Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 145 (2004) 3–14
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Automatic particle selection: results of a comparative study

Yuanxin Zhu,[a] Bridget Carragher,[a] Robert M. Glaeser,[b] Denis Fellmann,[a] Chandrajit Bajaj,[c] Marshall Bern,[d] Fabrice Mouche,[a] Felix de Haas,[e] Richard J. Hall,[f] David J. Kriegman,[g] Steven J. Ludtke,[h] Satya P. Mallick,[g] Pawel A. Penczek,[i] Alan M. Roseman,[j] Fred J. Sigworth,[k] Niels Volkmann,[l] and Clinton S. Potter[a,*]

[a] *Center for Integrative Molecular Biosciences and Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA*
[b] *Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA*
[c] *Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712, USA*
[d] *Computer Sciences Lab, Palo Alto Research Center, Palo Alto, CA 94304, USA*
[e] *Division of Electron Optics, FEI Company, Eindhoven, The Netherlands*
[f] *Department of Biological Sciences, Imperial College London, London SW7 2AY, UK*
[g] *Department of Computer Science and Engineering, University of California, San Diego, CA 92093, USA*
[h] *National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX 77030, USA*
[i] *Department of Biochemistry and Molecular Biology, University of Texas-Houston Medical School, Houston, TX 77225, USA*
[j] *MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*
[k] *Department of Cellular and Molecular Physiology, Yale University School of Medicine,New Haven, CT 06520, USA*
[l] *The Burnham Institute, La Jolla, CA 92037, USA*

Received 15 August 2003

Abstract

Manual selection of single particles in images acquired using cryo-electron microscopy (cryoEM) will become a significant bottleneck when datasets of a hundred thousand or even a million particles are required for structure determination at near atomic resolution. Algorithm development of fully automated particle selection is thus an important research objective in the cryoEM field. A number of research groups are making promising new advances in this area. Evaluation of algorithms using a standard set of cryoEM images is an essential aspect of this algorithm development. With this goal in mind, a particle selection "bakeoff" was included in the program of the *Multidisciplinary Workshop on Automatic Particle Selection for cryoEM*. Twelve groups participated by submitting the results of testing their own algorithms on a common dataset. The dataset consisted of 82 defocus pairs of high-magnification micrographs, containing keyhole limpet hemocyanin particles, acquired using cryoEM. The results of the bakeoff are presented in this paper along with a summary of the discussion from the workshop. It was agreed that establishing benchmark particles and using bakeoffs to evaluate algorithms are useful in promoting algorithm development for fully automated particle selection, and that the infrastructure set up to support the bakeoff should be maintained and extended to include larger and more varied datasets, and more criteria for future evaluations.

Keywords: Electron microscopy; Single-particle reconstruction; Automatic particle selection; Image processing; Pattern recognition

1. Introduction

Selection of individual particles from digitized electron micrographs begins to represent a labor-intensive bottleneck in single-particle cryo-electron microscopy (cryoEM) when the size of the dataset that is needed starts to exceed a few tens of thousand molecular images. The automation of particle selection has been a topic of interest for many years (for a review, see, Nicholson and Glaeser, 2001). Apart from the task of selection of images of spherical virus particles at relatively high defocus, computer algorithms alone have not been as effective as most users wish them to be. As a

* Corresponding author. Fax: 1-858-784-9090.
E-mail address: cpotter@scripps.edu (C.S. Potter).

1047-8477/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2003.09.033

result, current algorithms for automated (computer) selection are primarily used to select candidate particles, after which one manually edits (prunes) the list of candidates by visually inspecting and accepting—or rejecting—every one of the candidates. While *semi-automated particle selection* of this type is a big aid when one aims for datasets of ten or twenty thousand particles, the need to develop fully automated algorithms becomes rather important when one aims for datasets of a hundred thousand or even a million particles. Such large datasets are a prerequisite for cryoEM reconstructions that approach the resolution limit associated with large, single particles (Glaeser, 1999; Henderson, 1995). As a result, developing fully automated algorithms is an important research objective.

As is apparent from the papers presented at the recent Workshop (Multidisciplinary Workshop on Automatic Particle Selection for Cryo-electron Microscopy, The Scripps Research Institute, April 24–25, 2003), a number of research groups are making promising new advances on the difficult task of developing algorithms for fully automatic particle selection. It is also apparent that the outcomes achieved with alternative recipes must ultimately be evaluated by comparing how successful each one is when tested on a standard set of electron micrographs. The felicitous concept of holding a "bakeoff," in which each chef is restricted to using a common set of raw ingredients (micrographs), emerged from this metaphor of "algorithm as recipe." Thus, to initiate what could become a tradition, such a bakeoff was included in the program of this workshop, and the results are presented in this paper. In summary, 12 groups participated in the bakeoff, of which two groups manually selected particles and the others used automated algorithms. Table 1 includes a summary of the bakeoff participants in terms of representatives and group affiliations.

Ideally, to fully evaluate the performance of various approaches, we should use a range of particle datasets from the simplest spherical virus particles to the most difficult very-low-contrast asymmetrical particles. Unfortunately, such datasets are not readily available and thus, rather than deferring the problem to a later time, we chose to get started by using an available annotated dataset of images containing keyhole limpet hemocyanin (KLH) particles (Zhu et al., 2003). As a result, the performance of individual algorithms reported in this paper is limited to the selection of the KLH particles. It is understood that the KLH dataset is not "ideal" and that algorithms might perform completely differently on datasets that represent more (or less) challenging problems. However, in spite of these limitations, we believe that the results of the bakeoff provide a useful starting point for a discussion on how best to compare and evaluate algorithms and how to set up more general standard datasets for further evaluations. Thus, although selecting the KLH particles presented a relatively "easy" problem in particle selection, the bakeoff served as a common basis for us to better understand how to build benchmark particle datasets as well as how to set up criteria for evaluating methods of particle selection. Given the specific nature of the dataset, the major goal of the bakeoff is focused more on how to compare and contrast the results of different algorithms and less on the performance of individual algorithms.

Table 1
Bakeoff participants

Representative	Affiliation
Chandrajit Bajaj	University of Texas at Austin
Marshall Bern	Palo Alto Research Center
Fabrice Mouche	The Scripps Research Institute
Felix de Haas	FEI Company
Richard J. Hall	Imperial College London
Steven C. Ludtke	Baylor College of Medicine
Satya P. Mallick	University of California, San Diego
Pawel A. Penczek	University of Texas-Houston Medical School
Alan M. Roseman	MRC Laboratory of Molecular Biology
Fred J. Sigworth	Yale University
Niels Volkmann	The Burnham Institute
Yuanxin Zhu	The Scripps Research Institute

2. Materials and methods

2.1. Common dataset

An ongoing effort at the National Resource for Automated Molecular Microscopy (NRAMM) is to develop benchmark cryoEM datasets that can be used to test methods for automatic particle selection. As part of the effort, an *annotated dataset* of cryoEM images containing KLH particles has been established (Zhu et al., 2003) and was used for the bakeoff. The annotated dataset consisted of 82 defocus pairs of high-magnification images of KLH particles, locations of around 1000 side view particles in the images manually selected by Mouche (one of the participants), and a preliminary 3D reconstruction. The defocus pairs were acquired at a nominal magnification of 66 000× and a voltage of 120 kV, using the Leginon system (Carragher et al., 2000; Potter et al., 1999) and a Philips CM200 transmission electron microscope equipped with a 2048 × 2048 CCD Tietz camera. The first image of each defocus pair was acquired at near to focus conditions (e.g., 1 μm under focus) and the second one at farther from focus conditions (e.g., 3 μm under focus). The time interval between the two exposures was approximately 20 s due to the time required to read out the digital image from the camera. At this magnification, the pixel size is 2.2 Å on the specimen scale and the accumulated dose for each high-magnification image area was about 10 $e^-/Å^2$.

