstruction. Usually, a lot of projections are generated from a perspective after image registration, but some of them with low resolution may disturb following work. For evaluating the performance of our algorithm in widespread use, it is also test to reduce low-resolution images in cryo-ET projection image sets.

CONCLUSION

Selection of homogeneous images among large number of heterogeneous images is a very important step for improving the resolution of structural recovery of macro-molecular complexes captured by cryo-EM and cryo-ET. Our experiments on both simulated and real 2D cryo-EM and 3D cryo-ET image sets demonstrated that our SA approach achieved significantly better image selection performance compared to baseline methods in terms of resolution, accuracy, and convergence rate. Further improvements include designing better SA strategy, and resolution measures that are more robust to noise and missing values.

MATERIALS AND METHODS

Homogeneity of images

Image averaging is an important component in the 3D structural reconstruction, which involves the grouping of images so that each class represents a particular orientation of the particle of same structure.

In order to gain the average of images with higher resolution, image alignment [22] is essential since if the image is not aligned, *i.e.* the objects inside have different orientation and translations, even if they have same projection of same structure, the average image still does not reflect the true underlying structure.

In this paper, we define homogeneous images as the images with same orientation and in alignment. In other words, they can be used to generate the best average. On the contrary, heterogeneous image means that adding it into the image set will decrease the resolution of the average image.

Simulated annealing

Simulated annealing (SA) is a stochastic optimization algorithm inspired from the annealing theory of solid for settling combinatorial optimization problems [5]. Owning to existing random factors in its search process, it is not incline to converge in a local optima in the search space, compared to Hill Climbing algorithm [23].

Given a set of images, we first find a subset S of two images whose average achieves the highest resolution score. We then keep trying to add into S or replace in S an image with certain probability G. The calculation of G takes into account (1) the resolution score of the modified image set S', (2) current maximum resolution score, and (3) current number of iteration. The detailed procedure can be found in Algorithm 1, where the probability G is calculated as $G(r(S),r(S'),k) := exp(|r(S)-r(S')|/(k+1))$. G is defined according to the Metropolis Hastings rule. We use *"k* + 1" instead of "*k*" because k is always set to 0 at the beginning of the algorithm.

Algorithm 1

Simulated annealing

Require: A image set, S_0; Maximum size of image set n_{max}; Maximum number of iterations k_{max}; Minimum resolution score r_{min}; Iteration number M

Ensure: Optimal image subset S;

1: Calculate $(i^*, j^*) \leftarrow \underset{(i,j) \in S_0 \times S_0, i \neq j}{\arg\max} r(\{i, j\})$

2: $S \leftarrow \{i^*, j^*\}$

3: $n_{max} \leftarrow |S_0|/2$

4: **While** $(m < M)$ **do**

5: **While** $(k < k_{max})$ **do**

6: Randomly choose $t \in S_0 \backslash S$

7: **if** $|S| < n_{max}$ **then**

$S' \leftarrow S \cup \{t\}$
Randomly choose $\alpha \in (0,1)$
if $r(S') > r(S)$ or $G(r(S), r(S'), k) > \alpha$ **then**
$r(S) \leftarrow r(S')$
$S \leftarrow S'$
else
end if
end if
$k \leftarrow k + 1$
end while
for $i \in \{1, \ldots, |S|\}$ **do**
Randomly choose $t \in S$
$S^{(i)} \leftarrow S - \{t\}$
$t^{(i)} \leftarrow t$;
end for
Calculate $i_1, i_2, \ldots, i_{|S|}$ such that $S^{(i_1)} > S^{(i_2)} > \ldots > S^{(i_{|S|})}$
$S \leftarrow S - \left\{ t^{(i_1)}, t^{(i_2)}, \ldots t^{\left(i_{n_{\min}} - |S|\right)} \right\}$
$m \leftarrow m + 1$
end while
return S

Initial condition selection—As designed, SA algorithm selects two images (initial condition) in S at the beginning, and puts more other eligible images into S based on comparing the resolution of the average of these two images and others. Upon testing, SA algorithm is sensitive to the initial condition. It can perform very well with a good initial condition. Thus, it is very crucial to select a good initial condition. In essence, two homogeneous images are expected to select from So as far as possible. Originally, we select a pair of images whose average has highest resolution score among all pairs. However, the result of experiments shows it might not always be the case (see Fig. 5). Sometimes a pair of heterogeneous images are averaged to get a higher resolution than a pair of homogeneous images especially when a much low proportion of homogeneous images in S_0. Selecting the pair with highest resolution also costs a lot of time. Meanwhile, although two homogeneous images cannot always gain highest resolution score, we find the arithmetic mean of resolutions which are achieved by every pair of homogeneous images in S_0, $M_{\text{homogeneous}}$, is often larger than a threshold. It is larger than the arithmetic mean of resolutions achieved by heterogeneous pairs or mixed pairs (a homogeneous image and a heterogeneous image). For selecting homogeneous pairs as far as possible, we select two images whose average achieves the resolution larger than a threshold near $M_{\text{homogeneous}}$.

Contraction mechanism of SA—In initial experiments, we find that sometimes $F_{0.5}$ decreases with the increase of iteration, which indicates false positive images are selected into S. In order to address this issue, contraction mechanism is introduced in SA algorithm.

