FindEM—a fast, efficient program for automatic selection of particles from electron micrographs

A.M. Roseman*

Medical Research Council—Laboratory of Molecular Biology, Hills Road, England, Cambridge CB2 2QH, UK

Received 11 July 2003, and in revised form 29 October 2003

Abstract

The FindEM particle picking program was tested on the publicly available keyhole limpet hemocyanin (KLH) dataset, and the results were submitted for the “bakeoff” contest at the recent particle picking workshop (Zhu et al., 2003). Two alternative ways of using the program are demonstrated and the results are compared. The first of these approximates exhaustive projection matching with a full set of expected views, which need to be known. This could correspond to the task of extending a known structure to higher resolution, for which many 1000's of additional images are required. The second procedure illustrates use of multivariate statistical analysis (MSA) to filter a preliminary set of candidate particles containing a high proportion of false particles. This set was generated using the FindEM program to search with one template that crudely represents the expected views. Classification of the resultant set of candidate particles then allows the desired classes to be selected while the rest can be ignored. This approach requires no prior information of the structure and is suitable for the initial investigation of an unknown structure—the class averages indicate the symmetry and oligomeric state of the particles. Potential improvements in speed and accuracy are discussed.

Keywords: Particle picking; Automatic particle selection; Local correlation; Electron microscopy; CryoEM

1. Introduction

The FindEM program (Roseman, 2003) is based on projection matching using local real-space correlation with a series of templates. Real-space algorithms are traditionally relatively slow, but a fast Fourier based implementation of the real-space local correlation function has been introduced which is approximately two orders of magnitude faster for the projection matching application. This algorithm is referred to as the “Fast Local Correlation Function” or FLCF. The templates can completely represent the entire range of expected views of the 3D structure allowed by the resolution band limit, or a smaller set can be chosen to represent the range approximately. Local correlation with templates accurately and sensitively produces correlation maps with peaks indicating particle positions. These peak positions are then subject to a filtering operation in order to reject particles that have low correlation with the templates, or that are overlapping, and to assign the view type (according to which template matches best). An interactive graphical interface shows the selected particles overlaid on the micrograph and can be used to optimise the parameters controlling this filtering, such as correlation threshold and particle size. Once these parameters are chosen, any number of similar images can be automatically processed.

The approach is versatile and can be adapted according to a particular experimental situation. The keyhole limpet hemocyanin (KLH) dataset (available from the AMI group, The Scripps Research Institute, CA, USA, http://ami.scripps.edu/prtl_data/) is used to demonstrate two complementary approaches to particle picking using the FindEM program. KLH is a cylindrically shaped ~8 MD particle, a homo-oligomeric didecamer with D5 point group symmetry (Orlova et al., 1997).

There are two isoforms of KLH, KLH1 and KLH2. The short cylindrically shaped particles are assumed to be KLH1 and longer particles are assumed to be KLH2 or aggregates of KLH1. The short cylindrical particles will be referred to as KLH in this report from now on.
Because the KLH particles are preferentially oriented into end-views (presenting a 5-fold axis perpendicular to the image plane) and side-views (with the 5-fold axis in the plane of the image) two templates accurately describe most of the views, and form the “exhaustive” projection set. (See Figs. 1–3). Side-views selected by this method (Method 1) were submitted for the “bakeoff” contest, and the results compared very favourably with the human experts’ selections (Table 1, and Zhu et al., 2003b, Table 2).

The specific search for the side-views and end-views using an average side-view and end-view image produces very accurate results and allows the different views to be distinguished at the correlation stage. An alternative approach is to use a less specific procedure to generate a larger particle set that contains all possible particles in addition to many false ones. Then classification of this preliminary set of images by multivariate statistical analysis (MSA) (Frank and van Heel, 1982, van Heel and Frank, 1981) will reveal the major populations of views in the images, and also allow the classes containing false positives, or junk, to be easily excluded. In this instance the preliminary set was generated using the FindEM program to calculate correlation maps, using just one template and a low correlation threshold to select candidate particles. The method presented could, however, be used to filter any preliminary set of particles.

