Published in final edited form as:

J Struct Biol. 2007 October ; 160(1): 11–27.

Averaging Tens to Hundreds of Icosahedral Particle Images to Resolve Protein Secondary Structure Elements using a Multi-Path Simulated Annealing Optimization Algorithm

Xiangnan Liu, Wen Jiang^{*}, Joanita Jakana, and Wah Chiu

National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030

^{*} Department of Biological Sciences, Purdue University, West Lafayette, IN 47907

Abstract

Accurately determining a cryoEM particle's alignment parameters is crucial to high resolution single particle 3-D reconstruction. We developed Multi-Path Simulated Annealing, a Monte Carlo type of optimization algorithm, for globally aligning the center and orientation of a particle simultaneously. A consistency criterion was developed to ensure the alignment parameters are correct and to remove some bad particles from a large pool of images of icosahedral particles. Without using any *a priori* model, this procedure is able to reconstruct a structure from a random initial model. Combining the procedure above with a new empirical double threshold particle selection method, we are able to pick tens of best quality particles to reconstruct a subnanometer resolution map from scratch. Using the best 62 particles of rice dwarf virus, the reconstruction reached 9.6Å resolution at which 4 helices of the P3A subunit of RDV are resolved. Furthermore, with the 284 best particles, the reconstruction is improved to 7.9Å resolution, and 21 of 22 helices and 6 of 7 beta sheets are resolved.

Keywords

common line; multi-path; simulated annealing; icosahedral particles; subnanometer resolution; consistency criterion; cryo-EM; electron cryomicroscopy

1. Introduction

Electron microscopy of macromolecules and single particle 3-D reconstruction were introduced about 40 years ago (DeRosier and Klug, 1968). Based on the central section theorem, alignment parameters of different particles can be determined from the intersecting lines (known as common lines) in Fourier space (Crowther et al., 1970b). The particles with determined alignment parameters can be combined to generate a 3-D density map. Icosahedral virus reconstruction was the first application of this common line methodology. Due to the icosahedral symmetry, there are thirty-seven pairs of self common lines present in a single particle and sixty pairs of cross common lines between two particles. Such redundancy of data in an icosahedral particle allowed sufficient orientation sampling with only a few particle images to yield a subnanometer resolution 3-D map. Theoretically, the number of evenly

sampled particles needed for a reconstruction is directly proportional to the size of particles (D) and inversely proportional to the resolution (d) (i.e. $n = \pi D/d$) (Crowther et al., 1970a). While ice embedding and low dose were later used to preserve the particles and record the images, it has been believed that a significantly larger number of particles than the theoretical estimate would be necessary to obtain a sufficient signal to noise ratio (SNR) for building an equivalent resolution map. As the methodology was used to yield higher resolution structures, the number of particles needed has also necessarily increased. Several thousand particles are now typically cited for the 3-D reconstructions of icosahedral viruses at subnanometer resolutions at which secondary structure elements are resolved (see review: (Jiang and Chiu, 2006)).

Alignment parameter determination of particle images is the most crucial step in the 3-D reconstruction. The original common line orientation search method was implemented by Crowther and co-workers (Crowther et al., 1970a). The first step determined the center by cross-correlation with either a circularly averaged template or the same particle image rotated by 180° . The next step was to find the initial orientation with the pre-fixed center using self common line search, typically through exhaustive search with a 1° sampling step size in an asymmetric unit. The final step refined locally the center and the orientation using cross common lines between raw particle images and reference images that are projections of existing 3-D model at known orientations. The refinement often used a local optimization strategy, such as Simplex (Fuller et al., 1996; Nelder and Mead, 1965; Zhou et al., 1998). However, the refinement is effective only if the initial center and orientation are close to their correct values.

This conventional procedure has been productively used to solve many ice-embedded icosahedral virus particles to subnanometer resolutions, but several drawbacks exist. The most significant one appears in early steps during the separate searches of the center and orientation in sequential order. The errors in the earlier steps, such as an incorrectly identified center, will inevitably be carried into later steps without the capability of self-correction. The accuracy of the self common line method is also sensitive to the particle orientation. For instance, when the particle is oriented near the symmetry axes, many of the self common lines are clustered or even degenerate. It is well recognized that this self common line method performs poorly for images with small defoci ($< 2 \mu\text{m}$). This is due to the decreased signal amplitudes in the low frequency range ($< 1/30 \text{ \AA}^{-1}$), on which the self common line method relied primarily. As a result, a focal pair imaging strategy was used to circumvent these problems where particle orientations were estimated first from the out-of-focus images and used to refine the close-to-focus images with an attempt to get a higher resolution map (Fuller, 1987; Zhou et al., 1998).

In contrast, cross common line search does not have these problems of self common line method and should be able to accurately determine the particle orientation and center parameters simultaneously as shown previously (Zhou et al., 1998). However, the potential of the cross common line method is inadvertently limited by the narrow convergence range in the current software implementation that only supports local refinement (Zhou et al., 1998). This limitation is due to the astronomical number of search steps ($> 10^7$) if an exhaustive search in the whole asymmetric unit were undertaken even at 1 degree and 1 pixel step sizes for the 3 orientation and 2 center parameters. The exhaustive search is computationally too expensive to be used for the search of five parameters (center and orientation) simultaneously.

In our approach, the exhaustive search of cross common lines is converted to an optimization problem. The optimization is performed using a newly developed global optimization algorithm, multi-path simulated annealing (multi-path SA), which dramatically improves the search speed and the precision of these alignment parameters over conventional SA. This method needs only $\sim 10^4$ steps to search for the five parameters simultaneously. Since it is a

Monte Carlo algorithm, the result is not guaranteed to be accurate, or even close to optimal solution. A consistency measure is necessary to screen for the most probable result.

To arrive at the final map, an iterative procedure is used in which intermediate maps are reconstructed using approximations of the centers and orientations of the particles determined by our algorithm; the centers and orientations used in an intermediate map are refined in the next iteration to produce the next intermediate map. This procedure is carried out until convergence is reached. The validation of our method was carried out by applications of this algorithm on both simulated and real cryo-EM data of icosahedral virus particles of known structure. We also evaluated the minimum number of particles necessary to reach a subnanometer resolution map and detect secondary structure elements of the protein components.

2. Methods

2.1 The image alignment and 3-D reconstruction procedure

Figure 1a illustrates our 3-D reconstruction procedure. Usually a reconstruction procedure requires raw particles and an initial model as inputs. A final map is obtained after some rounds of refinements. In our method, an initial model is built by using a set of raw particles assuming random orientations and particle centers at image box center. In each round of refinements, since our algorithm is based upon the cross common line method in Fourier space, we are able to choose the resolution range (top left box of Figure 1a) for the raw particles in order to progressively refine the alignment parameters. We are also able to choose global or local search in each round of refinements (top left box of Figure 1a). In general we perform a few global searches followed by several local searches.

There are three layers of loops in Figure 1a: an outer, middle, and inner loop. The outer loop is used to execute multiple rounds of refinements mentioned above. The middle loop goes through the N raw particles. The inner loop is used to generate multiple (M) candidate solutions for each particle. Typically, M is set to a value between 7 and 10. A single execution of the step described in the shaded box of the inner loop of Figure 1a will generate one candidate solution, which will be detailed in Figure 1b. In preparation for alignment determination, we compensate for the effects of the contrast transfer function (CTF) on the raw particle image by applying the CTF to the reference images, rather than de-convolute the CTF from the raw particle images.

The set of M candidate solutions for the i^{th} particle is submitted to the consistency check (see section 2.3). If this solution set passes the consistency criterion, the particle will be considered “good”; otherwise, it will be treated as “bad” and set aside temporarily in this middle loop. For a good particle, the best of the consistent solutions will be chosen as this particle’s alignment parameters. The middle loop reduces the number of particle images from N to N' , which will be used for the subsequent 3-D reconstruction.