This KLH Dataset-1 is publicly available at: http://ami.scripps.edu/prtl_data/.

2.2. Bakeoff rules

An example of a defocus pair of images is shown in Fig. 1. The KLH didecamer appears in two main orientations, as rectangular side views and as circular top views. Images typically also contain intermediate views of broken molecules and aggregates of two or more particles. Bakeoff participants were required to select only side view KLH particles in the farther from focus image of each defocus pair. This requirement was imposed because no top view KLH particles were originally manually selected in the common dataset. It is widely accepted that using overabundant type of views (here the top views) may lead to later reconstruction artifacts (Boisset et al., 1998); therefore, the top views are usually not used for 3D reconstruction of KLH maps—the major driving force of automatic particle selection. This explains why the top view KLH particles were not annotated in the common dataset.

A call for participation and a specification for the bakeoff, including how to submit particle selection results, the deadline for submission, and the suggested method of assessing different results, were made known to the participants. Each participant was required to provide the center coordinates of selected particles in an ASCII file, each row of which records the coordinates of a particle, with the origin of the coordinate system being at the bottom-left corner of the image. Each participant was also asked to submit a text file containing any information that is important or would be helpful to other people in understanding the results. (More detailed information about the bakeoff can be found at: http://nramm.scripps.edu/seminars/2003/prtl_work/bake off.htm.)

2.3. Algorithms/criteria for particle selection in the bakeoff

As mentioned in Section 1, 12 groups participated in the bakeoff, of which Mouche and Haas manually selected particles using their own criteria and the others used automated algorithms. The 10 algorithms used by the other participants can be more or less grouped into two classes. Class I algorithms are based on cross-correlation using templates (references), generated from either a 3D reference structure or the averages of a set of manually picked particles. These are called template matching-based approaches, including Bern's, Ludtke's, Penczek's, Roseman's, and Sigworth's algorithms. Class II methods are based on feature recognition where algorithms work by way of recognizing local or global salient features inherent to particle images without the use of a 3D reference structure, called feature-based approaches, including Bajaj's, Hall's, Mallick's, Volkmann's, and Zhu's algorithms. Unlike the other feature-based approaches reported here, Mallick's algorithm uses machine learning as the basic tool to learn both discriminative features and a cascade of classifiers for particle detection (Mallick et al., 2004). There are distinct advantages to each of these approaches and these are described later in this section. Algorithms requiring a 3D model and those starting from pure features represent two different starting points to the task of particle selection or, in other words, stand at opposite ends of a continuum of methods for automatic particle selection. From this point of view, some participants' algorithms may be more

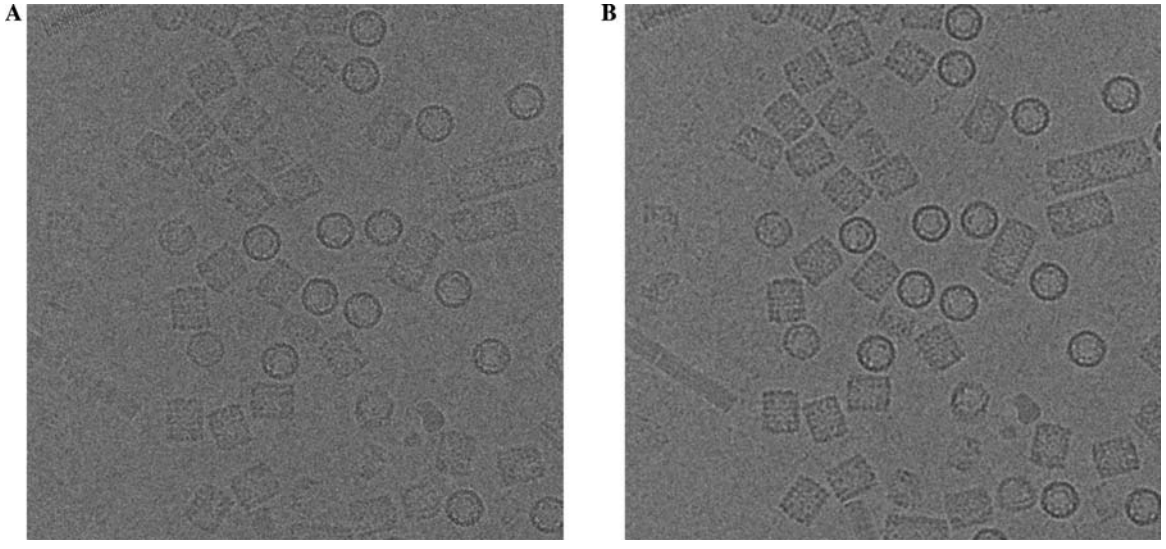

Fig. 1. An example defocus pair of images from the common dataset of a specimen of keyhole limpet hemocyanin (KLH). The images are acquired at a nominal magnification of 66 000× using a 2048 × 2048 pixel CCD camera. The image shown in (A) was acquired first at a near to focus condition (1 μm under focus) and the one shown in (B) was recorded second at much farther from focus (3 μm under focus).

accurately classified into somewhere between the two opposite ends, for instance, Hall's algorithm. For completeness, a brief description of all algorithms and the criteria used for the manual selections are given below.

2.3.1. Manual selection criteria

2.3.1.1. Mouche's criteria. The KLH didecamer presents two main orientations, a rectangular side view and a circular top view. From the 82 images obtained with the CCD camera, 1042 single particles were manually and interactively selected, using SPIDER and WEB (Frank et al., 1996). Only rectangular side views and intermediate orientations were selected. No aggregate or "single" particle showing a different length (shorter or longer) was manually picked. Furthermore, to avoid any reconstruction artifacts due to overabundant views, no circular top views were selected.

2.3.1.2. Haas' criteria. Particle picking was guided by rules as follows: (i) Only particles of the rectangular side view were selected; (ii) Particles should not overlap; (iii) Particles should not have any defects (dissociated, contaminated with ice crystal deposits, etc.); and (iv) Particles should be clearly visible (not too thick an ice layer).

2.3.2. Template matching approaches

Template matching is a basic technique used in many signal processing and image analysis applications for detection and localization of patterns in signal corrupted by noise. The technique is based on a linear image formation model, i.e., it is assumed that the observed signal is a sum of the original, uncorrupted signal and noise; the latter is further assumed to be stationary with zero average, with a known power spectrum, and to be uncorrelated with the signal. If the noise is white, the template matching reduces to the correlation technique for signal detection. For colored noise, one constructs a linear Matched Filter that takes into account the power spectrum of the noise. The popularity of the template matching technique is further enhanced by the fact that the matched filter can be shown to be an optimal Bayesian classifier, i.e., it minimizes the probability of the detection error.

