Available online at www.sciencedirect.com
ScienceDirect
Journal of Structural Biology 157 (2007) 168–173
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

SIGNATURE: A single-particle selection system for molecular electron microscopy

James Z. Chen *, Nikolaus Grigorieff

Howard Hughes Medical Institute, W. M. Keck Institute for Cellular Visualization, Rosenstiel Basic Medical Sciences Research Center, Department of Biochemistry, Brandeis University, 415 South Street, Waltham, MA 22454-9110, USA

Received 6 February 2006; received in revised form 12 June 2006; accepted 15 June 2006
Available online 17 June 2006

Abstract

SIGNATURE is a particle selection system for molecular electron microscopy. It applies a hierarchical screening procedure to identify molecular particles in EM micrographs. The user interface of the program provides versatile functions to facilitate image data visualization, particle annotation and particle quality inspection. The system design emphasizes both functionality and usability. This software has been released to the EM community and has been successfully applied to macromolecular structural analyses.

Keywords: Single-particle electron microscopy; Automatic particle selection

1. Introduction

Advances in single-particle electron microscopy (EM) have enabled the visualization of macromolecular structures at sub-nanometer resolution. This imaging technique analyzes 2D projections of the target molecule at various orientations to derive a 3D density model. The method is applicable to structural analysis of macromolecules with a mass of 200 kDa or more (Henderson, 1992). In single-particle EM, data processing begins with particle selection. The collected particle dataset is then subjected to particle alignment, particle classification, 3D reconstruction and model refinement. Conventionally, the particles are identified by manual annotation, which is difficult to reproduce and is prone to subjective bias. A computational algorithm will not only relieve researchers from the laborious and mundane task, but also generate objective and consistent results. In this publication, we present a computational screening system that strives to produce high quality particle datasets for EM structure determination.

Automated particle selection has been a subject of active research in the EM community (Nicholson and Glaeser, 2001; Zhu et al., 2004), and the majority of methods developed so far can be categorized into two classes: template-matching and pattern-recognition. The program FindEM (Roseman, 2003), an example for the template-matching method, calculates the local correlation between the micrograph and a set of predefined references to identify particle candidates. The program Selexon (Zhu et al., 2003), an example of the pattern-recognition approach, detects geometric features (e.g., edges, shapes) of the particles in the micrograph. Ideally, the goal of algorithmic particle screening is to fully automatically label particles from input electron micrographs without error. In practice, however, it has been recognized that the existing methods cannot entirely eliminate false-positives and false-negatives, as the algorithms are frequently fooled by edges, contaminants and other defects in electron micrographs. User intervention is still indispensable in order to obtain a high quality particle dataset. Therefore, in order to increase the efficiency and accuracy of particle selection

* Corresponding author.
E-mail addresses: jzchen@brandeis.edu (J.Z. Chen), niko@brandeis.edu (N. Grigorieff).

1047-8477/$ - see front matter
doi:10.1016/j.jsb.2006.06.001

for single-particle EM, improvements should come from both the algorithm design and software engineering—better computational algorithms can reduce the manual effort required for post-editing, and a user-friendly interface can expedite the process whenever the manual editing is called for.

The computational screening algorithm presented here is based on the template-matching method. It employs a hierarchical approach to improve the success rate of particle selection. Its user interface provides flexible functions to facilitate data visualization, particle annotation and quality inspection. In the following sections, the algorithm and the program implementation, SIGNATURE, will be introduced first. Its validation based on both synthetic and experimental micrographs will be presented. At the end, a few practical issues regarding the application of SIGNATURE will be discussed.

2. Methodology and validation

The proposed method selects particles from an EM micrograph according to a template image set defined *a priori*. The algorithm includes a set of hierarchical screening stages using various matrices: (1) the local-density-correlation function (LCF), (2) the spectrum-correlation function (SCF), and (3) inter-particle distance restraint.