The N' good particles resulting from the middle loop are used to generate a 3-D map using the *make3d* program from EMAN (Ludtke et al., 1999). If convergence is met (i.e., there is no improvement in the map resolution over the previous iteration of the outer loop), then we stop and output the final map. Otherwise the 3-D map will be used as a new template to generate reference projections for subsequent iterations of the inner and middle loops to improve the accuracy of the alignment parameter determinations. Subsequent iterations of the outer loop will progressively include higher resolution region data for alignment parameter determination, though the data will only be a fraction of the Nyquist frequency. However, the 3-D map is always generated using all of the data up to Nyquist frequency as commonly practiced in other reconstruction software [e.g. (Ludtke et al., 1999)] simply because one does not know a

priori how high the resolution of a reconstruction may be reached. Of course, image denoising can always be applied after the 3-D map is computed.

2.2 Multi-path simulated annealing in a global alignment search for a solution

To obtain an accurate solution for the alignment parameters, we have taken an approach of simultaneously searching all five alignment parameters using only cross common line. To make this feasible, we converted the problem to an optimization task instead of an exhaustive search. The conversion is mathematically equivalent to finding the minimum of an objective function. In this application, the objective function, called residual (R), computes the normalized summation of the discrepancies between the Fourier elements of each pair of the cross common lines whose locations are calculated (see Supplement 1) from the known orientation of the references and trial orientation of the raw particle (bottom center box in Figure 1b). There are different metrics for measuring the R value in our implementation. The R value in our test cases was calculated by averaging the cosines of the phase differences. As the normalized residual R approaches 0, the correlation between the common line pairs approaches the maximum. $R=1$ indicates negative correlation. If there is no correlation, $R=0.5$. The expected R value of a good solution should be less than 0.5. Usually, a smaller R value will be produced if the raw particle images are less noisy and lower resolution region data are used.

2.2.1 Choice of optimization algorithm—There are many global optimization algorithms (Hartmann and Rieger, 2004; Press et al., 2002). Simulated annealing (SA) (Kirkpatrick et al., 1983; Metropolis et al., 1953) is one of the most widely used stochastic optimization algorithms for global search. This optimization is equivalent to finding the lowest energy state of an objective function in a physical process of heating and slowly cooling a substance to form a highly ordered structure. Theoretically, SA can find the global minimum of any given objective function, provided sufficient CPU time, but in reality, there is no guarantee of finding the correct global minimum in a finite amount of time (Laarhoven and EHL, 1987). The problem lies in the possibility of getting permanently trapped in a local minimum rather than finding the global minimum. Conventional SA search starts from a single point; thereafter all the subsequently accepted points link to a single path. The conventional SA is often hampered by the fact that it uses a single path to perform the search. To avoid getting trapped in a local minimum along such a single path, our algorithm runs on multiple paths that communicate with each other during the search to mutually and coordinately approach the global minimum. We name this global optimization algorithm *multi-path SA*.

Figure 1b shows the flowchart of the multi-path SA algorithm in the context of cross common line search, which describes the details of the shaded rectangle on the top right corner of Figure 1a. The algorithm takes two inputs (left column of Figure 1b): a raw particle image and a set of reference projections. The output solution comes from the best solution found among all the paths. The annealing cycles of this algorithm are split into three stages with different annealing temperatures (see section 2.2.3). Because of the multiple annealing stages and multiple paths, the path selection and update schemes are different in the last stage compared to the first two stages. The following subsections will address the details of this multi-path SA.

2.2.2. Path initialization, selection and evolution—The multi-path SA starts from multiple points; thereafter each of the start points links to its subsequently accepted points to construct one of the multiple paths. Each of the start points is randomly initialized as one of the starting trial solutions. Normally the number of paths (i.e. the number of starting points), P , is set to 10–15. Since the multi-path SA does not strictly stay around the best trial solution, two solution lists are used to keep track of the best solutions found so far by each path (*best solution list*) and the current trial solutions of every path (*current solution list*). When the search starts, all the values of the randomly initialized start points are copied to each of the two lists,

and this is followed by a search in the whole space. Usually the global search involves only the data of lower resolution of a raw particle image.

Conceptually, the individual paths run simultaneously, in parallel. In the actual code implementation, the paths run consecutively in a single process. The selection of the path to be evolved next is quasi-random with weighted priority given in the first two stages (see section 2.2.3) to paths based on their residual values in the current solution list (middle center box in Figure 1b) with worse-performing paths getting higher priority. The path selection is random in the third stage.

When a solution S_p is selected from the current solution list, the algorithm will generate a trial solution S_p' by introducing a random change to S_p (lower center box in Figure 1b), which is limited by the corresponding step size (Supplement 2). The center is varied linearly; on the other hand, since the search will result in non-uniform sampling in the rotation space if the Euler angles are also updated linearly, quaternion (Hamilton, 1944; Hamilton, 1947; Kuffner, 2004) instead of Euler angles is used to represent a rotation (Supplement 2).

The common lines for the trial solution S_p' between the raw particle and reference projections can be located by the trial orientation and center (S_p') and the known parameters of the references, which are detailed in Supplement 1. The trial residual R_p' can be calculated by one of the residual metrics.

Once the residual R_p' of a trial solution S_p' is computed (bottom right corner box in Figure 1b), the algorithm decides whether this trial solution is accepted or rejected (lower right diamond in Figure 1b). The decision is made according to the classic Metropolis criterion (Metropolis et al., 1953), which always accepts a better solution S_p' but also accepts a worse solution S_p'

with a certain probability $P = \exp\left(\frac{-(R_p' - R_p)}{T}\right)$, where the control parameter, T , is current temperature in the cooling schedule. At high temperature, the trial point, S_p' , has a higher probability for accepting points with worse residual, which helps to jump out local minimum and locate the approximate position of the global minimum. As a result, large step sizes should be used to generate the trial point, S_p' , from current point, S_p . On the other hand, at low temperature the trial point, S_p' , has more chance to sample locally, which helps to find the deepest point of a minimum.

Path evolution involves the generation and acceptance a trial solution, but in our implementation it also involves a choice of the path to which an accepted trial solution should be appended. The choice is governed by the path update scheme that is in effect. In the first two stages (see section 2.2.3), paths update themselves only with the accepted trial solution originated from the same path, whereas in the third stage, an accepted trial solution will replace the worst-performing element of the current solution list and will then be the new current solution associated with that path. This essentially steers the worst performing path to follow other paths that have found better solutions. The best solution list is always updated based on each individual path (upper right box in Figure 1b). If the accepted trial solution has a residual smaller than any one achieved previously in its path, the corresponding solution will be updated in the best solution list. This will also be discussed in the following section.

2.2.3. The three annealing stages—To find the solution that corresponds to the global minimum, the multi-path SA is divided into three different stages with varying annealing temperatures, path update schemes, and numbers of trials for each stage (the box below the upper middle diamond in Figure 1b).

In the first stage, the annealing cycles at high temperatures with large step sizes. Each solution S_p in the current solution list travels globally through the entire search space, while being updated by the trial solution S_p' as it was in conventional SA. As a result of high temperature annealing, most of the trial solutions are accepted. At the end of this stage, the solutions in the best list for all the paths are evenly distributed among the whole orientation space and are taken as the starting solutions for the second stage (the second box below the upper center diamond in Figure 1b).

The second stage performs at lower temperatures with relatively small step size. The step sizes of both the center and orientation changes are gradually reduced to a few pixels and degrees. It remains a global search; however, since this step is done at a lower temperature, its search range is relatively limited, (i.e. searching more locally around each of the starting points). This stage is the most computationally expensive step among the three stages. With multiple starting paths, it is expected that at least one of the paths will approximately reach the global minimum. These “best solutions” are used as the starting solutions of the third stage (the second box below the upper center diamond in Figure 1b). Again, each of the current solutions is updated independently for each path.