The image formation model underlying the template matching approach corresponds well to the accepted model of the image formation process in the electron microscope in its linear, weak-phase approximation (Wade, 1992). Thus, if the necessary parameters of the EM transfer function could be estimated, the template matching would provide results that, at least in theory, could not be surpassed by the usage of any other linear method. The only remaining problem is of a practical nature: how to create a set of 2D template images that would be exhaustive, i.e., would contain all possible views of the known 3D template structure, but it would be sufficiently small to make the application of the method practical. Since in EM the goal is the detection of any of the possible 2D projections of the known 3D structure, the number of templates can be very large. Not only all possible projection directions have to be considered, but also all the possible in-plane orientations of projections have to be generated. In order to reduce the number of templates, two possible strategies have been suggested. In the first strategy (e.g., Sigworth's algorithm), principal component analysis is used to express the large number of original templates as linear combinations of a small set of eigenimages. In the second (e.g., Penczek's and Bern's algorithms), clustering techniques are applied to group the templates and a small number of class averages are then used as templates.

The main weakness of the template matching technique is that it results in a relatively high rate of false positives. Any objects in the field that have about the same size and average intensity as the templates will yield high correlation coefficients. Thus, further post-processing of the template matching results is necessary in order to improve the performance of this technique.

2.3.2.1. Bern's algorithm. The algorithm starts by projecting an initial reference 3D map in many different directions to produce synthetic 2D templates. The templates are clustered, and cluster averages are then cross-correlated with the micrographs, using the fast Fourier transforms (FFT) for speed, to give a set of candidate picks. Afterwards, the candidate picks are screened by scoring them using a probabilistic model of cryo-EM image formation; the score is the ratio of the probability of generating the candidate pick by cryo-EM imaging of a template to the probability of generating the candidate by a pure noise process. In scoring, the original synthetic templates are used, rather than the cluster averages. In principle, this algorithm allows the use of almost any noise model, even one learned from the data as in Mallick's algorithm, but in their bakeoff entry, Bern and his coauthors (Wong et al., 2004) used a simple noise model with independent, identically distributed pixels, whose distribution was determined empirically. This noise model gives an algorithm similar to classical matched filtering (Sigworth's algorithm), but with less emphasis on the "power term" (grayscale variance) of the candidate picks, although not as severely normalized as using the correlation coefficient (Roseman's algorithm).

Key parameters in the algorithm include the number of 2D templates and their Euler angles, the number of clusters of templates, and the score thresholds for accepting candidate picks. Bern et al. used 35 templates, 5 top views, and 30 side views (planar rotations of 0°, 6°, 12°,... of a master side view, which was rotationally averaged about the KLH's axis of symmetry); they used five clusters, one of which consisted of top views; and they set the score thresholds based upon manual

examination of the picking results for a few micrographs. The algorithm picked both side and top views, but only the side views (determined by which 2D template they matched with highest score) were included in the bakeoff entry. The processing time of the algorithm is approximately 2 min per micrograph.

2.3.2.2. Ludtke's algorithm. A particle with a good side view was selected as a reference. High-pass and low-pass filters were applied at 1 pixel and 70 pixels, respectively. A stack of reference images for use with "boxer/batchboxer," EMAN's interactive/batch particle selection tool (Ludtke et al., 1999), was generated by rotating the reference particle in 5° steps. Boxer and batchboxer both use a multi-reference correlation-based scheme with several thresholds. The correlation function is based on Alan Roseman's fast local correlation technique (Roseman, 2003). A single image was loaded into boxer and a reasonable set of threshold parameters were selected. This single set of thresholds was then used to automatically select particles out of all high-magnification images.

2.3.2.3. Penczek's algorithm. A template matching approach was used for particle selection, taking advantage of the existing, intermediate-resolution model of the structure. The approach comprised three steps. In the first step, a set of possible particle views was generated using the available reference structure. The template images were constructed as linear combinations of available particle views using the rotationally invariant K-means clustering technique (Huang and Penczek, 2004). Since the goal was to detect only the side views in the micrographs, only the side views of the reference structure were selected and their number was reduced to three using the clustering technique. Next, in-plane rotated copies of the template images were created using the step size of 10°. This resulted in 108 templates. In the second step, the noise characteristics for all micrographs were established using an automated contrast transfer function (CTF) estimation procedure where it is assumed that the radially averaged power spectrum of the whole micrograph, calculated using the method of averaged overlapping periodograms, yields a robust estimate of the noise power spectrum. Concurrently, the CTF parameters were automatically calculated based on estimated power spectra. Third, the noise power spectrum and the CTF parameters were used to construct a matched filter (Huang and Penczek, 2004). This was done by applying the appropriate CTF to the respective micrograph, the product was divided by the noise power spectrum, and finally the result was normalized using the fast Fourier-space technique to estimate moving average and moving variance using the window size corresponding to the particle size. The Fourier transform of the normalized result was multiplied by the Fourier transforms of appropriately padded template images to yield a set of cross-correlation functions. To speed-up the procedure, the input data were decimated twice. Detection criteria were: the maximum correlation coefficient with one of the reference images must be above a pre-selected threshold; the maximum correlation coefficient is accepted if there are no larger correlation coefficients within the neighborhood corresponding to the particle size.

2.3.2.4. Roseman's algorithm. The FindEM (Roseman, 2003) program was used to select the particles. It uses local correlation with templates to detect occurrences of objects similar to the templates in the micrograph fields. The advantage of the local correlation algorithm is that the density scaling between the template and the local region of the micrograph being compared is optimized, whereas the conventional correlation applies a global normalization and details beyond the local region of interest can distort the correlations.

There are two stages to the procedure. First the templates are made and the correlation maps are calculated. Initial templates were generated by averaging 20 hand picked particles from the first of the images in the series, which were optimally aligned using an iterative orientation and cross-correlation procedure. A template was created for each of the two predominant views, the side view and top view. Each template was correlated in turn with each micrograph image, covering all orientations of the template relative to the micrograph by successively rotating the template in steps of 4°. The final correlation map output, for each template, indicated the maximum correlation at each point, over all orientations. The images and templates were reduced in size by a factor of 4, and band-pass filtered in the range 30–2000 A.

In the second stage, peak positions from the correlation maps are extracted and filtered according to a correlation-coefficient threshold and interparticle distance criteria (or particle size). When peak positions from different templates coincided, the particle was assigned to the class of the template it correlated best with. The particle size was chosen to include side views that were almost touching, but not overlapping. The parameters were optimized by examining the particles chosen on ~5 images, using the graphical interface that is part of the FindEM package. This allows interactive adjustment of the parameters while the images are displayed with the selected particles overlaid. These parameters were then used to automatically select the particles from the set of 82 images. The procedure was reiterated once, submitting the average of all selected side views and top views as new templates. The particles detected as side views were submitted for the bakeoff.

A manual de-selection option is also available but was not used for the particle set submitted for the bakeoff, which was completely automatically generated. More details on the procedure and the exact parameters used are given in the accompanying paper (Roseman,

JSB, 2004). The time taken to find the ~1000 side view KLH particles for the bakeoff was 56 min per template, using a DEC alphaEV6 600 MHz computer.