The LCF measures the local density similarity between a micrograph and a particle template. A mask can be customized to exclude regions beyond the template particle. The LCF follows the formulation by Roseman (2003) and is rewritten here as

$$\mathrm{LCF}(x) = \frac{1}{N_T \sigma(I_x)} \langle M_T \otimes T, I \rangle_x, \tag{1}$$

where

$$\sigma^2(I_x) = \frac{1}{N_T} \langle M_T, I^2 \rangle_x - \left(\frac{1}{N_T} \langle M_T, I \rangle_x \right)^2, \tag{2}$$

and

$$\langle A, B \rangle_x = \sum_i A_i \times B_{i+x}. \tag{3}$$

In the above equations, I is the micrograph image and T is the particle template. M_T is the template mask and the term $(M_T \otimes T)$ represents the masked template image. N_T is the total number of effective pixels under the mask M_T. It is assumed that the template image has been normalized to $N(0, 1)$ (mean=0, s.t.d.=1) under the mask. The term $\langle A,B \rangle_x$ can be efficiently evaluated via Fourier transform (FT):

$$\langle A, B \rangle_x = FT^{-1}[FT(A) \times FT^*(B)]. \tag{4}$$

A relatively high LCF score alone cannot fully justify a particle's candidacy, as a good density correlation does not necessarily guarantee its shape similarity to that of the template. A more rigorous evaluation of resemblance is to examine the profile of the correlation function around the local maxima, since its distribution reflects the shape of the particle object. For a true particle location, the LCF pattern should be similar to the auto-correlation function (ACF) of the template image itself, assuming a clean background. This comparison can be quantified as yet another correlation evaluation between LCF and ACF. The auto-correlation function of an image is the inverse Fourier transform of its own power-spectrum, hence the name "spectrum-correlation function" (SCF):

$$\mathrm{SCF}(x) = \frac{1}{N_S \sigma(\mathrm{LCF}_x)} \langle M_S \otimes \mathrm{ACF},\ \mathrm{LCF} \rangle_x, \tag{5}$$

where

$$\sigma^2(\mathrm{LCF}_x) = \frac{1}{N_S} \langle M_S,\ \mathrm{LCF}^2 \rangle_x - \left(\frac{1}{N_S} \langle M_S,\ \mathrm{LCF} \rangle_x \right)^2, \tag{6}$$

and

$$\mathrm{ACF}(x) = \langle T, T \rangle_x \sim N(0, 1). \tag{7}$$

The definition of the SCF follows that of the LCF—the micrograph image I_x is substituted by LCF_x, the template image T by ACF, and the template mask M_T by the ACF mask M_S. The ACF in the above equation has also been normalized to $N(0,1)$.

Even though LCF and SCF appear very similar in their definition, they are independent metrics in measuring the similarity between an object and a template. The LCF function evaluates the intensity agreement between two images, but does not consider the order of pixel arrangement. Therefore, for a given template, two objects could have drastically different shapes, but still share the same LCF value. This ambiguity can be resolved by the SCF function, which enforces the intra-pixel relationship of the template (via auto-correlation) in the image to be matched. The difference and the complementary property of LCF and SCF can be illustrated by the following simple example (Fig. 1). A binary disc (1 inside the disc and 0 for the background) serves as a template and is matched to two objects: in image-A, half of the pixels in the original disc are randomly selected and reset to 0; in image-B, all the pixels in the lower half of the original disc are reset to 0. A direct LCF calculation (with a tight mask) produces 0.5 for both cases. The auto-correlation function of the template and its cross-correlation function with each of the objects are displayed below the respective images. Apart from a scaling factor, the cross-correlation function between the template and image-A is very similar to the auto-correlation function of the template itself. After image normalization under the mask, the SCF score is 0.99 for image-A, but only 0.56 for image-B.

To measure the overall quality of the image-template-matching, an S-factor, ranging between 0.0 and 1.0, is intuitively introduced as:

$$S(x) = \mathrm{LCF}(x) \times \mathrm{SCF}(x). \tag{8}$$

Since both LCF and SCF should be high for a true particle in the image, a high S-factor will consequently be associated with its location in the micrograph.