The third stage runs at very low temperature with small step size, while utilizing a different path updating scheme. To precisely determine the final solution, the step sizes are gradually reduced to very small values depending on the size of the particle and the desired accuracy. For all the paths to converge to the global minimum, the worst solution in the current solution list will be replaced by a new trial solution of better residual from any of the paths. Due to the small step sizes, low temperature, and the cross-path update scheme, the deepest minimum around the candidate solutions from the second stage will be identified. If one of the best solutions from the second stage is near the actual global minimum, it will be captured in the third stage and refined with very high precision.

2.3. Using the consistency criterion to improve confidence in a solution

Multi-path SA produces a very precise result. However, there is still no guarantee that this result actually is the global minimum. To increase confidence in the result, a *consistency criterion* has been used (center shaded diamond in Figure 1a). For the same raw particle image and reference images, the same search procedure as shown in the top right corner shaded box in Figure 1a and detailed in Figure 1b was repeated independently multiple (M) times in the inner loop of Figure 1a, which would generate M candidate solutions (the idea is similar to the multi-start refinement in X-ray crystallography (Brunger et al., 1998a)). The M solutions could be quite different from each other due to the inherent properties of the Monte Carlo algorithm. If some of the results are consistent within given tolerable errors, confidence is increased in the assertion that the consistent solutions are near the global minimum (i.e. close to true center and orientation). For example, we can repeat the search seven times (i.e. $M=7$) and require that four of the seven solutions should be consistent within error limits of 2.0 pixels and 2.0 degrees with respect to center and orientation (lower right side boxes in Figure 1a). The specific allowable error limits can be derived based on the fact that no real space point in the structure should be off by more than $2d/A$ pixels from its actual value, where d is the expected resolution and A is pixel size.

2.4. Running the multi-path SA for a local alignment refinement

In addition to the *global search* which uses only low resolution region data, a *local refinement* mode was also developed to achieve a high accuracy solution with the inclusion of higher resolution data. The local refinement takes place after several iterations of global alignment searches because the solution for the alignment parameters of a particle derived from the global searches can still be further improved though it is quite accurate. The multi-path SA

for local refinement runs the same three annealing stages as the ones in global search (upper center box in Figure 1b) but the some internal annealing control parameters are different and the search ranges are limited to stay close to the starting points. The starting points can be read in from previously determined solutions of a particle image. At low resolution, the handedness of some particles could be ambiguous compared with the reference model (*i.e.* sometimes the residuals of the two handedness-related orientations are close). Therefore, the handedness-related orientations associated with the read-in solutions are also considered as part of the starting points. All of these starting points initiate both the best and current solution lists (upper center box in Figure 1b).

2.5 Selecting better quality particles

In general, when the number of particle images is fixed, better quality images result in a superior map. By the same token, given a map, the same resolution can be achieved with fewer particle images provided they are of higher quality.

Quality of a particle image can be affected by the conformational homogeneity, preservation of the icosahedral symmetry of the particle, and the imaging conditions (astigmatism, beam-induced movement (Saad et al., 2001)). All of these quality factors will be reflected in the value of the residual. Also accurate alignment parameters given by a computational algorithm will result in smaller residual. Usually a particle image with a smaller defocus value will have a better signal to noise ratio in the high resolution region (Saad et al., 2001).

According to the distribution of the residuals of the accepted particles in each selected micrograph with smaller defocus value, better quality particle images can be empirically selected by a double threshold method. The selection requires the residual R of a particle image to be better than two residual thresholds $\bar{R}_1 - x\sigma_1$ and $\bar{R}_2 - y\sigma_2$ (*i.e.* $R \leq \bar{R}_1 - x\sigma_1$ and $R \leq \bar{R}_2 - y\sigma_2$), where $\bar{R}_1$ and $\bar{R}_2$ correspond to the average residual value of all the particles in a micrograph and the average residual of all the particles in all the micrographs, respectively; σ_1 and σ_2 represent the variances of residuals respectively; and the variables, x and y , affect the number of better particle images to be selected. For instance, larger x and y values would impose a more stringent selection criterion. The values of x and y are usually set to same values in the empirical formula above, but they can be individually adjusted.

3. Results

Rice dwarf virus (RDV) is a double shelled particle containing a double-stranded RNA genome (Zhou et al., 2001). The shell diameter is about 700 Å and has a total mass of greater than 26 MDa. Both outer and inner shells have icosahedral symmetry. The outer capsid is revealed as a T=13/ icosahedral shell which contains 260 trimers of P8 (46 kDa). The T=1 inner capsid is composed of 60 dimers of P3A and P3B (114 kDa) which have identical sequences but slightly different conformations. The genome is composed of 12 double-stranded RNA segments. The structure of RDV has been solved by cryo-EM (Zhou et al., 2001) and X-ray crystallography (Nakagawa et al., 2003). In this study, RDV was used as the test case to validate the algorithm development.

3.1 Simulated data of rice dwarf virus

For testing purposes, 2000 simulated images were randomly projected from a RDV atomic model (PDB ID: 1UF2). Different CTF parameters from the micrographs shown in Figure 2 were applied, the centers were shifted, and different levels of noise were added to the projected images. The noise was added as white noise which has constant power in all frequencies. It is different from the real noise present in raw cryo-EM particles images. In general, when the

noise level is in the range of $20\sigma_{\text{img}}$ to $25\sigma_{\text{img}}$ (σ_{img} is the variance of a projection image), the visual contrast is equivalent to the real cryo-EM images.

For a small noise level (i.e. $5\sigma_{\text{img}}$), the newly implemented multi-path SA algorithm can find center and orientation of all images with average errors less than 0.1 pixel and 0.2 degrees respectively (Table 1). For a higher noise level (i.e. $25\sigma_{\text{img}}$), it can determine the parameters for more than 95% of the images with average errors less than 0.4 pixels and 0.5 degrees (Table 1). This level of accuracy for both center and orientation is sufficiently good to allow 3-D reconstruction of RDV to 3 Å resolution.

3.2 Cryo-EM images of RDV

The electron micrographs were taken from RDV particles embedded in vitreous ice at -170°C in a JEM4000 electron cryomicroscope with a LaB₆ gun, operating at 400kV with a dose of $10\text{--}13\text{ e}/\text{\AA}^2$ (Zhou et al., 2001). The nominal microscope magnification was set at 50,000 x. The micrographs were recorded as focal pairs for our original common line method that used self common line and cross common line in tandem (Zhou et al., 2001). However, only the first close-to-focus images were used in the current study. The defocus values of these close-to-focus images varied from 0.32 to $2.87\mu\text{m}$, and were mostly in the range of 0.5 to $2.0\mu\text{m}$. This set of data was originally digitized with a Zeiss scanner that led to the 6.8 Å map by using the focal pair processing strategy (Zhou et al., 2001). In the current study, all the micrographs were re-digitized with a MRC-built scanner with a step size of $4\mu\text{m}$ (Henderson et al., 2007). Each of the boxed-out 4,865 particle images of size 1200×1200 pixels from 100 digitized micrographs was averaged three times for subsequent data processing. The final sampling of the particle images was done at $2.4\text{ \AA}/\text{pixel}$. Figure 2 illustrates some typical RDV particles with different defocus values.

We first built an initial random model (the light shaded box with dash border in Figure 1a), and then performed 5 iterations (the outer loop in Figure 1a) to globally refine the 3-D map from a random initial model using increasingly higher resolution data (up to $27\text{\AA}$). 7-12 reference images were used in the global searches. Finally local refinement was employed to include higher resolution data (up to $10.6\text{\AA}$) in the last few iterations (the outer loop in Figure 1a) of refining the 3-D map of RDV.

In each of the 5 iterations of global search mentioned above, the multi-path SA was repeated seven times to generate seven candidate solutions (i.e. $M=7$, inner loop in Figure 1a) for each particle, and then the consistency criterion (middle loop in Figure 1a) was applied to the seven candidate solutions. If four of the seven solutions were consistent within an error of 2.0 pixel and 2.0 degree with respect to the center and orientation, the particle was considered good and used in the 3-D reconstruction.