2.3.2.5. Sigworth's algorithm. This is a model-based, multiple-reference detector that uses a white Gaussian noise model. We first get an estimate of the circularly averaged power spectrum of the background, and build an inverse filter to "whiten" the noise. Each data image is processed by this filter. From the 3D model we build a large set of representative projections, and filter them with the CTF and the same inverse filter to make the references. From these we use singular value decomposition (SVD) to build a small set of eigenimages spanning the set. FFT-based cross-correlations are done with the eigenimages to save time, but then the results are converted to, in effect, cross-correlations with the references. Two statistics are computed: (1) the maximum correlation with one of the references, to give a "motif amplitude" and (2) a weighted sum of the power spectrum of the residual, after subtraction of the best-fit reference, from the putative particle image. Thanks to the pre-whitening of the original image, both statistics have predictable distributions.

In the case of the KLH particles, images were first binned to reduce them to 512×512 in size. Power spectra from "empty" regions of some of the images were used to construct the inverse filter. The references were 64 rotated "side views" of the KLH particle. From the SVD the first 13 eigenvectors were kept. Allowable values for the two statistics were chosen by comparison with manual picks. After setting up the references, picking took ~15 s per image using Matlab on a fast PC.

2.3.3. Feature-based approaches

In comparison to template matching where a large number of image pixels are used, feature-based approaches usually rely on a small set of local or global salient features of particle images, including geometric features such as positions of corners, line segments, contours, etc., and statistical features such as moments, and so on. Procedures for feature-based approaches vary widely but three major components may be identified: the definition of a discriminative feature set, the extraction of these features from an image, and the recognition algorithms. In addition to less demanding computational requirements, in principle, distinctive features invariant to scale, rotation illumination variations, and/or 3D projection can be extracted for fast object detection (Lowe, 1999) and thus it is quite desirable for the task of particle selection. The main weakness of feature-based approaches is that it may be difficult to extract distinctive features pertinent to a specimen when dealing with very low-contrast images.

2.3.3.1. Bajaj's algorithm. The algorithm is designed for detecting circle-like and rectangle-like particles, but it is also possible to extend it to other types of particles if certain geometric features can be derived from the shapes of the particles (e.g., icosahedral viruses). The method is fast, fully automatic, and reference-free. The steps are as follows: (1) Detect the edges using Canny edge detector (Canny, 1986); (2) Remove the connected components of edges that contain too few edge pixels; (3) Compute the Voronoi diagram (VD) and distance transform (DT) of the edges obtained from the last two steps (Guan and Ma, 1998); (4) Use the distance transform map to detect and refine the circles; (5) Use the Voronoi diagram and distance map to detect and refine the rectangles; (6) Let the detected circles compete with the detected rectangles, with the assistance of distance maps. It is assumed that the size of the circle and the rectangle are fixed for all detected particles. However, it could be possible for us to improve our algorithm to detect the particles with flexible sizes.

The false positive rate (FPR) and the false negative rate (FNR) for the side views of KLH are listed in Table 2. In case of the top views, the FPR and FNR are 13.6% and 2.6%, respectively, evaluated against visual detection. This method was tested on SGI Onyx2 with single processor (400 MHz MIPS R12000) and the total computational time is about 20 s for each image with a size of 1024×1024 pixels. About 18, 2, and 6% of the total time are used for edge detection, edge cleaning and computations of DT&VD, respectively. The rest of the total time is used for the particle detection (including circle detection, rectangle detection and circle-rectangle competition).

2.3.3.2. Hall's algorithm. The algorithm was developed as a general method for automated selection of particles, independent of shape, size, image quality, and the availability of a model. Selection is carried out in two stages; the first being a template matching stage using a rotationally averaged sum of a small number of manually picked particles. The cutoff used at this point is such that no particles are missed resulting in a very large number of false positives. The second stage involves calculation of a feature vector for each picked region and clustering using a self-organizing map (SOM) (Kohonen, 1989). The feature vector is made up of 16 features, including four statistical characteristics of the total distribution of gray values, four textural characteristics (Lata et al., 1995), and eight morphometric characteristics calculated from the largest continuous object found when the image is segmented based on local variance (van Heel, 1983). The SOM can be automatically interpreted, giving an optimal number of clusters for the data; it is then up to the user to select which clusters contain particles. The method was developed on very noisy low-contrast micrographs, and has been demonstrated on RNA polymerase data that proved difficult to pick by eye (Hall and Patwardhan, 2004).

2.3.3.3. Mallick's algorithm. This is a feature-based discriminative learning approach that learns important features derived from the so-called integral image of the original particle image using a set of representative examples including both particle and non-particle images (Mallik et al., 2004). The core learning algorithm, Adaboost (Freund and Schapire, 1995), has been successfully used in the domain of face detection by Viola and Jones (2001). The approach can be divided into an off-line learning phase followed by on-line particle detection. The result of the learning phase is to produce a two-category classifier which takes as input a window of a digital micrograph (e.g., a 50×50 pixel sub-image) and classifies it as either containing a particle or not containing a particle. During on-line detection, a detection window is scanned over an input micrograph, and for each location (pixel), the sub-image covered by the window centered at that location is classified as particle/non-particle. As there will usually be positive responses at multiple, neighboring locations for each particle, the results are post-processed using connected component analysis (Horn, 1986), and the mean of each component is reported at the location of a particle. If the detector is trained to only detect particles in a particular 2D orientation in the image plane while particles in a micrograph may appear at any orientation, then the detector is scanned multiple times. During each scan, particles in a particular orientation are detected; either the detector is "rotated" with each scan, or else the detector is fixed, but the image is rotated. The processing time on a micrograph from the common dataset, decimated to 512×512 pixel, at eight different orientations was about 6 s on a 1.3 GHz Pentinum M processor. The algorithm is fast, generic, and is not limited to any particular shape or size of the particle to be detected.

2.3.3.4. Volkmann's algorithm. The particle selection algorithm relies on the use of reduced representations. In this approach, the underlying motif is approximated by a small number of locations that capture the intensity characteristics of the motif (Volkmann, 2004). Reduced representations can be constructed from models or directly from the data. The reduced representation for this application was constructed from the average of 75 hand selected side views. This representation was then used for real-space template matching. One advantage of the reduced representation strategy is the gain in speed. In this application, a box of 240×240 pixels containing a particle side view can be efficiently reduced to 40 locations. For real-space scoring functions, this is a gain in speed of a factor of better than 1000, for four times compressed images the speed gain is still about 100. For this application, a model-free three-step procedure was used. First, the reduced representation template was constructed directly from the data, second the real-space template matching module was run on the micrographs using this reduced representation, and third a peak recognition program was run for the actual identification of particles in the peak image. Peaks corresponding to real particles tend to be sharper than those corresponding to random noise or different views. The peak recognition software only picks peaks above a certain threshold that do exceed a certain degree of sharpness. These parameters (threshold and degree of sharpness cutoff) need to be adjusted to optimize performance. Here, two micrographs were picked randomly and the parameters were adjusted to minimize false positives. Recently, a fourth step was added to the procedure to increase the number of picked good particles while still keeping the false positives to a minimum. Tests indicate that this additional step leads to significant improvements over the implementation used for the bakeoff (Volkmann, 2004).