The two approaches presented are complementary, and represent two starting points for particle selection.

Fig. 1. Templates. Three different sets of side and end-view templates (left and right, respectively) are shown. Top: the average of 20 images of side and end-views from the first far-from-focus image. Middle: alignment of ~1000 side-views and end-views from a preliminary particle selection from the far-from-focus images. Bottom: side and end-view templates selected from a preliminary particle selection from the near-to-focus images (also ~1000 of each). The templates from 20 images are quite good, but clearly 1000 images give an improvement. The templates from the near-to-focus images do show a little more detail.

Fig. 2. Correlation maps. Top: the CCD first image from the far-from-focus series. Middle: correlation map with the side-view template. Bottom: correlation map with the end-view template. The higher peaks correspond to the location of the correct particle type. Interestingly, the side-view template forms rings of relatively high correlation around end-views (but lower than the correlation peaks with the side-views). There are higher peaks at the centres of these rings when correlating with the end-view template.
The first method, Method 1, represents an exhaustive accurate template search where all the expected views are well defined and represented, providing a high level of discrimination. Strictly, preliminary knowledge of the 3D structure is required for this, though in practice, and as in this example, the required representative views can be otherwise obtained. In the second method, Method 2, no information about the structure is assumed, as an approximate template is used to represent all of the views. The discrimination by correlation is weaker and a broad range of images representing the distribution of particle views is produced. Examination of the class averages produced by MSA reveals the composition of the population of molecules, and new views or conformations can be discovered.

The template based method including an angular search is very general and can be applied to any shapes of particles, since the correlation is equally valid for any shape of template. The contrast transfer function (CTF) of the electron microscope is not explicitly considered in the correlation procedure. The templates calculated from the representative images on the micrographs have CTF effects and the noise model incorporated, as would a 3D model calculated directly from them. If the templates are derived from a 3D model calculated incorporating a CTF correction procedure, then the low-pass filter limit must be set within the first CTF zero crossing (so there is no reversal in contrast between the projections from the model and the images from the micrographs) or the model or projections must be convoluted with the CTF.

2. Methods

The far-from-focus (2 μm defocus) and near-to-focus (0.3 μm defocus) sets of 82 CCD images, 2048 × 2048 pixels each, of KLH molecules were retrieved from the AMI database (KLH Dataset I).
1. Each CCD image was binned by a factor of 4 (resulting in the near-to-focus images in Method 1 from the local correlation algorithm in the FindEM program. The advantage over conventional correlation is that the template and image densities are optimally scaled within the local real-space region around each particle. The templates were rotated through 360° in 4° intervals. The final output correlation map contains the map of the maximum correlation over all orientations.

(B) The peaks in the correlation maps were automatically extracted and filtered according to the following steps (for each image in turn)

1. A minimum value correlation threshold was applied to generate a preliminary list of particle positions. This was done for each of the two templates to give two lists.

2. The highest value in each correlation peak was found and the shoulder values within a user-defined peak radius were excluded, for each list.

3. Where peak positions from the correlation with both templates coincided (within the defined peak radius) then the particle view-type was assigned to the template with the higher correlation.

4. Particles that lie within a user-defined diameter of one another were deselected. This was to reject overlapping particles that are not suitable for 3D structure calculation. The particle diameter can be individually defined for each template. Overlaps between the same particle view-type and different particle view-types were determined.

2. Twenty end and side-views were selected manually (using Ximdisp, Smith (1999)) from the first image in the series, and these were optimally aligned using five iterations of orientation and translational correlation alignment to the running average, to generate a high signal-to-noise template for the end and side-views, using SPIDER (Frank et al., 1996) procedures.

3. Each template was correlated with each CCD image using the local correlation algorithm in the FindEM program. The advantage over conventional correlation is that the template and image densities are optimally scaled within the local real-space region around each particle. The templates were rotated through 360° in 4° intervals. The final output correlation map contains the map of the maximum correlation over all orientations.