To produce each of the seven candidate solutions for each particle image, approximately 700, 1100 and 700 iterations (the loop in Figure 1b) were used in the three stages of the multi-path SA algorithm, respectively, as described in the Method. We searched the orientation within a range slightly larger than an asymmetric unit, and the center search was limited to within 40 pixels of the center of a boxed-out image in the X and Y directions. Twelve projections were used as reference images. The CTF modulations were applied to reference projections. As mentioned previously, here the residual R was calculated by averaging the cosines of the phase differences. All FFT data were given equal weights for the residual calculation. Residual usually varied from 0.3 to 0.45 depending on the resolution range of the data used.

In the last few iterations of locally refining the 3-D map from the result of the global searches, the search range, number of references, tolerable error of the consistency criterion, and number of trials for each annealing stage were varied (see section 3.2.3).

For visualization and for measuring resolution, the high frequency region data of the reconstructed map was low-pass filtered using the *lp* option of *proc3d* program in EMAN, and the outside noise and interior dsRNA density were removed using soft masks. The resolutions of the 3-D reconstructed maps were measured against the known X-ray structure (PDB ID: 1UF2) (Nakagawa et al., 2003) using the 0.5 criterion in the Fourier Shell Correlation (FSC) (Harauz and van Heel 1986) curves. The secondary structure elements were identified computationally by SSEHunter (Baker et al., 2007) and visualized with Chimera (Pettersen et al., 2004). Further details of the whole refinement and reconstruction procedure are presented below.

3.2.1 Initial model—The current implementation does not require a user-supplied initial model. It builds an initial model internally (the light shaded box with dash border in Figure 1a), with random orientation assignments for each of the 4,865 particles and the center of a particle is assumed to be at the center of the particle image box. The required number of particles for building initial model can be as few as 10. In our study, we assigned random orientations for the whole dataset, instead of picking a few particle images. Icosahedral symmetry was enforced in the reconstruction. The initial model appears like a ball because of the random orientations used for the raw particles. This initial model provides a centered icosahedral ball with an approximately correct particle size. The central slice of the initial model is shown in Figure 3a(0), which shows the two capsid shell proteins having higher density than the inner RNA. The corresponding FSC is shown in Figure 3b(0), indicating a resolution $\sim 100\text{\AA}$.

3.2.2 Global alignment searches with different data resolution ranges—The first iteration of the 3-D map refinement (the outer loop in Figure 1a) of RDV started with very low resolution data for the global alignment search. We used only data lower than $136\text{\AA}$ resolution, i.e. data from the first 7 pixels of the FFTs of the raw image and references. About 83% of the 4,865 particles passed the consistency criterion at this resolution. From the central slice of the reconstructed map of this iteration (Figure 3a(1)), it can be seen that the map's quality is slightly improved over the random initial model and the resolution was measured at $\sim 45\text{\AA}$ (Figure 3b(1)). Two rings can now be seen and the gap between the two rings splits the capsid shell and the inner RNA.

The second iteration used slightly higher resolution (up to $64\text{\AA}$) data (i.e. data from the first 15 pixels of the FFTs of the images). This map (Figure 3a(2)) shows two inner RNA rings and other features on the outer capsid shell, although they are still blurred. The resolution of this map is about the same as that of the first iteration (Figure 3b(2)).

The same data range (up to $64\text{\AA}$) as in the second iteration of 3-D map refinement was also used in the third iteration. The map (Figure 3a(3)) begins to reveal the trimer appearance with a resolution $\sim 33\text{\AA}$ (Figure 3b(3)).

In the fourth iteration, we only used data that were lower than $38\text{\AA}$ resolution (the first 25 pixels of the FFTs) for refinement. However, the resolution of the reconstructed map ($12.5\text{\AA}$) improved substantially in this iteration. As a result, more features of the capsid shells became apparent (Figure 3a(4)).

We observed that alignment parameters can be determined very accurately using very low resolution data. To further emphasize this fact, one more iteration (5th iteration) for global alignment search was done by including data only at low resolution ($27\text{\AA}$) for refinement. It generated a map with very clear high resolution features at a resolution of approximate $8.2\text{\AA}$ (Figures 3a(5) and 3b(5)). This suggests that the orientation determination is rather accurate up to this point.

For each of the five iterations of global search performed, we used the same 4,865 raw particles; roughly 83–87% of the aligned raw particles passed the consistency criterion each time. However, notable differences can be found over the course of the process by comparing the alignment parameters of each individual particle from one iteration to the next. Between the 1st and 2nd iterations, the alignments of only 20 particles were in agreement within 3 pixels for center and 3 degrees for orientation. The alignments of around 200 of the particles were in agreement (the same criteria above) between the 2nd and 3rd iterations, and between the 3rd and 4th iterations. But between the 4th and 5th iterations, about 2600 particles were in agreement (the same criteria above).

3.2.3 Local alignment refinements—Once a map reached a sufficiently high resolution (as in the maps from the 4th or 5th iterations of the global search described above), the local alignment refinement procedure was launched, which then converged very quickly to the final map in just a few iterations.

Essentially, the local alignment refinement is the same procedure as the global alignment search, except for the fact that the starting points are the results of the earlier global searches instead of being randomly assigned (top middle box of Figure 1b). Also, the step sizes for randomly varying the center and orientation (lower middle box of Figure 1b) are limited to 5 pixels and 5 degrees or even smaller, using previous solutions as possible starting points. Their flipped handedness was also considered as starting points. As a result, the search was restricted to the small ranges around these starting solutions. To increase the signal to noise ratio, we have used up to 20 reference images. The tolerated errors of the consistency criterion were reduced to 1 pixel and 1 degree, respectively. The tolerable errors in the local refinement must be limited because there are many local minima due to the high resolution data included in refinement. To refine the solutions, the multi-path SA still ran in 3 stages, but the first two stages do not use much CPU time compared to the global alignment search (for example, 400, 400, and 600 trials per particle used for the three stages, respectively).

Two iterations of local refinements were done. We used data that ranged from 190 Å to 16 Å for iteration #6, and from 63 Å to 10.6 Å for iteration #7. Figure 3a(6–7) demonstrates the corresponding center slices with high resolution features. The FSC curves (Figure 3b(7)) indicate 7.3 Å resolution as measured against the crystal structure blurred to 5 Å and the inner dsRNA was masked out after the two maps were scaled by *Foldhunter* (Jiang et al., 2001). However, the FSC curve (Figure 3b(7 half-half)) between the two half-dataset reconstructions indicates 6.3 Å resolution.

3.2.4 The final map with 4200 particle images—Figure 4a shows the final reconstructed map from the 7th iteration of the 3-D refinement which was reconstructed from 4200 good particles out of 4,865 boxed raw particles. To clearly visualize the structure details, Figure 4b shows the stereo view of the outer capsid P8 subunit and Figure 4c-d show the two side views of the inner capsid subunit P3A. Crystal structures of both subunits are superimposed on the cryo-EM map exhibiting their excellent matches. The surface of the secondary structure elements in P3A subunit appears bumpy indicating the high quality of the map as expected at this range of resolution.

3.3 Map qualities as a function of number of particles

In our study, by only using the micrographs with defocuses smaller than 1.6 μm and by varying the parameters x and y which were defined in section 2.5, we were able to select different subsets of best quality particles from the total set of 4200 good particle images which were used to generate the map shown in Figure 4. The sizes of the respective subsets varied from 1755 to 23 particle images, with the expectation that the average quality of particles in a subset

increases as the sizes of the subsets decrease. For each individual subset, we reconstructed a density map from scratch: build initial model from randomly assigned orientation and go through the multiple iterations similar to the process used for the whole dataset as discussed in section 3.2. During the processing of the different subsets, we found it only took 6 to 9 iterations (outer loop of Figure 1a) to get a converged 3-D map from scratch for the subsets with greater than 100 best particle images. For subsets of 81 and 62 best particle images, approximately 15 iterations were required for convergence. For the subset of 41 best particle images, it took about 20 iterations to converge. We were not able to obtain a converged map from the dataset with the 23 best particles from scratch.