2.3.3.5. Zhu's algorithm. A two-stage framework is use for automatic selection of KLH particles. Under this framework, a cryoEM image is first decimated to generate a much smaller sized image with a coarser resolution but increased signal-to-noise ratio. Candidate particles in the decimated image are detected using edge and contour information, particularly the Hough transforms (Zhu et al., 2003). Afterwards, candidate particles in the original full-resolution image are extracted by projecting the coordinates of particles in images with a coarser resolution. The candidate particles are then subject to a second stage of processing—pruning false alarms. In this stage, a correlation-based template matching method is applied to effectively reject low-quality particles or junk, using templates generated by aligning and averaging the candidate particles. With this two-stage framework, computational efficiency is achieved through the coarse-to-fine strategy while the high accuracy relies on the refinement in the second stage. The time required for picking side view KLH particles depends on the number of particles in an image, but is roughly 1 min per image.

3. Results and discussion

As described in Section 1, due to the specific nature of the dataset, the major goal of the bakeoff focuses more on how to compare and contrast the results of different algorithms and less on the performance of individual algorithms. As we know, even for experts, the final set of particles selected from the same set of images may vary from person to person. Even for the same expert, one's criteria of determining whether to pick a particle may change with time (that is, from image to image) during a single experimental session. For this reason, we currently assess the results from different participants by comparing one result against another's, measured by the false negative rate (FNR) and false positive rate (FPR). Taking one participant's result as the truth set and another's

Table 2
Confusion matrix generated using the results provided by the bakeoff participants

Test / Truth		Bajaj	Bern	Mouche (Manual)	Haas (Manual)	Hall	Ludtke	Mallick	Penczek	Roseman	Sigworth	Volkmann	Zhu
Bajaj (1269)			33.9 11.5	24.7 8.3	31.0 7.0	42.2 24.3	51.9 21.0	28.0 9.8	52.9 25.2	17.4 14.0	37.4 5.1	38.5 9.2	24.0 11.4
Bern (948)		11.5 33.9		16.2 23.8	21.5 21.0	36.3 37.7	43.1 30.3	17.7 23.1	48.4 38.8	10.3 30.3	26.4 16.7	29.9 22.8	17.1 28.0
Mouche (1042)		8.3 24.7	23.8 16.2		11.7 2.3	27.4 22.0	43.4 23.7	14.2 11.7	46.8 30.7	2.4 16.6	23.2 4.5	27.4 12.2	9.7 13.7
Haas (944)		7.0 31.0	21.0 21.5	2.3 11.7		26.2 28.2	41.1 28.4	12.2 18.4	44.0 33.9	1.5 23.9	18.4 8.4	22.9 15.7	8.8 21.3
Hall (969)		24.3 42.2	37.7 36.3	22.0 27.4	28.2 26.2		52.0 39.9	30.1 33.2	55.9 46.6	19.3 35.8	35.3 25.2	39.3 31.7	25.7 33.7
Ludtke (775)		21.0 51.9	30.3 43.4	23.7 43.4	28.4 41.1	39.9 52.0		23.0 41.2	48.3 50.0	20.3 49.4	27.1 32.7	32.3 39.1	23.5 45.4
Mallick (1015)		9.8 28.0	23.1 17.7	11.7 14.2	18.4 12.2	33.2 30.1	41.2 23.0		46.7 32.5	7.0 22.6	25.8 10.3	30.1 17.9	14.5 20.5
Penczek (799)		25.2 52.9	38.8 48.4	30.7 46.8	33.9 44.0	46.6 55.9	50.0 48.3	32.5 46.7		23.7 50.0	38.4 41.3	39.7 44.0	30.2 49.1
Roseman (1219)		14.0 17.4	30.3 10.3	16.6 2.4	23.9 1.5	35.8 19.3	49.4 20.3	22.6 7.0	50.0 23.7		33.1 2.7	34.9 7.8	17.5 7.8
Sigworth (838)		5.1 37.4	16.7 26.4	4.5 23.2	8.4 18.4	25.2 35.3	32.7 27.1	10.3 25.8	41.3 38.4	2.7 33.1		12.3 14.6	6.8 28.1
Volkmann (861)		9.2 38.5	22.8 29.9	12.2 27.4	15.7 22.9	31.7 39.3	39.1 32.3	17.9 30.1	44.0 39.7	7.8 34.9	14.6 12.3		11.5 30.0
Zhu (1109)		11.4 24.0	28.0 17.1	13.7 9.7	21.3 8.8	33.7 25.7	45.4 23.5	20.5 14.5	49.1 30.2	7.8 17.5	28.1 6.8	30.0 11.5	
Median/Mean	FNR	11.4/13.3	28.0/27.9	16.2/16.2	21.5/22.0	43.4/44.5	33.7/34.4	20.5/20.8	48.3/47.9	7.8/10.9	27.1/28.0	30.1/30.7	17.1/17.2
	FPR	33.9/34.7	21.5/25.3	23.2/21.7	18.4/18.7	27.1/28.9	30.1/33.6	23.1/23.8	33.9/35.4	30.3/29.8	10.3/15.1	15.7/20.6	28.0/26.3
Standard Deviation	FNR	7.0	7.1	8.6	8.0	6.0	6.7	7.3	4.1	7.9	7.7	8.0	7.8
	FPR	11.3	12.8	14.2	14.4	8.6	11.9	13.0	8.2	12.4	12.7	12.4	13.2

Note. (1) The two values in each table cell represent the false negative rates (FNR) and false positive rates (FPR), respectively, in percentage. (2) The numbers in parentheses represent the total number of particles picked by the corresponding participant.

as the test set, as illustrated in Fig. 2, particles which are selected in the truth set, but fail to be selected in the test set, are false negatives whereas particles selected in the test set but not in the truth set are false positives. Algorithms that can achieve both a low FNR and FPR are considered as having a higher performance and thus more desirable. Given an algorithm, the FNR in general changes in the direction opposite to the FPR. Therefore, one has to make a tradeoff between having a lower FNR with a higher FPR or the opposite based on the requirement of the application at hand. For the selection of side view KLH particles to enter into the bakeoff, the participants made their own decisions as to whether to select more particles (which usually means higher false positive rates) with fewer false negatives or vice versa.

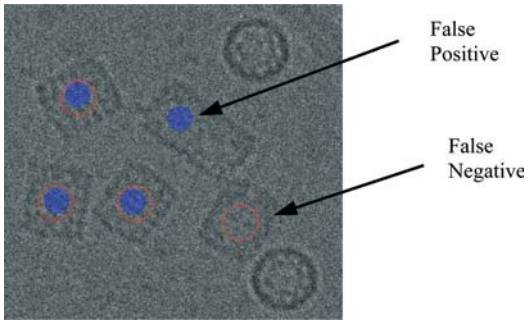

Fig. 2. Illustration of the definition of false negatives and false positives. Particles selected in the truth set are outlined with a red circle and those in test set with a blue dot. Particles selected in the truth set but not in the test set are false negatives. Particles selected in the test set but not in the truth set are false positives.