After each iteration, the algorithm checks every images i in S, and reduces i if S-{i} can obtain higher resolution. In multiple iterations, previously eliminated images will be evaluated again and added to S if higher resolution resulted.

Then Baseline methods and SA algorithm with contraction mechanism are tested on 2D datasets with SNR= 0.5, 10. SA significantly outperformed the two baseline methods (see Fig. 6), and it has a better performance on the dataset with larger SNR.

Baseline methods

Image matching—Image matching is a typical method of image denoising. For electron microscope images, many researches are centered on image matching for achieving high-quality images such as feature-based matching [24,25], template-based matching [26] and image block matching [27]. Different methods measure the similarity between images by different indicators, such as the geodesic distance [27], statistical features [24], crosscorrelation features [28]. In essence, the image matching algorithm is to classify the whole image set upon “similarity” and find the subsets of more similar images.

In this paper, for ease of comparison between SA and GA, we also define the similarity of images based on resolution. That is, if the average of two images can realize a higher resolution, they are thought to more similar.

Given a set S_0 of images, we first select a pair of images, a and b, whose average has highest resolution score among all pairs. The selected image pair is used as the initial image subset S. This tries to find two homogeneous images, and guarantees a good start of the algorithm. We then regard a and b as the target image, respectively. The algorithm constantly selects an image i in $S_0\backslash\{a,b\}$ such that the average of the target image and i achieves highest resolution. Then we update $S \leftarrow S \cup \{i\}$, and set i as new target image. We repeat the addition procedure until the image selected has been in S. And S is the solution subset.

Genetic algorithm—Genetic algorithm (GA) is another classical optimization algorithm, which has remarkable global searching ability [29]. It is also adept at finding the maximum or best solution of an objective function. Therefore, in this problem, we consider to capitalize on GA to search for best solution of the resolution-based objective function.

GA of image selection regards an image subset S as a candidate solution which consists of certain images. Each image in the dataset is labelled as a number. Each candidate solution is represented by a sequence of “0” and “1” where “1” indicates the image with corresponding number exists in this candidate solution set, and “0” means the opposite.

The fitness of a candidate subset is assessed by the resolution of its image average. Fit candidate subsets are selected for future generation to produce updated candidate subsets. In each generation, a fixed proportion of candidate subsets undergo random cross-over and mutation. The newly generated candidate subsets are subject to fitness assessment. The generation process is stopped when the fittest candidate subset remains unchanged over successive generations or the maximum number of iterations is reached. This algorithm works like a process of natural selection, where “survival of the fittest in natural selection,

survival of the fittest". Following this process, the candidate solution set with best average is expected in the final generation. GA is implemented in a similar way as in [6].

ACKNOWLEDGEMENTS

We thank Dr. Ming Sun for suggestions and Mr. Shan Zhou for initial exploratory studies. We thank Ms. Xindi Wu for helping with manuscript editing. This work was supported in part by U.S. National Institutes of Health (NIH) grant (P41 GM103712). MX acknowledges support from Samuel and Emma Winters Foundation. XZ was supported by a fellowship from Carnegie Mellon University's Center for Machine Learning and Health. RJ is a RONG professor at the Institute for Data Science, Tsinghua University.

REFERENCES

1. Frank J (2006) Three-dimensional Electron Microscopy of Macromolecular Assemblies. New York: Oxford University Press
2. Lučić V, Rigort A and Baumeister W (2013) Cryo-electron tomography: the challenge of doing structural biology *in situ*. J. Cell Biol, 202, 407–419 [PubMed: 23918936]
3. Sali A, Glaeser R, Earnest T and Baumeister W (2003) From words to literature in structural proteomics. Nature, 422, 216–225 [PubMed: 12634795]
4. Liao HY and Frank J (2010) Definition and estimation of resolution in single-particle reconstructions. Structure, 18, 768–775 [PubMed: 20637413]
5. Van Laarhoven PJM and Aarts EHL (1987) Simulated annealing In: Simulated Annealing: Theory and Applications, pp. 7–15. New York: Springer
6. Xu M, Tocheva EI, Chang Y-W, Jensen GJ and Alber F (2015) *De novo* visual proteomics in single cells through pattern mining. Structure, 27, 679–691.e14
7. Xu M, Beck M and Alber F (2012) High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. J. Struct. Biol, 178, 152–164 [PubMed: 22420977]
8. Powers DM, (2011) Evaluation: from precision, recall and F-measure to roc, informedness, markedness and correlation. J. Mach. Learn. Tech 2, 37–63
9. Snyder DL, O'Sullivan JA, Murphy RJ, Politte DG, Whiting BR and Williamson JF (2006) Image reconstruction for transmission tomography when projection data are incomplete. Phys. Med. Biol, 51, 5603–5619 [PubMed: 17047273]
10. Wong W, Bai X, Brown A, Fernandez IS, Hanssen E, Condron M, Tan YH, Baum J and Scheres S (2014) Cryo-em structure of the Plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. eLife, 3, e03080
11. Zhao Y, Zeng X, Guo Q and Xu M (2018) An integration of fast alignment and maximum-likelihood methods for electron subtomogram averaging and classification. Bioinformatics, 34, i227–i236 [PubMed: 29949977]
12. Holland J (2012) Genetic algorithms. Scholarpedia, 7, 1482
13. Yang Z, Fang J, Chittuluru J, Asturias FJ and Penczek PA (2012) Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure, 20, 237–247 [PubMed: 22325773]
14. Frank J (2009) Single-particle reconstruction of biological macromolecules in electron microscopy —30 years. Q. Rev. Biophys, 42, 139–158 [PubMed: 20025794]
15. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I and Ludtke SJ (2007) EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol, 157, 38–46 [PubMed: 16859925]
16. Scheres SH (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol, 180, 519–530 [PubMed: 23000701]
17. Elad N, Clare DK, Saibil HR and Orlova EV (2008) Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J. Struct. Biol, 162, 108–120 [PubMed: 18166488]