Figure 5 illustrates the P3A subunits of RDV density maps (segmented with a known P3A mask) reconstructed from subsets of no more than 284 particles, superimposed with the X-ray structure; their resolutions, which ranged from 7.9 to 9.8Å, were measured against the X-ray structure. The maps reconstructed from both 1470 and 1755 particles were practically identical to that reconstructed from the whole data set of 4200 particles and are not shown in the figure. With the best 284 particles, 20 out of 21 expected α helices and 6 of 7 large β sheets of the subunit P3A can be identified by *SSEHunter* from the reconstructed map at 7.9Å resolution. With the best 181 particles, the map's resolution is slightly lowered (8.3Å) and 16 helices and 4 sheets were identified. When the number of best particles is reduced to anywhere from 101 to 180 particles, most of the helices and some sheets in the P3A subunit of the reconstructed map can be identified by *SSEHunter*. With 62 and 41 particles, *SSEHunter* identified 4 and 1 helices respectively, though visually the density distribution and connectivity appear well matched to the crystal structure.

3.4. Maps generated with various subsets of 284 particles of different qualities

In general, better quality particle images should yield a better quality map. To evaluate if our method of selection (section 2.5) is capable of properly identifying better quality particles we tested three more 284-sized datasets. 284 particles were randomly selected from the 4200 good particles; 284 particles were randomly selected from the 665 bad particles; and 284 particles were randomly selected from among all the 4,865 boxed-out particles. Models were constructed starting each time from random initial models. These three test-sets were all expected to have poorer quality than the subset of 284 best particles since their selection was not based on the selection method that produced what we considered the best particles.

In the case of the 284-subset of best particles, after the 3rd iteration, the quality of the reconstructed map had undergone a substantial change. After 6 or 7 iterations, it converged to a final map at 7.9Å resolution based on the FSC curve against the known X-ray structure with the clear visualization of the secondary structure elements. The segmented P3A subunit superimposed with the identified secondary structure (20 helices and 6 sheets) is shown in Figure 6a.

In the case of the subset of 284 good (not necessarily best) particles the reconstruction converged to a map at 9.2Å resolution (Figure 6b) in which 15 of the 21 helices and 4 of 7 sheets (≥ 2 strands) were correctly detected in the P3A using *SSEHunter*. Compared with the map from the best particle set (Figure 6a), most of the helices were positioned correctly, but the quality of the density map was noticeably reduced as also reflected in the map resolution.

In the case of the subset of 284 particles from all the boxed out particles the reconstruction converged to a lower resolution of about 10.6Å resolution (Figure 6c), which was not high enough to identify any of the expected secondary structure elements reliably.

In the case of the subset of 284 bad particles, which were not included to generate the map shown in Figure 4, an approximately 52Å resolution map was built from scratch by exploiting

the fact that these images nevertheless contained some low-resolution information. This map did not provide any significant structure details (Figure 6d; all P3A subunits in this section were segmented by a known P3A mask). One may only see that the map has a double-shelled capsid and that the bumpy regions on the outer capsid display the location of the trimers.

4. Discussion

4.1 Multi-path SA is a new algorithm for single particle 2-D alignment

Although conventional SA was previously applied in cryo-EM data (Ogura and Sato, 2006), it differed from our current algorithm: (1) The previous implementation only produced an initial model; (2) it accommodated no more than 200 particle images; (3) boxed-out particle images had very limited size (i.e. smaller than 100 x 100 pixels); (4) the running speeds were slow. Obviously, the conventional SA has to be improved to make it more practical for a large data set containing high resolution information. Our multi-path SA, determines each particle image's alignment independently by searching for the cross common lines in Fourier space. In this way the CPU demand is a linear function of the number of particle images. The search speed is determined by what resolution data is used instead of being limited by the image size; the alignment of a particle can be accurately determined using only the lower resolution data.

Our staged multi-path SA, with the quasi-random path selection and cross-path update scheme, collectively and globally aligns a particle image using a small number of intelligently evolving paths. This is different from other multi-start methods that restart themselves to perform increasingly localized, successive searches (Dixon and Szegő, 1975; Törn, 1978), or other consecutive-path SA methods that jump out of local minima by initiating a new path with an increase in temperature while keeping track of the best local minimum found (Lech and Hua, 1991). Our multi-path SA algorithm is different from running the conventional SA multiple times from the same model but with different initial velocities (Brunger et al., 1998b; Rice et al., 1998). And it is different from other algorithms that start with multiple starting-points and perform grid searches over the entire orientation space with multiple random starting angles and then use subsequent local optimization refinement (Grigorieff, 2007).

4.2 Low resolution region data are useful

We saw that the very low resolution region data can be used to approximate the size and shape of the RDV (Figure 3a(0)) even without knowing any orientation information. In the case of icosahedral particles, generally there are strong spherical harmonic peaks at the low resolution regions as seen in solution X-ray scattering experiments (Saad et al., 2001; Thuman-Commike et al., 1999). In RDV, the double shell organization might contribute such strong scattering at low resolutions. Low resolution region data provides useful information for the characteristic features of the particle. We used very low resolution data ($\leq 64 \text{ \AA}$) for global searches in the first three iterations (the outer loop in Figure 1a) to refine the initial model to an approximately $33 \text{ \AA}$ solution (Figures 3a(1–3), 3b(1–3)). When we search a particle's alignment parameters based on the low resolution models from the preceding iteration, high resolution and noisy data would obviously be useless and may even be harmful.

Low resolution region data also helps the multi-path SA algorithm more effectively to determine alignment parameters of a particle image. The algorithm's efficiency depends on the landscape of the objective function. A landscape that is generally smooth with some mountains and valleys is advantageous for finding an approximation to the global minimum (i.e. the true solution), as there are comparatively few local minima; this situation corresponds to only including low resolution data in the objective function. If the landscape is comparatively flat on the global scale but oscillating on the local scale, which corresponds to the case in which abundant amounts of high resolution data are included in the objective function, it becomes

difficult for the program to find the global minimum because as soon as the algorithm jumps out of a local minimum, there is a high probability of landing in or near another nearby local minimum. During the global coarse search, we thus only use low resolution data to ensure that the landscape of the objective function varies slowly but is not entirely flat. The global minimum sits in the bottom of a big valley instead of being in a small hole. A side benefit of this is that the amount of data included in computation is small and speeds up the search.

Low resolution region data can be used to accurately determine the center and orientation of a raw particle image. With the inclusion of all of the data up to Nyquist frequency for building a map, the actually achieved resolutions as measured by the FSC curves are in general significantly better than the resolution limit of the data included in orientation and center search. In the fourth iteration of the global search (outer loop in Figure 1a) for orientation and center, only data up to 38Å was used. The corresponding reconstruction reached approximately 12.5Å from these aligned particle images (Figure 3a(4), 3b(4)). For the 5th iteration, the refinement using data up to 27Å resulted in a map exhibiting feature of ~8Å (Figure 3a(5), 3b(5)). When the reference model quality reached a critical level of resolution (e.g. 33Å in the third iteration), the achievable resolution in the reconstruction could be dramatically improved well beyond the resolution range of data included in orientation and center search. This is analogous to the phenomenon of phase transition in physics (Smith and Ness, 1987). The alignment parameters of most of the raw particles would not change much after the “phase transition” (section 3.2.2). It looks like 33Å resolution data is enough to determine the RDV structure correctly. After the “phase transition”, local alignment refinement should take place within one or two more iterations of the global alignment search on 3-D refinement.

4.3 Discriminating a solution and identifying a particle's quality

For discriminating a solution, the consistency criterion performs a consistency check instead of setting a residual threshold for all the raw particles. Different particles may have different noise levels, defocus values, or other factors which may influence the final residual. There is no uniform threshold to determine whether a solution is right or wrong. There is also no guarantee that the alignment parameter solution resulting from the consistency criterion is the true one. Nevertheless, after applying the consistency criterion, the tests on both simulated and real data show that the result is very accurate.