The specific procedure used for calculating both FNR and FPR of one participant's result against another's is described below. One participant's picks are taken as the truth set and then the other's as the test set. A false negative is found if a particle is picked in the truth set but its pixel distance to its nearest neighbor in the test set is larger than a pre-defined threshold d_T. Likewise, a false positive is found if there is a particle in the test set whose pixel distance to its nearest neighbor in the truth set is larger than the pre-defined threshold d_T. The FNR is then calculated based on the total number of particles in the truth set, while the FPR is calculated from the total number of particles in the test set. In addition, to establish consistency between algorithms, particles from both sets whose pixel distances to the border of the image are less than a pre-defined threshold b_T were removed before the computation of the FNR and FPR. Given the average width of the side view KLH particles as b_T (134 pixels) and half of this width as d_T (67 pixels), a confusion matrix was generated, shown in Table 2.

Among the many observations that can be made from Table 2 it is clear that the two manual selection results are noticeably different from one another. Taking Haas's selections as the truth set, the FNR and FPR of Mouche's results are, respectively, 2.3 and 11.7%. In another words, Mouche only picked 922 out of the 944 KLH particles selected by Haas in the common

dataset though he selected 98 more particles. An example image outlined with particles selected by the two participants is shown in Fig. 3A. For the 13 particles picked by the two participants in the example image only eight of them were selected by both of them. Three out of the five other particles merit a further discussion.

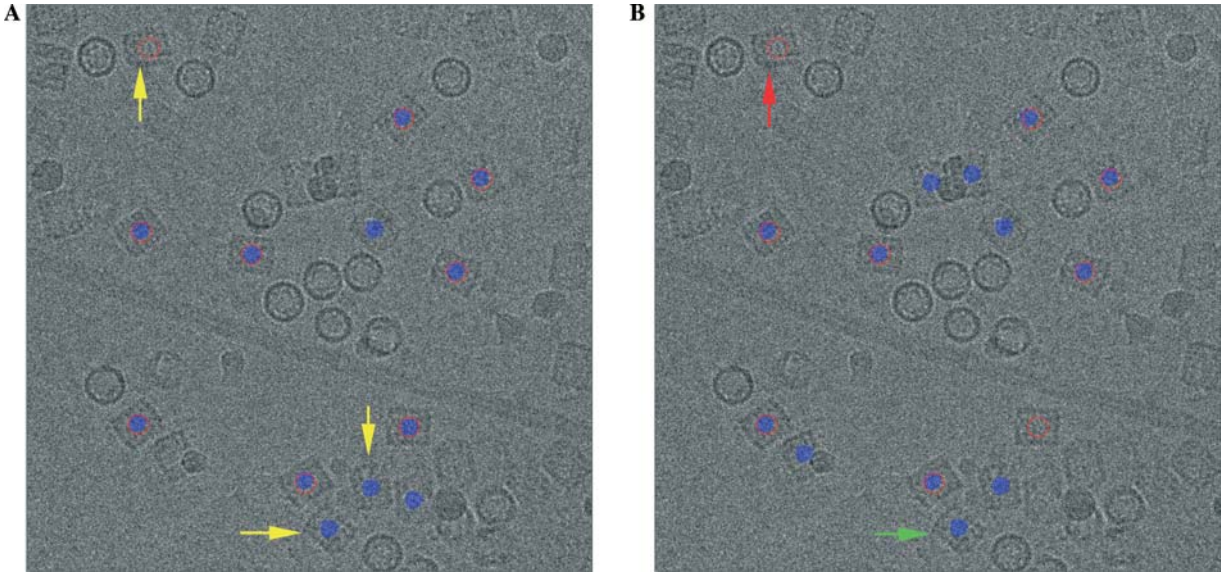

Fig. 3. Illustration of the comparisons between particles selected by two bakeoff participants on the same example image. (A) Particles outlined with a red circle were selected by Haas and those outlined with a blue dot were picked by Mouche. For the 13 particles picked by the two participants only eight of them were selected by both of them. The three particles pointed to by yellow arrow signs are visually undistinguishable, but Haas only selected one of them and Mouche only selected the other two. (B) Particles outlined with a red circle were again selected by Haas and those outlined with a blue dot were picked by Zhu's algorithm. The particle pointed to by a green arrow sign is visually better than the one marked by a red arrow sign, but only the latter one was manually selected. The example image reveals that the criteria used by a person as to whether to select a specific particle may vary with time and images.

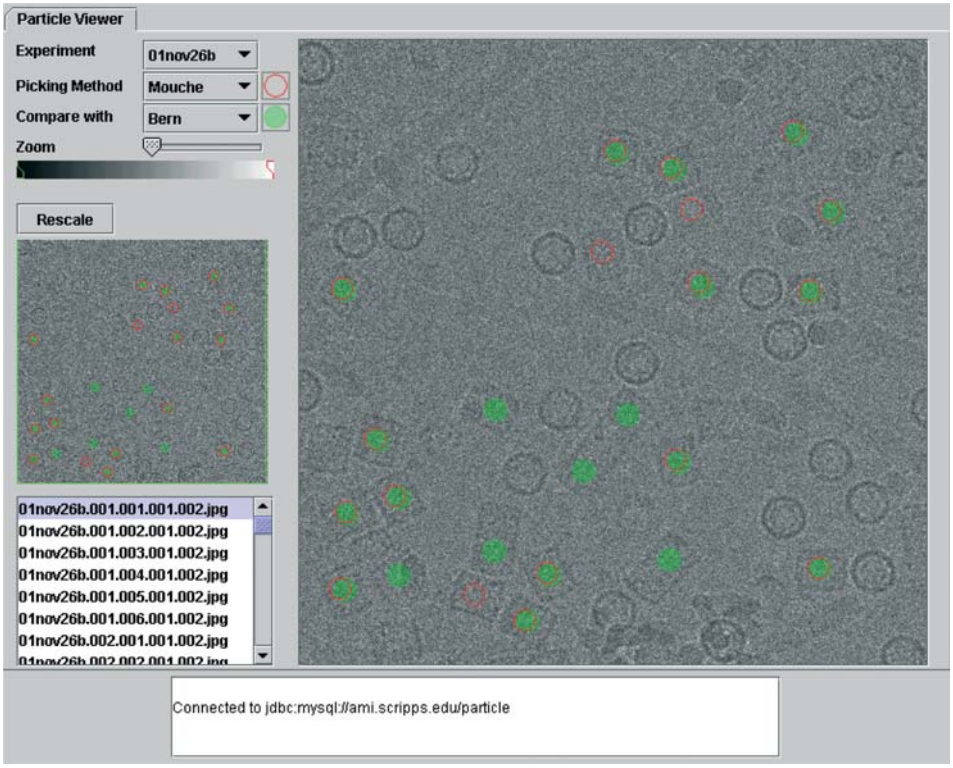

Fig. 5. A screenshot of the interface of the web-based particle selection tool. Users may use the tool in two different modes: to pick particles or to compare one selection against another. The screenshot shows one selection overlaid with green dots and the other outlined by red circles.