18. Sorzano COS, Bilbao-Castro JR, Shkolnisky Y, Alcorlo M, Melero R, Caffarena-Fernández G, Li M, Xu G, Marabini R and Carazo JM (2010) A clustering approach to multi-reference alignment of single-particle projections in electron microscopy. J. Struct. Biol, 171, 197–206 [PubMed: 20362059]
19. Heymann JB, Conway JF and Steven AC (2004) Molecular dynamics of protein complexes from four-dimensional cryo-electron microscopy. J. Struct. Biol, 147, 291–301 [PubMed: 15450298]
20. Dubochet J, Adrian M, Chang JJ, Homo JC, Lepault J, Mcdowall AW and Schultz P (1987) Cryoelectron microscopy of vitrified specimens. Springer Berlin Heidelberg
21. Oikonomou CM, Chang Y-W and Jensen GJ. (2016) A new view into prokaryotic cell biology from electron cryotomography. Nat. Rev. Microbiol, 14, 205–220 [PubMed: 26923112]
22. Szeliski R, (2007) Image alignment and stitching: A tutorial. In: Foundations Trends R in Computer Graphics Vision, 2, 1–104
23. Lim A, Rodrigues B and Zhang X (2007) A simulated annealing and hill-climbing algorithm for the traveling tournament problem. Eur. J. Oper. Res, 174, 1459–1478
24. Hall RJ and Patwardhan A (2004) A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. J. Struct. Biol, 145, 19–28 [PubMed: 15065670]
25. Mallick SP, Zhu Y and Kriegman D (2004) Detecting particles in cryo-em micrographs using learned features. J. Struct. Biol, 145, 52–62 [PubMed: 15065673]
26. Sigworth FJ (2004) Classical detection theory and the cryo-EM particle selection problem. J. Struct. Biol, 145, 111–122 [PubMed: 15065679]
27. Ouyang J, Liang Z, Chen C, Fu Z, Zhang Y and Liu H (2018) Cryo-electron microscope image denoising based on the geodesic distance. BMC Struct. Biol, 18, 18 [PubMed: 30554569]
28. Sorzano COS, Recarte E, Alcorlo M, Bilbao-Castro JR, San-Martín C, Marabini R and Carazo JM (2009) Automatic particle selection from electron micrographs using machine learning techniques. J. Struct. Biol, 167, 252–260 [PubMed: 19555764]
29. Larose DT (2006) Data Mining: Methods and Models. Hoboken: Wiley