A particle image is treated as “bad” if the set of M solutions from the inner loop of Figure 1a fails to pass the consistency criterion; this simply means that the algorithm did not find the global minimum of the corresponding objective function, which measures similarity in the common lines. Compared with the landscape of the objective function of a good particle, a bad particle will yield a landscape with a comparatively shallow global minimum. As a result, it is harder for the algorithm to find consistent solutions for a bad particle. It is also possible that the criterion may occasionally fail for some good particles for one reason or another. The results from both the 284 good particles and the 284 bad ones in section 3.4 confirm that the consistency criterion can identify a good or bad particle effectively.

The reconstructions of different best particle datasets (section 3.3) showed that a few very high quality particle images could reconstruct a good density map from scratch if these high quality particle images can be identified. The question remains, “How can we identify these high quality particle images?” By high quality, it is meant for particle images that they would contribute to coherent averaging, leading to a high resolution structure. The quality of a particle is affected by the image's SNR at high resolution, conformational homogeneity, and icosahedral symmetry preservation. Though the current electron microscopes can generate relatively high resolution micrographs, a computational method is still needed to select the highest quality particle images.

When we analyzed the best 284 particle images selected by the empirical method of section 2.5, it was found that these particles came from just 18 micrographs. Figure 7 shows scatter plots of logarithmic peak values of SNR vs. $1/d^2$ where d is resolution. The SNR, which is generated by the *ctfit* program of EMAN, is defined as the fitted CTF multiplied by the Gaussian approximation of the fall-off function and then divided by the fitted background (Ludtke et al., 1999). The SNRs of the particles from the 18 best micrographs were plotted as red symbols, and the SNRs of the particles from the remaining 82 micrographs were plotted in green symbols. Observing the SNRs of these two sets of the micrographs, we saw that the red symbols were generally distributed at the upper end of the range of SNRs. The corresponding linear fits were also shown in Figure 7 (red line for the 18 best, green for the remaining). The two fitted lines show that the B-factor based on SNR using the definition of Rosenthal and Henderson (Rosenthal et al., 2003) for the 18 micrographs containing the best 284 particles is $404.5\text{\AA}^2$, as opposed to a value of $501\text{\AA}^2$ for the particles in the remaining micrographs. This observation indicates that these selected best particles have better quality as a whole (the average SNR of the 18 micrographs is about 3 times that of the remaining micrographs at $7\text{\AA}$ resolution). This supports the notion that SNR of particle images is a good indicator for the likelihood of achieving a high resolution map (Saad et al., 2001).

4.4 Model-bias-free reconstruction

The model bias problem is often a concern in cryo-EM image processing (Grigorieff, 2000). Different initial models could lead to significantly different final reconstructions. To arrive at unbiased models, different approaches have been taken. For instance, the *multi-refine* program has been used in EMAN, where several initial models of different conformations or multiple models with varying amount of added noise would give rise to the same final map (Brink et al., 2004; Chen et al., 2006). Alternatively, Yan and coworkers (Yan et al., 2007) proposed the Random Model (RM) method and showed that random models can be effective starting models. They split whole dataset into two halves to generate two independent models (starting from two individual random initial models) which have to agree with each other. In our approach, we simply started with several random initial models, and a converged final map was produced after several iterations of global searches followed by local refinements.

Our approach builds the random initial model using particles with randomly assigned orientations, which results in a spherical shape model with the approximate size of the corresponding virus particle. The features of the random model are quite different from the actual virus and thus should not bias the final solution. It thus allows us to assess reconstructions without the model bias concern while using different numbers of particle images of similar quality (Figure 5), or the same number of particle images with variable qualities (Figure 6). All reconstructions (except the one from the bad particles) converged to correct structures, even though the maps have different resolutions and different identifiable structural features. We also tested some other virus samples using random initial models. All of these tests led to convergent maps with similar structural features (unpublished). For the user-supplied models, which were even significantly different from the real structure in size or shape, the algorithm was still able to yield the converged maps. These observations above strongly suggest that there is no model bias problem in our approach.

4.5 Relation of number of particles to resolution in single particle reconstruction

Crowther, DeRosier and Klug (Crowther et al., 1970a) pointed out the minimum number of projection views required to get a certain resolution map, based on geometrical considerations, by implicitly assuming noise-free and evenly spaced data. Subsequently, estimates of the number of particles needed for a reconstruction in the presence of noise have been proposed by taking into account the average SNR of the data (Glaeser, 1999; Rosenthal et al., 2003). Under the proposed theoretical formulation of Rosenthal and Henderson, we would expect to

see a linear relationship in the plot of $\log(Nd)$ vs $1/d^2$, where N is the number of particles and d is the resolution.

Borgina and co-workers were the first to show that even using nine best particle images, α -helical features could be visualized from the reconstruction of the pyruvate dehydrogenase icosahedral core (Borgina et al., 2004). It is not known if the resolutions of the reconstructed maps can be measured reliably from FSC curves with a small number of particles in their study. Furthermore, it is not clear if a correct reconstruction can be obtained when alignment parameters of these particles are determined from scratch and not based on the ones determined as part of a larger dataset.

In our study, we randomly selected different numbers of particles from the 4,200 good particles that were used to generate the map shown in Figure 4. The green symbols of Figure 8 show the relationship between the numbers of good particle images and the resolutions of the corresponding reconstructed maps as defined in the theoretical formula (Rosenthal et al., 2003). This data does demonstrate a linear relationship (green line in Figure 8) with a computed B_{overall} factor of $\sim 826 \text{ \AA}^2$. However, for the best quality particle datasets, which were partially shown in Figure 5, the relationship of the particle numbers and resolutions (red symbols of Figure 8) does not exhibit the same linear relationship as the one in the green symbols of Figure 8. Although no overall linear relationship exists in the best particle datasets, two segmented linear relationships can still be found. The lower part of the red symbols can be fitted to a straight line (red line in figure 8) with $B_{\text{overall}}=425 \text{ \AA}^2$; and the data points in the upper part (from particle number 379 to 972) can be fitted to a line with the $B_{\text{overall}}=625 \text{ \AA}^2$ (cyan line of Figure 8).

The B_{overall} factor can be contributed from three independent sources: computational (B_{comp}), experimental (B_{expt}) and biological (B_{bio}). We independently determined the experimental B factor, B_{expt} , based on the SNRs of micrographs (Figure 7). Its value is influenced by the Fourier amplitude fall-off due to the coherence of the beam, specimen drift, and the detector's modulation transfer function. The accuracy of its estimate depends on the fitting of the parameters as defined previously (Saad et al., 2001). The biological B factor, B_{bio} , is affected by the symmetry preservation of a particle and the conformational homogeneity among different particles. The analysis of Figure 8 shows different values of B_{overall} in various subsets of particle images (Figure 5) which could be due to differences in imaging or particle quality. Interestingly, these differences in B_{overall} are in agreement with our selection of some particle images as better than the others (Figure 7). Furthermore, the B_{overall} deduced in Figure 8 for the small numbers (< 200) of the best particles is almost the same value as the B_{expt} (Figure 7) of the micrographs that contained the best particles. This suggests that the computational and the biological B factors are negligible for the small subset (< 200) of best particles in this case study. Therefore, increasing the quality of particle images (i.e. smaller B_{expt}) is the way to improve the resolution or to reduce the number of particles needed to achieving a certain resolution. However, for some other specimens, one may additionally need to handle the conformational homogeneity (Chen et al., 2006) and/or symmetry preservation computationally. Our study supports the notion made previously (Borgina et al., 2004) that only a small number of high quality particle images are enough to detect secondary structure elements, but the exact number of particle images may differ from case to case as demonstrated in the current investigation. The minimum number of particles could approach the theoretical limit, $\pi D/d$ (Crowther et al., 1970a), if we have very small B_{overall} factor. However, we were unable to get a converged map with fewer than 41 best particles. It seems that there is a lower bound on the number of particles necessary to generate a correct reconstruction from random initial model, as an inherent feature of the statistics in a particular set of image data.