The three particles, marked by yellow arrow signs, have visually almost the same image quality, but Haas only selected one of them and Mouche only selected the other 2. This particular example and the overall difference between the two selection sets further demonstrates that different experts may use different criteria when manually selecting particles in the same set of micrographs and that the criteria used by a single expert may vary with time or from image to image. Moreover, the criteria used by an expert as to whether to select a specific particle may be biased sometimes. As shown in Fig. 3B, where the above example image was outlined with the particles selected both manually (by Haas) as well as automatically (by Zhu's algorithm), the particle pointed to by a green arrow sign is visually better than the other particle pointed to by a red arrow sign, but the former was automatically targeted but not manually and the latter was manually selected but not automatically. Had the former one been manually selected, the FPR of the machine algorithm would be even lower on this dataset. As proposed in a previous work (Zhu et al., 2003), this actually raises the question "*How do we build truth datasets of single particles to evaluate machine algorithms?*" This question was intensively discussed during a special session at the workshop. One suggestion was that the basis for particle selection should be determined purely by the final 3D reconstruction. However, it was pointed out that the final 3D reconstruction could be biased by the initially selected particles. Another proposal was that evaluation of particle selection algorithms should be independent of the later 3D reconstruction and all particles should be selected without regard to contamination, broken shape, etc. Apparently, no immediate answer was available to this question and a consensus was not reached.

The more seriously one takes the goal of evaluating the success of automated particle selection, however, the more one also begins to question the success of semi-automated selection (as described in Section 1) or even fully manual (human) selection of particles. Unless a bakeoff is done with synthetic data in which the coordinates of all particles are known in advance, there is always a high probability that there will be some human error in selecting the true particles that represent the "gold standard" that is needed for making such a comparison. Two suggestions for dealing with the potential ambiguity emerged in the workshop discussion. The first suggestion was that the "gold standard" reference-data used in future bakeoffs could be annotated to indicate: (1) the level of human confidence that is attached to the selection of each particle, e.g., "certain," "probable," and "unsure" and (2) the reasons why some of the candidate particles were not included in the human selections (distorted; broken or incomplete; too close to other particles, etc.). The second suggestion was that all new (candidate) particles, which were not identified as being part of the original "gold standard" dataset, should be used to produce a three-dimensional (3D) reconstruction on their own. If another reconstruction is produced with the same number of particles from the "gold standard" dataset, and if both reconstructions are generated with the same number of cycles of refinement, one could then use the Fourier shell correlation to evaluate the quality of the data contained in a dataset that consists exclusively of excess, candidate particles.

A second question one would naturally raise is *how these algorithms perform in selecting side view KLH particles*. Although the performances of different algorithms varied, most algorithms achieved human-level performances. High performances were achieved by both template matching-based approaches (e.g., Roseman's algorithm) and feature-based approaches (e.g., Mallick's algorithm). As listed in Table 2, several algorithms can select over 90% of the particles that have been manually picked either by Haas or Mouche, including Bajaj's, Roseman's, and Zhu's, with false positive rates ranging from 15 to 30%. The lowest false negative rate reported in the Table is 1.5% with a false positive rate of 23.9% by Roseman's algorithm, taking Haas's selections as the truth set. The lowest false positive rate in the Table is 4.5% with a false negative rate of 23.2%, achieved by Sigworth's algorithm, taking Mouche's results as truth set. Compared to manual selections, the highest false negative rate was 46.8% by Penczek's algorithm with the false positive rate of 30.7%. This level of performance seems poor in comparison to most of the other algorithms. After further examination, we found that the high false positive rate is due to the fact that the algorithm did not successfully separate top view particles from side view ones, as shown in Fig. 4. Since top view particles are considered false positives in the bakeoff, a high threshold had to be used in selecting side view particles, which in turn led to a high false negative rate. If the selection of top view particles had been included in the bakeoff, the algorithm would have a better performance. This also explains why Ludtkes' algorithm did not perform well in the bakeoff.

A third question that arises is just *how good the process of automated particle selection needs to be, before it is good enough for routine use*. Two points are important in this regard: (1) how efficient is a given algorithm in selecting most of the particles that a human operator would select, and (2) how many false positives (non-particles) are included in the dataset? Most experimentalists will take a pragmatic view on how efficient the automated data selection process needs to be: if it takes less time to collect additional micrographs than it does to manually select the same number of particles from existing micrographs, then most would prefer to collect a larger number of micrographs and let the computer do the boring job of selecting particles. As a rough guide, at least, most would agree that automated particle selection would be well

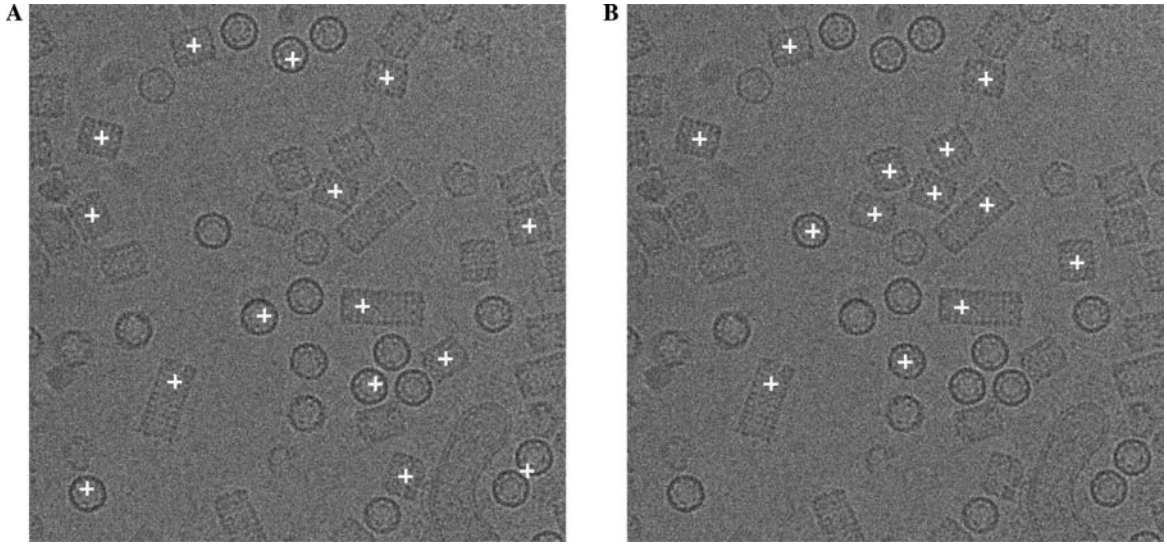

Fig. 4. Illustration of why two of the algorithms did not perform as well as the other ones. A particular example image is outlined with the "side view" particles selected by Penczek's algorithm (A) and those by Ludtke's algorithms (B), respectively. Clearly, the two algorithms did not separate effectively the side views from the top views. Top views are considered false positives in the bakeoff. If the selection of top view particles had been included in the bakeoff, the algorithms would have higher performance.

received as soon as it could routinely select 75% of the particles (i.e., a false negative rate of 25% or below) that an experienced human operator would pick. An important caveat will be that the automated selection process must not systematically miss the selection of one or more sets of views. There is at present no strong evidence on how many false positives can be included in a dataset without corrupting the reconstruction to an unacceptable extent. Most experimentalists would be uncomfortable to use a dataset that is known to contain (or to be likely to contain) 50% or more false positives. Most would surely do a manual editing of such a dataset before proceeding with the 3D reconstruction. On the other hand, however, most would agree that having fewer than 10% false positives in the initial dataset would be quite acceptable. The false positives, if they are structurally uncorrelated with the true positives, will only add to the noise in quadrature (as does the noise that is already present in the images of the true particles). The actual situation is even better than that argument would indicate, however, since many of the false positives that are present in the original dataset will also be deleted early in the data analysis, either because they show up as outliers in a classification step or because they do not adopt stable values of the orientation or position parameters in successive cycles of refinement. The general sense therefore seemed to be that automated particle selection would be likely to become popular once it could be shown to meet or exceed the 25%/10% rule described above. Obviously, further improvements in performance beyond that point would only further cement the acceptance of any given selection tool.