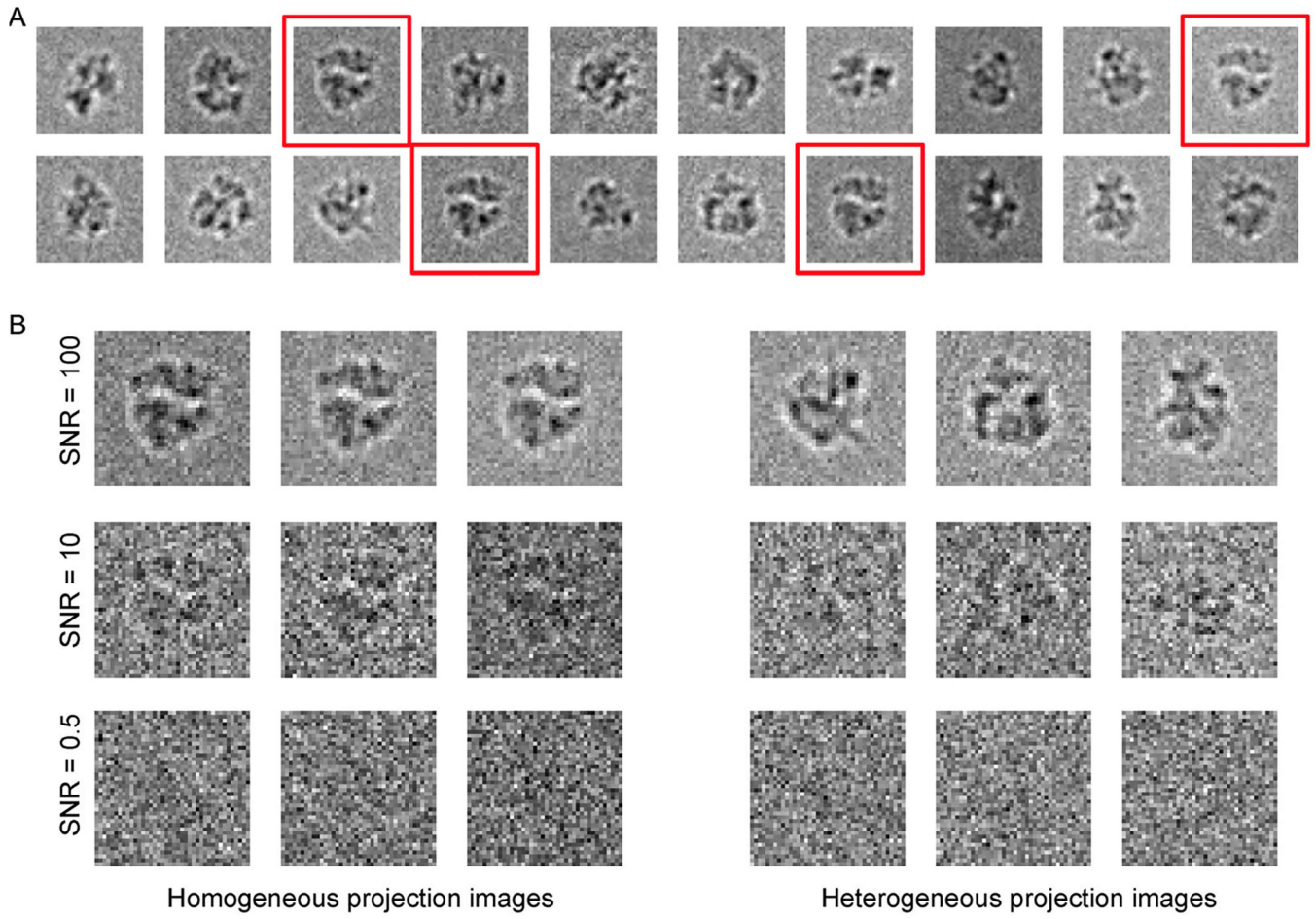

Figure 1. Examples of Cryo-EM image data for visualization purpose.

(A) Example of selecting homogeneous Cryo-EM images containing ribosome molecules of different orientations. The images are de-noised for visualization purpose. (B) Example of 2D Cryo-EM homogeneous and heterogeneous images containing projections of ribosome macromoleculars. Images with SNR = 100 are only for displaying the projections structure of the particle conveniently.

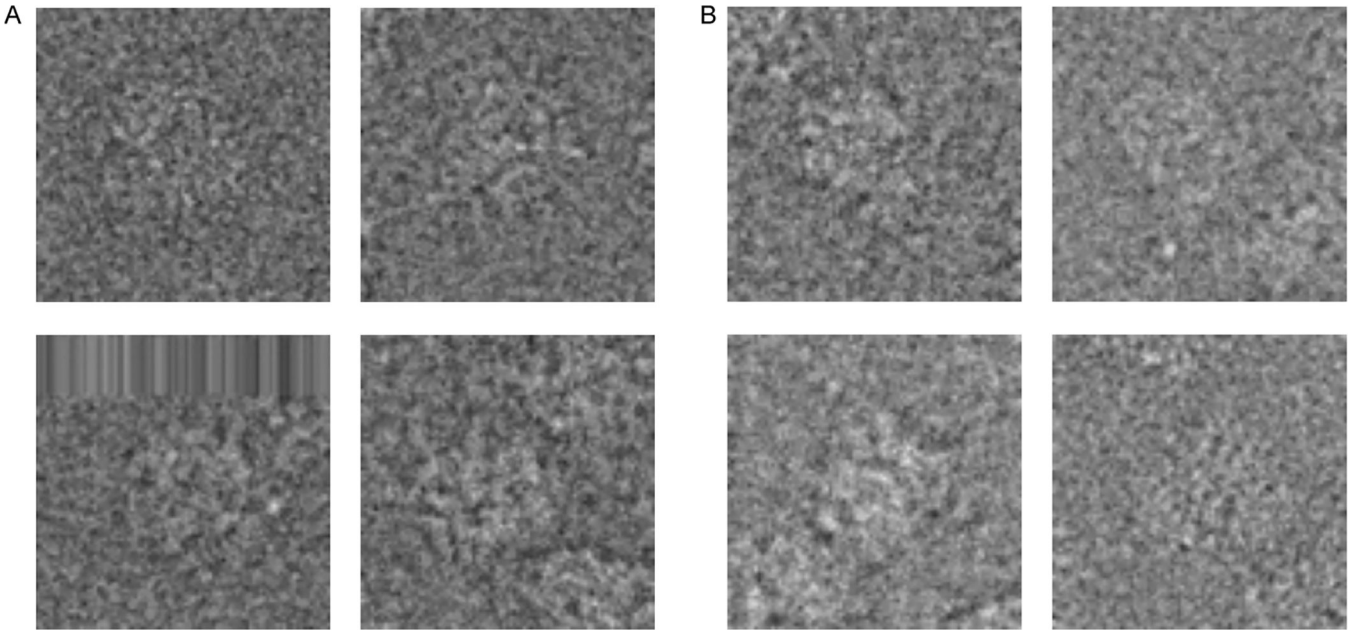

Figure 2. Examples of homogeneous and heterogeneous images in 2D experimental dataset.
The sub-figures (A) and (B) are obtained from different view directions.