4.6 Refining more parameters and extending to non-icosahedral symmetry

Currently, the multi-path SA is only used for searching the center and orientation of a raw particle image simultaneously; however, it can easily accommodate more than the five parameters. The algorithm is essentially a global optimizer. The CPU demand of this algorithm will not increase exponentially as the number of refining parameters increases. Other parameters such as defocus, scaling, astigmatism, etc. could be included in the future.

The program currently works only for icosahedral symmetry. Theoretically, lower symmetry particles may be accommodated by increasing the number of reference projections in cross common line search. This approach can potentially be applied to solve other lower symmetry and asymmetrical structures.

5. Conclusions

The new multi-path SA algorithm is a Monte Carlo type of global optimization algorithm. It approaches a global minimum of an objective function from multiple paths with three specially designed annealing stages and a special path update scheme. The image processing was converted to a global optimization problem based on cross common line search. The new algorithm greatly reduced the computational cost compared with exhaustive cross common line search and made it practical to use a single step of cross common line search to replace the three step process of conventional common line strategy.

The consistency criterion was developed to discriminate whether a solution is right or wrong. The increased confidence in the solutions accelerates the speed of convergence and yields surprisingly accurate center and orientation information, even when using very low resolution data. The reconstruction procedure is model bias free. Using only the selected best tens to hundreds of 2-D particle images without a user supplied initial model, we are able to see the secondary structure from the reconstructed 3-D maps at subnanometer resolution. This new algorithm could be extended to non-icosahedral symmetry and possibly extend the capability to refine other experimental parameters such as defocus of individual particle image.

The 3-D maps and segmented sub-units shown in figure 5 have been deposited to EBI under accession numbers _____.

The software is freely available as part of the EMAN software distribution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH (R01GM070557 and P41RR02250) and the Agouron Institute. We thank Dr. Richard Henderson at MRC for allowing us to use his scanner and the technical assistance of Dr. G McMullan. We thank Jonathan Respress, Hari Damodaran, Kurt Welgehausen and Deepy Mann for editorial assistance in preparation of the manuscript.

References

- Baker ML, Ju T, Chiu W. Identification of secondary structure elements in intermediate-resolution density maps. *Structure* 2007;15:7–19. [PubMed: 17223528]
- Borgina MJ, Shi D, Zhang P, Milne JL. Visualization of α -helical features in a density map constructed using 9 molecular images of the 1.8 MDa icosahedral core of Pyruvate Dehydrogenase. *J Struct Biol* 2004;147:136–145. [PubMed: 15193642]

- Brink J, Ludtke SJ, Kong Y, Wakil SJ, Ma J, Chiu W. Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. *Structure (Camb)* 2004;12:185–91. [PubMed: 14962379]
- Brunger AT, Adams PD, Rice LM. Recent developments for the efficient crystallographic refinement of macromolecular structures. *Curr Opin Struct Biol* 1998a;8:606–11. [PubMed: 9818265]
- Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 1998b;54:905–921. [PubMed: 9757107]
- Chen DH, Song JL, Chuang DT, Chiu W, Ludtke SJ. An Expanded Conformation of Single-Ring GroEL-GroES Complex Encapsulates an 86 kDa Substrate. *Structure* 2006;14:1711–22. [PubMed: 17098196]
- Crowther RA, DeRosier DJ, Klug A. The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proc Roy Soc Lond A* 1970a;317:319–340.
- Crowther RA, Amos LA, Finch JT, DeRosier DJ, Klug A. Three dimensional reconstructions of spherical viruses by Fourier synthesis from electron micrographs. *Nature* 1970b;226:421–5. [PubMed: 4314822]
- DeRosier DL, Klug A. Reconstruction of three-dimensional structures from electron micrographs. *Nature* 1968;217:130–134.
- Dixon, L.; Szegő, G. Towards global optimization. North-Holland; Amsterdam: 1975.
- Fuller SD. The T=4 envelope of Sindbis virus is organized by interactions with a complementary T=3 capsid. *Cell* 1987;48:923–34. [PubMed: 3829124]
- Fuller SD, Butcher SJ, Cheng RH, Baker TS. Three-dimensional reconstruction of icosahedral particles--the uncommon line. *J Struct Biol* 1996;116:48–55. [PubMed: 8742722]
- Glaeser RM. Review: electron crystallography: present excitement, a nod to the past, anticipating the future. *J Struct Biol* 1999;128:3–14. [PubMed: 10600552]
- Grigorieff N. Resolution measurement in structures derived from single particles. *Acta Crystallogr D Biol Crystallogr* 2000;56(Pt 10):1270–7. [PubMed: 10998623]
- Grigorieff N. FREALIGN: high-resolution refinement of single particle structures. *J Struct Biol* 2007;157:117–25. [PubMed: 16828314]
- Hamilton WR. On a new species of imaginary quantities connected with a theory of quaternions. *Proc Royal Irish Acad* 1944;2:424–434.
- Hamilton WR. On quaternions. *Proc Royal Irish Acad* 1947;3:1–16.
- Harauz G, van Heel M. Exact filters for general geometry three dimensional reconstruction. *Optik* 1986;73:146–156.
- Hartmann F, Rieger AKH. New optimization algorithms in physics. 2004
- Henderson R, Cattermole D, McMullan G, Scotcher S, Fordham M, Amos WB, Faruqi AR. Digitisation of electron microscope films: six useful tests applied to three film scanners. *Ultramicroscopy* 2007;107:73–80. [PubMed: 16872749]
- Jiang, W.; Chiu, W. Cryoelectron microscopy of icosahedral virus particles. In: Kuo, J., editor. *Methods in Molecular Biology, Electron Microscopy Methods and Protocols*. Humana Press; Totowa, NJ: 2006. p. 345–363.
- Jiang W, Baker ML, Ludtke SJ, Chiu W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J Mol Biol* 2001;308:1033–44. [PubMed: 11352589]
- Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science* 1983;220:671–680. [PubMed: 17813860]
- Kuffner, JJ. Effective sampling and distance metrics for 3D rigid body path planning. *Proc IEEE Int'l Conf on Robotics and Automation*; New Orleans. 2004. p. 3993–3998.
- Laarhoven, P.; EHL, A. *Simulated Annealing: Theory and Application*. Dordrecht: Reidel Publ Co; 1987.
- Lech, M.; Hua, Y. Vector quantization of images using neural networks and simulated annealing. In: Juang, BH., et al., editors. *Proc IEEE Neural Networks for Signal Processing*; IEEE Service Center, Piscataway, NJ. 1991. p. 552–561.
- Ludtke SJ, Baldwin PR, Chiu W. EMAN: Semi-automated software for high resolution single particle reconstructions. *J Struct Biol* 1999;128:82–97. [PubMed: 10600563]