In addition to the confusion matrix, the bakeoff results were also loaded into a web-based particle selection tool, developed at NRAMM. Using the tool, users cannot only select particles in a set of micrographs managed from a database, but also compare the results of two different selections. The comparison is visualized by superimposing two different kinds of icons, each associated with a particular selection, onto the selected particles. Fig. 5 shows a screenshot of the interface of the tool where particles selected by one algorithm were overlaid with green dots and those by the other algorithm were outlined by red circles. The URL of the web-based particle selection tool is http://ami.scripps.edu/leginon/particle_viewer/. Readers can visually compare one bakeoff participants' results against another's by exploring the site. (Note: in order to keep bakeoff results from being changed by a third party, readers in the public domain are only allowed to view particles selected by the bakeoff participants.)

4. Summary and conclusions

Particle selection is critical and could become a bottleneck in moving toward high-throughput high-resolution structure determination of macromolecules using cryoEM. Automatic selection of asymmetric particles in low-contrast cryoEM images is an unresolved challenging problem. This in turn demands a rapid development of fast and accurate algorithms for this purpose. To expedite the algorithm development and to reveal the state of the art in automatic particle selection, a bakeoff was held in which 12 representative groups in the field submitted results of particle selection, either manually or automatically, using a common image dataset containing KLH particles. The results were then tabulated in a confusion matrix where both the false positive rates

and false negative rates were calculated for each participant's results against every other result. In addition, images outlined with particles picked by different participants were made publicly available using a web-based particle selection tool.

The 10 different algorithms tested in the bakeoff can be more or less grouped into two categories: those based on template matching and usually requiring a initial reference structure and those based on image feature recognition without the requirement of 3D reference structures. Several approaches from both categories achieved a high performance in selecting side view KLH particles in the common dataset. Although selecting KLH particles is a relatively "easy" problem to approach, as the particles are large, symmetric and readily visible, the bakeoff did serve as a common basis for a productive discussion at the workshop and a starting point toward establishing representative benchmark particle datasets as well as setting up criteria for evaluating algorithms for automatic particle selection. It is agreed that both well-annotated benchmark particle datasets and agreed-upon criteria for evaluating particle selection methods are essential aspects to the overall success of fully automated particle selection. Therefore, it was agreed that the infrastructure set up to support the bakeoff should be maintained and extended to include larger and more varied datasets and more criteria for future evaluations.

Selecting different particles may require different approaches. Given the specific nature of the dataset, algorithms that work well in selecting KLH particles in the bakeoff might perform completely differently on other datasets. The generalization of the ability of various approaches reported in this paper will remain to be tested in the future.

Acknowledgments

The Multidisciplinary Workshop on Automatic Particle Selection for Cryo-electron Microscopy held at The Scripps Research Institute, April 24–25, 2003, La Jolla, California and the bakeoff event were supported by the National Resource for Automated Molecular Microscopy which is supported by the National Institutes of Health though the National Center for Research Resources' P41 program (RR17573).

References

Boisset, N. et al., 1998. Overabundant single-particle electron microscope views induce a three-dimensional reconstruction artifact. Ultramicroscopy 74, 201–207.

Canny, J., 1986. A computation approach for edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8, 679–698.

Carragher, B., Kisseberth, N., Kriegman, D., Milligan, R.A., Potter, C.S., Pulokas, J., Reilein, A., 2000. Leginon: an automated system for acquisition of images from vitreous ice specimens. J. Struct. Biol. 132, 33–45.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., Leith, A., 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. 116 (1), 190–199.

Freund, Y., Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory. Eurocolt'95, Springer-Verlag, Berlin, pp. 23–37.

Glaeser, R.M., 1999. Review: Electron crystallography: present excitement, a nod to the past, anticipating the future. J. Struct. Biol. 128, 3–14.

Guan, W., Ma, S, 1998. A list-processing approach to compute Voronoi diagram and Euclidean distance transform. IEEE Trans. Pattern Anal. Machine Intelligence 20 (7), 757–761.

Hall, R.J., Patwardhan, A., 2004. A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. J. Struct. Biol. 145, 19–28.

Henderson, R., 1995. The potential and limitations of neutrons, electrons, and X-rays for atomic resolution microscopy of unstained biological macromolecules. Q. Rev. Biophys. 28, 171–193.

Horn, B., 1986. Robot Vision. MIT Press, Cambridge, MA.

Huang, Z., Penczek, P.A., 2004. Application of template matching technique to particle detection in electron micrographs. J. Struct. Biol. (in press).

Kohonen, T., 1989. Self-Organization and Associative Memory, third ed. Springer-Verlag, Berlin-Heidelberg/New York/Tokio.

Lata, R.K., Penczek, P., Frank, J., 1995. Automatic particle picking from electron micrographs. Ultramicroscopy 58, 381–391.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision, Corfu, Greece, pp. 1150–1157.

Ludtke, S.J., Baldwin, P.R., Chiu, W., 1999. EMAN: semiautomated software for high-resolution single particle reconstructions. J. Struct. Biol. 128, 82–97.

Mallick, S.P., Yuanxin, Z., Kriegman, D., 2004. Detecting articles in cryo-EM micrographs using learned features. J. Struct. Biol. (in press).

Nicholson, W.V., Glaeser, R.M., 2001. Review: automatic particle detection in electron microscopy. J. Struct. Biol. 133, 90–101.

Potter, C.S., Chu, H., Frey, B., Green, C., Kisseberth, N., Madden, T.J., et al., 1999. Leginon: a system for fully automated acquisition of 1000 micrographs a day. Ultramicroscopy 77, 153–161.

Roseman, A.M., 2003. Particle finding in electron micrographs using a fast local correlation algorithm. Ultramicroscopy 94, 225–236.

Roseman, A.M., 2004. FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. J. Struct. Biol. (in press).

van Heel, M., 1983. Detection of objects in quantum-noise-limited images. Ultramicroscopy 7, 331–342.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hawaii, vol. I, 8–14 December, 2001, pp. 511–518.

Volkmann, N., 2004. An approach to automated particle picking from electron micrographs based on reduced representation templaces. J. Struct. Biol. (in press).

Wade, R.H., 1992. A brief look at imaging and contrast transfer. Ultramicroscopy 46, 145–156.

Wong, H.C., Chen, J.D., Mouche, F., Rouiller, I., Bern, M., 2004. Model-based particle picking for cryo-electron microscopy. J. Struct. Biol. (in press).

Zhu, Y., Carragher, B., Mouche, F., Potter, C.S., 2003. Automatic particle detection through efficient Hough transforms. IEEE Trans. Med. Imaging 22, 1053–1062.