(B) The peaks in the correlation maps were automatically extracted and filtered according to the following steps (for each image in turn)

1. A minimum value correlation threshold was applied to generate a preliminary list of particle positions. This was done for each of the two templates to give two lists.

2. The highest value in each correlation peak was found and the shoulder values within a user-defined peak radius were excluded, for each list.

3. Where peak positions from the correlation with both templates coincided (within the defined peak radius) then the particle view-type was assigned to the template with the higher correlation.

4. Particles that lie within a user-defined diameter of one another were deselected. This was to reject overlapping particles that are not suitable for 3D structure calculation. The particle diameter can be individually defined for each template. Overlaps between the same particle view-type and different particle view-types were determined.

Table 1
Comparison of FindEM results with “expert” manual pickers results as submitted to the particle picking “bakeoff”

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
<th>Mouche (Manual)</th>
<th>Haas (Manual)</th>
<th>Roseman (FindEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouche (1042)</td>
<td>—</td>
<td>11.7</td>
<td>2.4</td>
<td>234 (18.3%)</td>
</tr>
<tr>
<td>Haas (944)</td>
<td>2.3</td>
<td>—</td>
<td>16.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Roseman/FindEM (1219)</td>
<td>11.7</td>
<td>23.9</td>
<td>—</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>2.4</td>
<td></td>
<td>2.4</td>
<td>1.5</td>
</tr>
</tbody>
</table>

False positive rate and false negative rates are reported as percentages (FNR/FPR). This table is extracted from the table summarising results submitted to the “bakeoff” particle selection contest (Zhu et al., 2003b, Table 2). The list of particles generated using FindEM (Method 1) is in very good agreement with those chosen manually by the two human experts, Mouche (Mouche et al., 1999; Zhu et al., 2003a) and Haas. FindEM included 97.6% of those chosen by Haas and 98.5% of those chosen by Mouche. The false positive rate of 17% and 24%, respectively indicates that some extra particles were found. In the Haas vs Mouche comparison the ~10% FNR is because Mouche found about 100 more particles than Haas did, and this also explains the higher FPR of FindEM vs Haas, than vs Mouche. This might be partly because Mouche deliberately included tipped side-views, while Haas did not. The total number of particles picked is in brackets, due to the application of standardised edge elimination criteria for the “bakeoff.”

There are two stages to the procedure, (A), a template based correlation step is used to find occurrences of the particles in the images, and (B), a filtering step is used to reject particles with low correlations or that are too close to one another.

2.1. Method 1. Correlation with two templates, representing the end and side-views

(A) Image preparation and correlation

1. Each CCD image was binned by a factor of 4 (resulting sampling 8.8 Å/pixel) and band-pass filtered in the range 30–2000 Å.

Table 2
Comparison of FindEM results using different methods

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
<th>Method 1. (FindEM)</th>
<th>Method 1’—FindEM (1154)</th>
<th>Method 2—FindEM (1232)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1—FindEM (1282)</td>
<td>—</td>
<td>136 (10.6%)</td>
<td>234 (18.3%)</td>
<td>284 (23.1%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>264 (22.8%)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Results are reported as numbers of particles and percentages. Method 1 is assumed as “truth” for this comparison (FNR/FPR). Method 1 is the far-from-focus dataset analysis using two templates. Method 1’ analyses the near-to-focus images by the same method. The FNR of 20% indicates that 80% of the particles, as compared with Method 1, were found. The FPR of ~10% may be a small nuisance compared to the alternative of collecting all images as defocus pairs. If this is avoided every image collected can have an optimal electron dose, and the data collection rate is doubled. Method 2 is the procedure using classification to select side-view images from the set generated using one side-view template with a low correlation threshold. The FPR and FNR (~20% each) indicates reasonable agreement between the two procedures, given the limitations discussed about absolute objective evaluation of the particle picks, and the ~20% FPR of Method 1 vs the expert picks. An explanation for some of the false positives in Method 2 is that more tipped side-view classes were included, and that some of these images may have been below the correlation threshold applied in Method 1. The total number of particles picked is in brackets. The same correlation thresholds were used to select the particles from the near-to-focus images in Method 1 as for Method 1. It is possible that these could have been optimised and better results obtained. On inspection of some of the near-to-focus images it is apparent that the particles have suffered less radiation damage than in the second image taken (the far-from-focus one) and are better preserved. Consequently, the particle picking program has the potential to be more accurate on the near-to-focus images, even though the contrast is lower.
This procedure was re-iterated once (starting at step A3), by generating higher signal-to-noise templates from the picked particles, which were then used to improve the correlation maps.