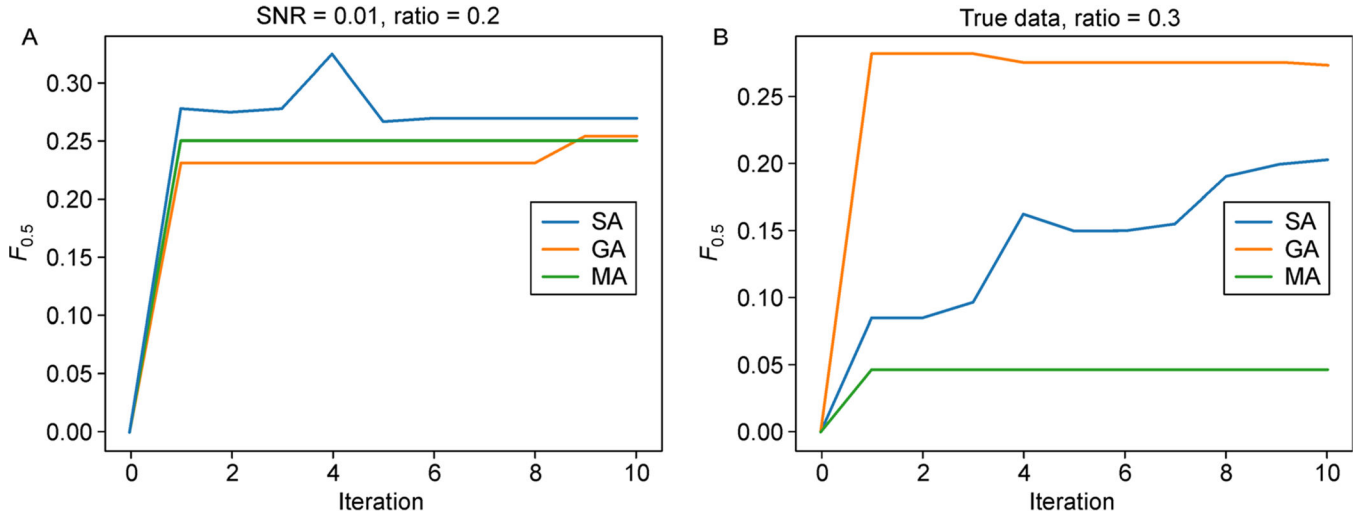

Figure 3. Results of comparing experiments of three algorithms on 3D simulated Cryo-ET image dataset with SNR = 0.01 and 3D experimental Cryo-ET image dataset.

(A) Simulated data. (B) Experimental data.