- Metropolis N, Rosenbluth AW, Rosenbluth M, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys* 1953;21:1087–1092.
- Nakagawa A, Miyazaki N, Taka J, Naitow H, Ogawa A, Fujimoto Z, Mizuno H, Higashi T, Watanabe Y, Omura T, Cheng RH, Tsukihara T. The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. *Structure (Camb)* 2003;11:1227–38. [PubMed: 14527391]
- Nelder JA, Mead R. A simplex method for function minimization. *Computer J* 1965;7:308–313.
- Ogura T, Sato C. A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. *J Struct Biol* 2006;156:371–86. [PubMed: 16949302]
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 2004;25:1605–12. [PubMed: 15264254]
- Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical recipes in C++. Cambridge University Press; NY: 2002.
- Rice LM, Shamoo Y, Brunger AT. Phase improvement by multi-start simulated annealing refinement and structure factor averaging. *J Appl Cryst* 1998;31:798–805.
- Rosenthal PB, Crowther RA, Henderson R. An objective criterion for resolution assessment in single-particle electron microscopy. *J Mol Biol* 2003;333:743–745.
- Saad A, Ludtke SJ, Jakana J, Rixon FJ, Tsuruta H, Chiu W. Fourier amplitude decay of electron cryomicroscopic images of single particles and effects on structure determination. *J Struct Biol* 2001;133:32–42. [PubMed: 11356062]
- Smith, JM.; Ness, HCV. Introduction to Chemical Engineering Thermodynamics. McGraw-Hill; NY: 1987.
- Thuman-Commike PA, Tsuruta H, Greene B, Prevelige PE, King J, Chiu W. Solution x-ray scattering based estimation of electron cryomicroscopy imaging parameters for reconstruction of virus particles. *Biophys J* 1999;76:2249–61. [PubMed: 10096920]
- Törn, AA. A search clustering approach to global optimization. In: Dixon, L.; Szegö, G., editors. Towards global optimization. North-Holland, NY: 1978. p. 49-62.
- Yan X, Dryden KA, Tang J, Baker TS. Ab initio random model method facilitates 3D reconstruction of icosahedral particles. *J Struct Biol* 2007;157:211–225. [PubMed: 16979906]
- Zhou ZH, Chiu W, Haskell K, Spears H Jr, Jakana J, Rixon FJ, Scott LR. Refinement of herpesvirus B-capsid structure on parallel supercomputers. *Biophys J* 1998;74:576–88. [PubMed: 9449358]
- Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, Dong G, Lu G, Chiu W. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nat Struct Biol* 2001;8:868–73. [PubMed: 11573092]

Figure 1.

Diagrams for the 3-D refinement procedure and the multi-path simulated annealing algorithm. The novel features are shown in the shaded boxes. (a) The refinement procedure includes three loops and one step to build an initial random model. The outer loop is used to iteratively refine a 3-D map. At the beginning of the outer loop, the data resolution range used in common line search can be chosen; and also the multipath SA algorithm can set to execute either global or local search. The inner loop is used to generate M candidate solutions which will be passed to the middle loop for consistency check. Each of the M solutions is produced by multi-path SA. The consistency measure identifies whether the raw particle image is good, in which case it is used in the corresponding iteration of 3-D refinement. The middle loop goes through each of the N raw particle images and then generates N' ($<N$) particles with consistent solutions. The N' particles are used for building a 3-D map using the *make3d* program of EMAN. (b) The multi-path SA algorithm runs in two modes: *global alignment search* and *local alignment refinement*. The loop goes through three annealing stages based on annealing temperature with a different path update scheme in the third stage for updating the current solution list. The current solution list is sorted by residual after each update. Paths are quasi-randomly picked, with some preference given to the worst residual paths in the first two stages. Once a path is picked, a trial solution is derived by randomly varying its centers and orientations. The trial residual is calculated based on the trial common lines. The acceptance of a trial solution is governed by the Metropolis criterion. An accepted trial solution replaces the corresponding solution in the current solution list in first two stages, or the worst solution in the list in the last stage. The best solution list always keeps track of the best solution found so far in each individual path. The final solution will be the best element in the best solution list.

Figure 2. Typical RDV raw particle images recorded at different defoci. (a) $\Delta z = 0.322\mu\text{m}$, (b) $\Delta z = 0.414\mu\text{m}$, (c) $\Delta z = 0.680\mu\text{m}$, (d) $\Delta z = 0.919\mu\text{m}$, (e) $\Delta z = 1.594\mu\text{m}$, (f) $\Delta z = 2.286\mu\text{m}$.

Figure 3.

The center slices (a) and Fourier Shell Correlation (FSC) curves (b) of the 3-D map refined from the initial random model to 7th iteration with 4,865 raw particle images. Sub-Figures 3a (0–7) represent the center slices from initial random model (0) to the map of the 7th iteration (7). In sub-figure 3b, FSC curves 0 to 7 were measured against the known X-ray structure (PDB ID: 1UF2) with respect to the initial random model to the map of the 7th iteration and the FSC curve labeled “7 half-half” was generated by measuring the correlation of the split half-half data. The Fourier transform data used for 2-D alignment and the corresponding 3-D map’s resolution are listed as follows (# of iteration, FFT data used for refinement, map resolution): (0) (initial, random, ~100 Å); (1) (1st, <136 Å, 47.7 Å); (2) (2nd, <63.5 Å, 47.7 Å); (3) (3rd, <63.5 Å, 32.9 Å); (4) (4th, <38.1 Å, 12.5 Å); (5) (5th, <27.2 Å, 8.2 Å); (6) (6th, 190 - 15.8 Å, 7.6 Å); (7) (7th, 63 - 10.6 Å, 7.3 Å); (7 half-half) (7th, 63 - 10.6 Å, 6.3 Å).

Figure 4.

Final reconstructed map of RDV with 7.3Å resolution from 4,200 good particles out of 4,865 raw particle images. (a) Entire capsid; (b) Stereo view of the P8 subunit with superimposing X-ray structure; (c-d) The two side views of the P3A subunit with superimposing X-ray structure.

Figure 5.

Structures were built from scratch with different numbers of best particles. The P3A subunits of RDV segmented with a known mask were superimposed with X-ray structures (PDB ID: 1UF2). Legend at the bottom of each sub-figure shows number of best particles used to build the 3-D map, the corresponding map resolution, and number of helices and sheets identified by SSEHunter.

Figure 6.

The P3A subunits superimposed with the results of SSEHunter from the maps built with various subsets of 284 particle images. (a) Best 284 particles, 7.9Å resolution; (b) 284 randomly picked from the 4,200 good particles, 9.2Å resolution; (c) 284 randomly picked from the whole dataset, 10.6Å resolution; (d) 284 randomly picked from the bad (rejected) 665 particles in Figure 4, 52.4Å resolution. The P3A subunits are masked out with a known P3A mask.

Figure 7.

Log(SNR) of micrographs as a function of $1/d^2$, where d is resolution, for the 18 micrographs containing the best 284 particle images and the 82 remaining micrographs. Red symbols correspond to the 18 best micrographs, green to the 82 remaining micrographs. Linear fits are shown in red line ($B = 404.5 \text{ \AA}^2$) and green line ($B = 501 \text{ \AA}^2$) for the 18 and 82 micrographs, respectively using the formulation previously proposed (Rosenthal et al., 2003). Here, the symbols (“x” or “+”) represent the peaks of the SNR curves which are generated by *ctfit* program of EMAN.

Figure 8.

Relationship between the number of particles and the maximum resolution achieved from the corresponding dataset using the previously proposed formula (Rosenthal et al., 2003). Each data point in green is computed from different number of particles randomly picked from the pool of 4200 good particles. The data points in red were the best particle datasets mentioned in section 3.3. Along any one of the two color data points from bottom to top, each of the points represents the following number of particles: 41, 62, 81, 101, 128, 147, 181, 207, 226, 284, 379, 446, 512, 648, 843, 972, 1470, 1755, 4200. The green points can be fit as a straight line with a B factor $826\text{\AA}^2$. The red data points can be partially fit into two lines with B factors $425\text{\AA}^2$ (red line) and $625\text{\AA}^2$ (cyan line) respectively.

The average center and orientation errors of the simulated images with respect to different noise levels. The simulated images were projected from a RDV model which was blurred from the X-ray structure, and CTF parameters extracted from some typical cryo-EM micrographs (Figure 2) were applied. The centers of the simulated images (size: 400X400 pixels) were shifted by up to 40 pixels. Different levels of white noise were added to the images in units of variance of the image (σ_{img}).

Table 1

Noise level (σ_{img})	5	10	15	20	25	30
Avg. center error (pixel)	0.0627	0.1121	0.1676	0.2269	0.2847	0.3586
Avg. orientation error (degree)	0.2300	0.2628	0.3015	0.3607	0.4234	0.4885
Success rate	1	1	0.9994	0.9994	0.9886	0.9302
False positive rate	0	0	0.0006	0.0006	0.0108	0.0685