The user-defined parameters required in stage B were optimised using the FindEM interactive graphical interface written in Tcl-Tk, which allows one to adjust the parameters and see the resultant particle selection overlaid on the image. About five images were examined using the graphical interface, and the parameters chosen were used to automatically select the particles from all 82 images. The actual parameters used for the results presented are: sampling 8.8 Å, particle diameters 300 Å, correlation coefficient thresholds 0.32, peak width 150 Å. The particle selection was not very sensitive to small adjustments in these parameters.

2.2. Method 2. Correlation with one template (side-view) and MSA classification

For Method 2, where a selection of candidate particles is submitted for MSA classification, part A is the same as for Method 1, but the correlation with only one template (in this example, the side-view) is required. For part B (extracting peaks from the correlation maps) some parameters were changed. The correlation threshold was dropped to 0.25. The diameter parameter was reduced to 150 Å because some of the false positives included (deliberately) by dropping the correlation threshold were close to real particles and it was not desirable to exclude these as overlapping particles. The selected images of particles were extracted and brought into a common orientational alignment using the angle found by the correlation search, which includes a complete in-plane orientation search. The oriented particle images were subjected to MSA classification as implemented in the IMAGIC program (van Heel et al., 1996). Class averages resembling side-views were selected by eye. Python scripts were used to manipulate text files to trace particle numbers between lists in order to compare the results of the two methods.

3. Results and discussion

3.1. Method 1—multiple template matching

The side and end-view templates used to correlate with the images are shown in Fig. 1, and the corresponding correlation maps with the first image in the series are shown in Fig. 2. The particles selected are indicated in Fig. 3, which shows a screen display from the graphical interface to the program.

The automated procedure selected 1282 side-views from the 82 CCD images, and the particle selection compares quite favourably with that of the “expert” human pickers, see Table 1. The program found >97% of the particles selected by the experts, and about 20% additional particles or false positives. A 3D structure was calculated from the images to a resolution of 25 Å (Fig. 4), using a projection matching procedure implemented in SPIDER (essentially as described in Roseman et al., 1996). The calculated 3D structure closely resembles the structure displayed on the AMI database web site and the low resolution views in the published report on the structure of the KLH1 didecamer (Orlova et al., 1997).

The limitation in the achievable accuracy of particle picking is illustrated in Fig. 3. In this image a side-view that is contaminated with a small bit of ice or “junk” is included in one case (particle 2), but not another (particle 3). Depending on the level of contamination and resultant distortion of the side-view, side-views that have minor ice contamination are sometimes selected, and sometimes not. Though it may be possible to include another “contamination detection” test, these relatively few cases are probably best dealt with in the 3D refinement stage. Similarly, damaged particles can only be reliably identified once a 3D structure is known. Contaminated or otherwise distorted particles are genuine particle views, which could arguably form part of a preliminary particle set. However, they will ultimately be rejected from the dataset and therefore it is not useful to deliberate over these marginal cases.

Method 1 was repeated using the near-to-focus (0.3 µm) images, producing a similar result. About 80% of the particles chosen were also contained in the set selected from the far-from-focus image set. This result is presented as Method 1′ in Table 2. The templates used in the correlations are shown in Fig. 1.