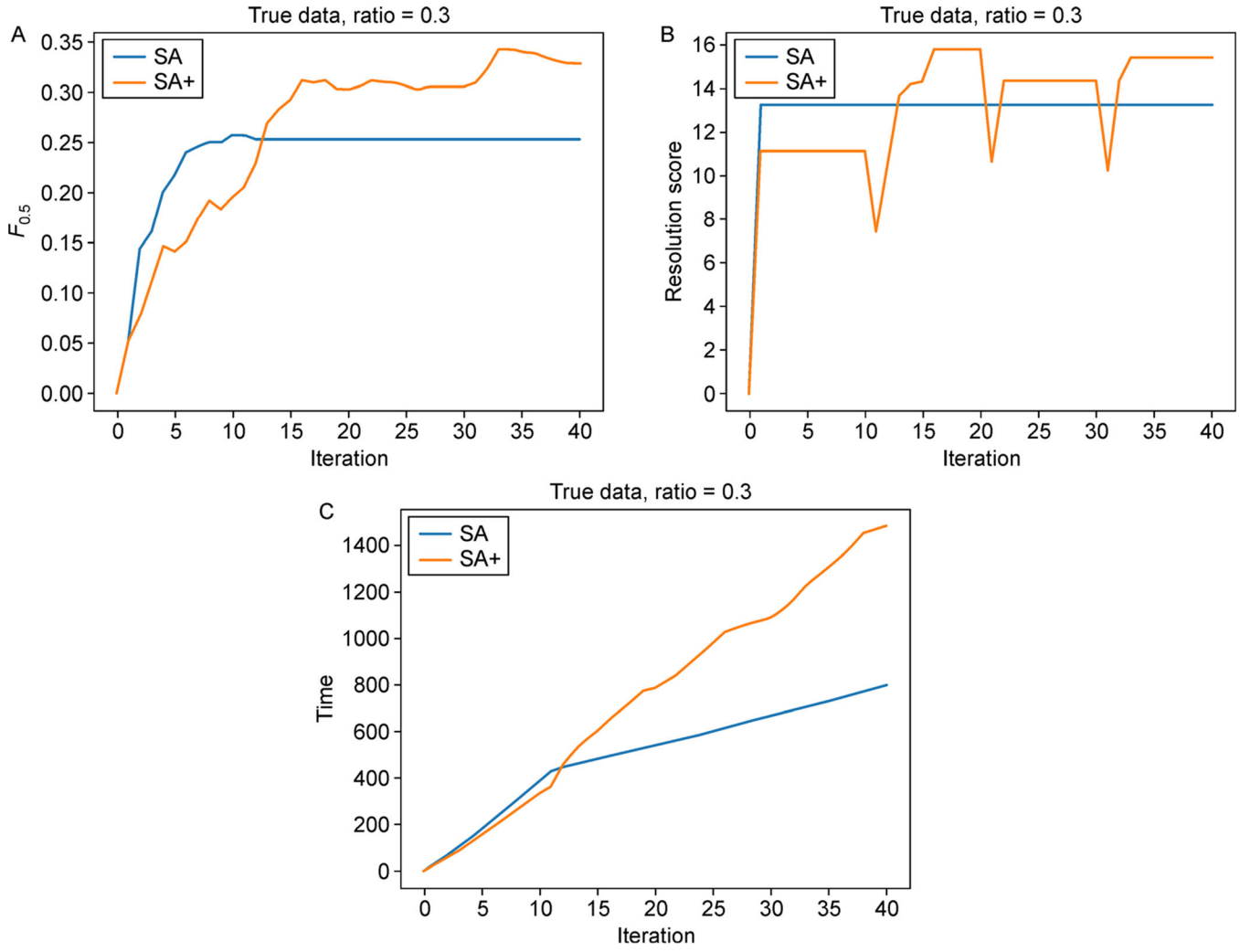

Figure 4. Results of comparing experiments of SA and SA+ on 3D experimental Cryo-ET image dataset.

(A) $\boldsymbol{F}_{0.5}$. (B) Resolution score. (C) Time.

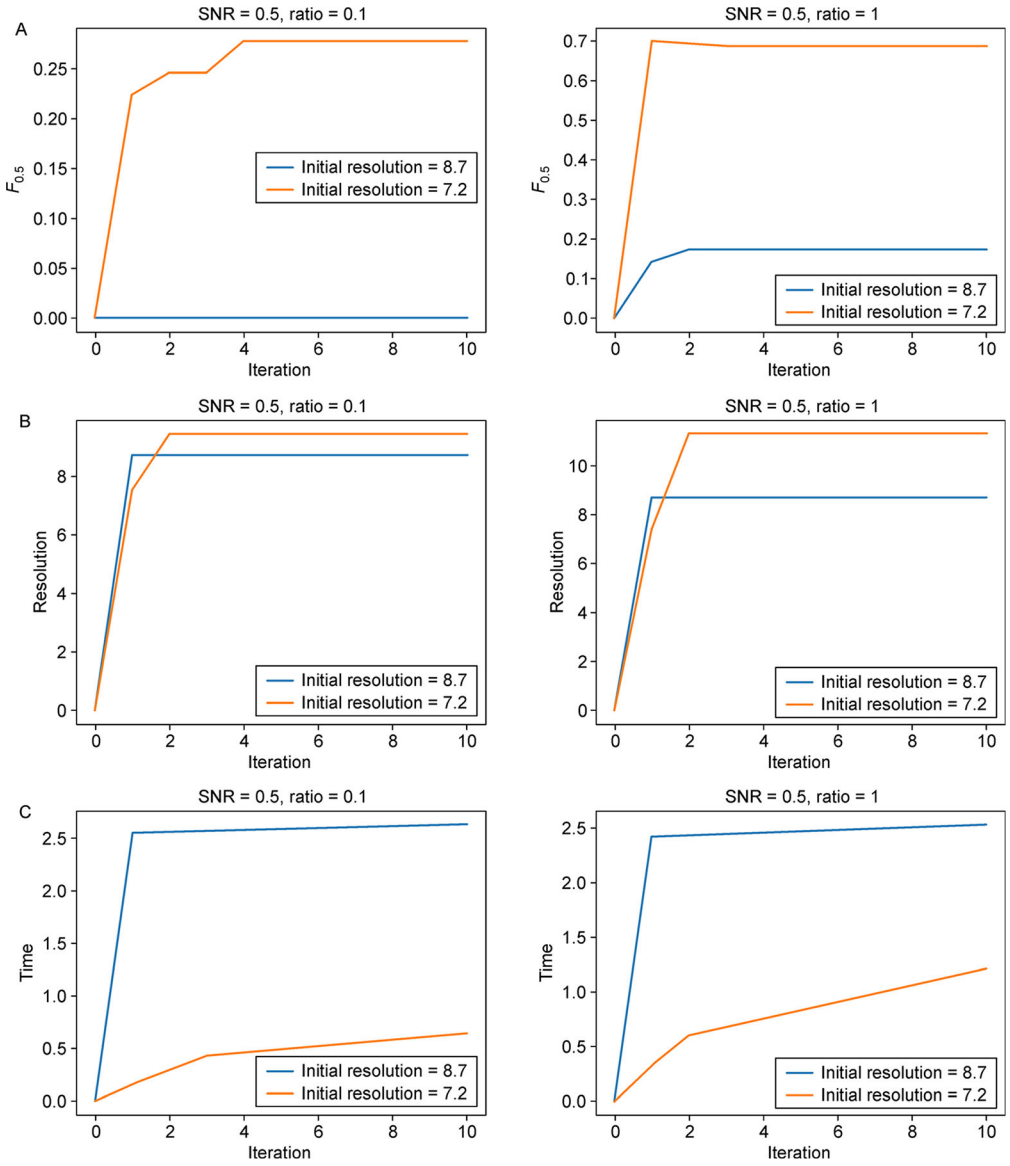

Figure 5. Results of the average value of 5 tests of three methods on 2D CEM image dataset with SNR = 0.5 and ratio = 0.1, 1.

(A) $\boldsymbol{F}_{0.5}$. (B) Resolution score. (C) Time costs.

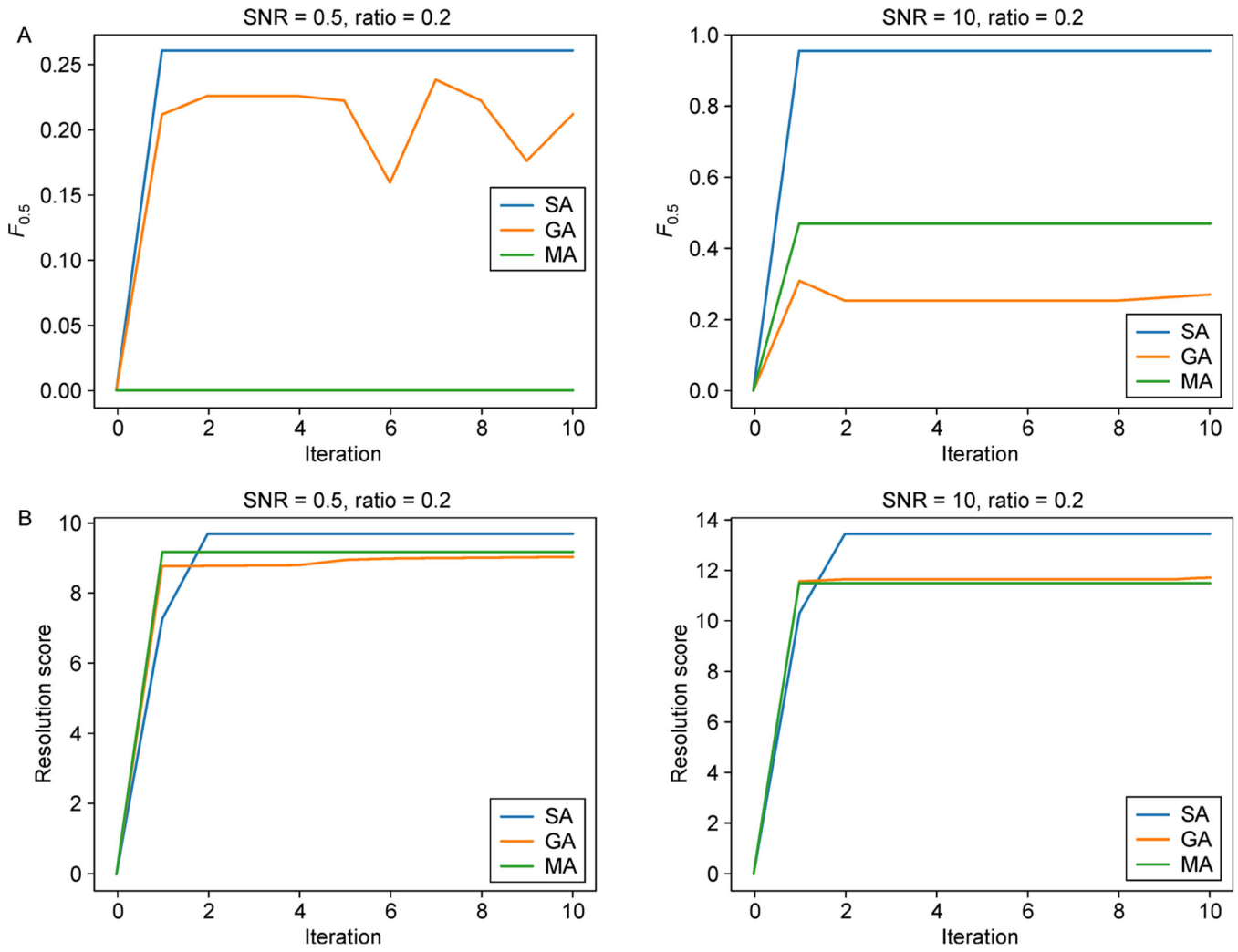

Figure 6. Results of the average of 5 tests of three methods on 2D cryo-EM image dataset with SNR = 0.5, 10.

(A) $\boldsymbol{F}_{0.5}$. (B) Resolution score.