3.2. Method 2—MSA classification of large set of candidates

The class averages obtained from classification of the ~3000 images in the preliminary particle set is shown in Fig. 5. The particles were brought into common in-plane orientation alignment before MSA, using the relative angles for which the highest correlation coefficient with the template was found. A range of class averages and particle types are evident. Different categories of side-views, end-views, half molecules, and other non side-view classes can be identified. Classes of side-view with different orientation, or that are tipped (with the D5 symmetry axis not in the image plane) can be distinguished. Tipped classes of end-view, distorted side-views, junk are also present. The side-view classes selected for the comparison with Method 1 are listed in the legend to Fig. 5. Classes 2, 8 and 14 are typical of good side-view classes; 15 is a tipped side-view class; 1, 4, 6, 7 are good end-view classes; 13 looks like a slightly tipped end-view class. Half molecules have aligned to...
either the top or bottom side of the side-view template, as seen in classes 31 and 37, respectively. Class 48 is an example of side-views recognised within longer particles, which would be rejected.

The selected particles were in agreement within a level of 80% when compared to Method 1 (Table 2), illustrating that MSA and classification to sort out image views from a larger selection of candidates can be a powerful method for analysing images even when little is known about the structure. Since the selection for expected views is weaker in Method 2 (than Method 1), it is more likely that unexpected views could be observed, if present, and therefore this procedure could be a quick way to select particles for a new structure. However, a
caution is that it could be a disadvantage not to engage in manually selecting the first few thousand particles, in order to learn about the specimen.

Method 2 was also repeated without using the relative orientations for the particles found by the FindEM correlation program. Instead the randomly oriented particles were brought into optimum relative alignment using an iterative correlation scheme. Cycles of orientation and cross-correlation of the extracted images to the running average (three cycles) were repeated until the average converged (five iterations). This procedure allows similar views to line up in a consistent fashion relative to the common average. The particle list obtained was very similar to Method 2, showing that this method can be applied to virtually any preliminary set of particles, and that preliminary orientation alignment to a distinctly asymmetric template is not a prerequisite.

3.3. Potential improvements

The algorithm used in the FindEM program is based on using local correlation to find correlation matches to given template images within a larger electron micrograph image. The local correlation function gives a map of normalised correlation coefficients, with optimised scaling of the template and target image in the local region of overlap, and is equivalent to a least squares optimisation. Improvement is possible in two areas though. In more demanding applications such as smaller particles, those with a greater distribution of distinct views, imaged closer-to-focus, and at higher voltage, more templates will be required to effect as accurate a search. Then, if many 1000’s of particles are required, some increase in speed will be necessary in order to complete the task in a reasonable time. Another possible
improvement is to introduce a statistical weighting (on a real-space pixel-wise basis) into the measure-of-fit, in order to weight the data in the images or projections according to their reliability.

3.3.1. Speed improvement

The time taken to find these \( \sim 1000 \) particles from 82 images was 56 minutes per template, using a DEC alpha EV6 600 MHz processor. A rough estimate of the number of particles required to calculate a map to 4Å, the level of resolution where an atomic model of a protein could be built, is on the order of \( 10^6 \). This is based on the number of independent asymmetric units used to calculate atomic structures of proteins from electron micrographs of 2D arrays (e.g. Henderson et al., 1990; Kühbrandt et al., 1994; Nogales et al., 1998). This number of particles corresponds to a 1000-fold scale-up, and the computing time required would be 2*56*1000 min. A modest array of 10 processing units could pick the \( 10^6 \) particles in 7 days. The speed of the calculation depends on the total image size, not on the particle size.

The particle picking program presented is therefore capable of accurately selecting enough of these KLH particles for high resolution structural analysis. A particle of similar size, but without symmetry (e.g. a ribosome), would need several hundred templates to explicitly represent all possible views and a corresponding level of speed-up to the procedure would be desirable. It is possible that a 10-fold speed improvement could be achieved through some code optimization, use of faster processors and (if all else fails) more processors. In addition another \( \sim 10 \)-fold improvement could be achieved by optimal choice of the spatial and angular sampling levels. In Fig. 4 it can be seen that groups of the projections appear to be very similar, and that a carefully chosen subset could comprise an accurate representation. The weighted form of the FLCF mentioned above could enhance the power of such a compact set of representative projections.