Table 1

Average values of resolution scores, $F_{0.5}$ values and time costs of 5 tests of three methods after 10 iterations on 2D Cryo-EM simulated image dataset with SNR = 0.5

(A) SA

Ratio	Resolution score	$F_{0.5}$	Time (s)
0.1	9.91	0.19	2.34
0.2	9.70	0.26	0.59
0.3	10.35	0.49	0.78
0.4	10.18	0.53	0.66
0.5	10.19	0.58	0.69
0.6	10.51	0.67	0.64
0.7	10.83	0.68	0.81
0.8	10.90	0.71	0.92
0.9	10.99	0.78	0.81
1.0	11.53	0.78	1.30

(B) GA

Ratio	Resolution score	$F_{0.5}$	Time (s)
0.1	8.89	0.13	13.97
0.2	9.01	0.21	6.90
0.3	9.23	0.29	7.91
0.4	9.21	0.34	7.11
0.5	9.38	0.45	8.18
0.6	9.55	0.41	8.00
0.7	9.63	0.43	9.42
0.8	9.88	0.50	10.68
0.9	9.99	0.48	8.83
1.0	9.93	0.57	7.27

(C) MA

Ratio	Resolution score	$F_{0.5}$	Time (s)
0.1	9.08	0.00	0.82
0.2	9.15	0.00	1.64
0.3	9.39	0.12	0.89
0.4	9.48	0.10	1.18
0.5	9.66	0.08	1.33
0.6	9.84	0.13	1.52
0.7	10.03	0.06	3.42
0.8	10.23	0.05	1.78
0.9	10.41	0.05	1.60
1.0	10.49	0.13	1.75

Table 2

Average values of resolution scores, $F_{0.5}$ values and time costs of 5 tests of three methods after 10 iterations on 2D Cryo-EM experimental image dataset

(A) SA			
Ratio	**Resolution score**	$F_{0.5}$	**Time (s)**
0.1	9.83	0.18	70.36
0.2	9.96	0.26	91.76
0.3	10.23	0.54	110.37
0.4	10.17	0.52	198.33
0.5	10.17	0.57	226.09
0.6	10.43	0.64	422.26
0.7	10.79	0.67	133.07
0.8	10.98	0.70	198.11
0.9	10.99	0.73	174.91
1.0	11.46	0.76	315.59
(B) GA			
Ratio	**Resolution score**	$F_{0.5}$	**Time (s)**
0.1	8.74	0.02	1175.66
0.2	9.24	0.13	1184.08
0.3	9.31	0.27	899.59
0.4	9.36	0.37	829.94
0.5	9.35	0.36	1112.65
0.6	9.37	0.39	962.33
0.7	9.62	0.44	973.59
0.8	9.66	0.44	1096.13
0.9	9.77	0.48	1011.56
1.0	10.07	0.50	913.11
(C) MA			
Ratio	**Resolution score**	$F_{0.5}$	**Time (s)**
0.1	8.46	0.00	153.56
0.2	8.64	0.00	129.02
0.3	8.95	0.00	147.59
0.4	9.12	0.00	202.12
0.5	9.34	0.00	178.83
0.6	9.56	0.07	244.15
0.7	9.86	0.00	220.64
0.8	10.13	0.00	289.64
0.9	10.33	0.05	235.43
1.0	10.43	0.04	220.02

Table 3

Average values of resolution scores, $F_{0.5}$ values and time costs of 5 tests of three methods after 10 iterations on 2D Cryo-EM experimental image dataset with projection images of the particle from another view direction

(A) SA			
Ratio	**Resolution score**	$F_{0.5}$	**Time (s)**
0.1	9.93	0.21	81.35
0.2	10.16	0.20	79.76
0.3	10.33	0.49	115.98
0.4	10.27	0.59	179.13
0.5	10.37	0.57	198.35
0.6	10.53	0.66	227.90
0.7	10.87	0.68	215.89
0.8	10.78	0.73	233.01
0.9	11.19	0.72	259.37
1.0	11.46	0.79	201.39
(B) GA			
Ratio	**Resolution score**	$F_{0.5}$	**Time(s)**
0.1	8.94	0.05	1233.16
0.2	9.54	0.20	1458.01
0.3	9.31	0.18	1198.20
0.4	9.56	0.29	938.34
0.5	9.66	0.35	949.33
0.6	9.77	0.39	1087.64
0.7	9.62	0.47	1145.01
0.8	9.86	0.49	1086.37
0.9	9.77	0.48	1010.65
1.0	10.37	0.51	1002.29
(C) MA			
Ratio	**Resolution score**	$F_{0.5}$	**Time (s)**
0.1	9.06	0.00	113.89
0.2	8.94	0.00	115.76
0.3	9.45	0.00	139.01
0.4	9.51	0.00	101.99
0.5	9.64	0.00	158.59
0.6	9.65	0.00	147.02
0.7	9.83	0.03	208.29
0.8	9.82	0.03	277.20
0.9	10.22	0.05	207.98
1.0	10.43	0.04	217.35

Table 4

Results of comparing experiments of GA, SA and SA+ on 3D simulated Cryo-ET image dataset with SNR = 0.01 and ratio = 0.2

Method	Iteration times	Resolution score	$F_{0.5}$	Time (s)
GA	20	10.43	0.35	11215
SA	20	11.63	0.57	54
SA+	20	11.72	0.57	85
SA	30	11.63	0.57	72
SA+	30	11.72	0.70	110

Table 5

Results of comparing experiments of GA, SA and SA+ on 3D experimental Cryo-ET image dataset

Method	Iteration times	Resolution score	$F_{0.5}$	Time (s)
GA	20	8.82	0.28	46184
SA	20	14.20	0.30	784
SA+	20	14.34	0.32	788
SA	30	15.79	0.31	1087
SA+	30	15.79	0.36	1151