3.3.2. A more sensitive FLCF with weighting

An improvement to the local correlation function would be to include a statistical weighting related to every point on the template or image. The correlation coefficient would then be equivalent to a weighted least squares solution. The weighting could in practice be derived from the variance of the 3D model projected to make the template, or the variance of particles averaged to make the template. This might allow more sensitive discrimination of particles on the micrographs, because the contribution of more reliable features would have a higher weighting, and variable features would contribute to a lesser extent. It could also allow the range of views to be accurately represented with a more compact set of templates.

The local correlation function, \( C_L(x) \), is defined in Roseman (2003) as,

\[
C_L(x) = \frac{1}{P\sigma_{MT}(x)} \sum_{i=1}^{N}(S_i M_i T_{i+x}),
\]

where \( S \) is the search object, \( M \) is the binary mask, \( T \) is the target map, and \( P \) is the number of points under the mask. \( N \) is the number of points in the target map and \( \sigma_{MT}(x) \) is the standard deviation of the target map points under the mask at position \( x \).

The correlation coefficient is closely related to the least squares residual between the search object and target map densities, having optimised the linear numerical scaling and an additive constant in order to minimise this residual. Similarly, the weighted form of the least squares residual can be represented by explicitly weighting corresponding pixel elements in the search object and target map with weights \( w_i \).

\[
C_{WL}(x) = \frac{1}{P\sigma_{MT}(x)\sigma_{\Delta}(x)} \sum_{i=1}^{N}(w_i S_i M_i w_i T_{i+x}),
\]

\[
C_{WL}(x) = \frac{1}{P\sigma_{MT}(x)\sigma_{\Delta}(x)} \sum_{i=1}^{N}(w_i^2 M_i S_i T_{i+x}).
\]

Therefore this calculation can be achieved by pre-multiplying the relevant mask and search object terms with the pixel-wise weights \( w_i \), and the extra computation time required will be negligible.

4. Conclusion

Absolute objective evaluation of a particle selection is difficult to achieve. The final selection of particles included in the 3D structure calculation will always be determined in the fine refinement against the current 3D structure, and at this stage wrong particles, damaged particles or distorted images will be excluded by objective criteria. Before a 3D structure is known it is not possible to define what a "good" view of a particle is, and it is not possible to discriminate damaged or distorted particles (or those in an alternate conformation), or views of wrong particles, from potential alternative views of the required object. Bearing in mind this context, some level of false positives in the initial particle selection can be tolerated—the most damage they will do is time wasting. Also, considering the huge saving in human labour that automatic particle picking offers, it can be tolerable that a small fraction of particles is missed—the processing of a few extra images will compensate.

In spite of the difficulties mentioned above, the results presented here indicate a high level of accuracy (\( \sim 95\% \)) while the level of false positives is reasonable (\( \sim 20\% \)). It is possible that these results could be improved by increasing the angular sampling, i.e. using more templates representing more views.
The template based method using local correlation provides a versatile, robust, relatively fast and efficient system for automatic particle selection from electron micrographs. Use of the program in two alternative ways has been demonstrated. The system and tools provided are versatile and allow optimum procedures to be developed for a particular problem. For example, an iterative strategy to use Method 1 following Method 2 to generate initial templates would be quite efficient. Another specialised implementation could be to search specifically for more rare particle views, once a 3D model has been obtained.

For most specimens, selection based on a few templates will be adequate. Exhaustive 3D projection matching is feasible for acquisition 10 000’s of particles, though some of the options for speed-up will need to be considered if very large datasets (>100 000’s of particles) are desired. The program is available on request from the author.

Acknowledgments

I am grateful to R. Anthony Crowther for helpful advice and comments on the manuscript. My appreciation goes to the organisers and participants of the automatic particle selection workshop for a stimulating meeting. The KLH image datasets were provided by the National Institutes of Health through the National Center for Research Resources’ P41 program (RR17573). Support from the Agouron Institute is also acknowledged.

References