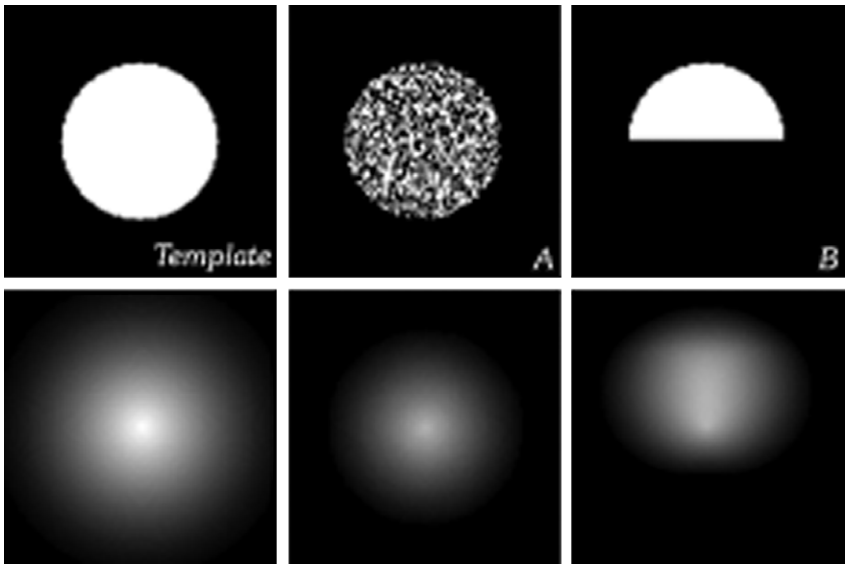

Fig. 1. The LCF function measures pixel-wise intensity correlation between two images, but is insufficient in determining the overall shape similarity of the objects.

In a complete micrograph screening, a stack of templates at various orientations are systematically searched. For each particle template (T_i) at each rotation angle (α), the $\mathrm{LCF}_{i,\alpha}$ and $\mathrm{SCF}_{i,\alpha}$ are calculated and consecutively screened with the user-prescribed thresholds. The S-factors of image locations passing both tests will be recorded into the map $S_{i,\alpha}(x)$. Once all the templates have been rotationally searched, the highest S-factor at each pixel in the micrograph is stored into the map $S(x)$:

$$S(x) = \max\{S_{i,\alpha}(x)\}. \tag{9}$$

The local maxima (in a region corresponding to the particle's size) of $S(x)$ indicates potential particles.

Because valid particles should be clearly separated, a distance restraint is applied at the end to avoid any overlap. A disc comparable to the diameter of the particle is drawn around each peak of $S(x)$. Whenever two or multiple discs conflict, all the corresponding peaks will be removed from the particle selection.

As a further effort to reduce false-positives, an algorithm termed "digital-gel-filtration", which is a figurative reference to size-exclusion chromatography, has been implemented. The method originates from the observation that the total density mass of particle projections should be invariant of the particle orientation when the image is properly normalized. In practice, however, because the image data is often quite noisy, and elastic, weak-phase electron scattering is merely a theoretical approximation for EM image formation, there will be a spread in the density mass histogram of the selected particles. A band in the histogram can be identified by the user, and the particles inside the band tend to be more homogeneous. The digital-gel-filtration function provides a guidance to improve the overall quality of a particle dataset, and it has been incorporated as an optional tool.

The above algorithm has been integrated in the program SIGNATURE, in which all parameters can be modified through a graphical user interface (GUI, Fig. 2). The program GUI also supports manual particle selection/inspection and provides informative data visualization functions. In addition to the interactive operation, this program can run in a "number crunching" computing mode that enables batch-processing and therefore can greatly enhance productivity through scripting and distributed computing. In order to reduce false-positives, a pre-screening procedure has also been introduced to automatically mask out edges and regions with abnormal variance in a micrograph. The masked area will not be processed in particle screening. A set of drawing tools enables the user to block out apparent non-particle regions in the micrograph. When the

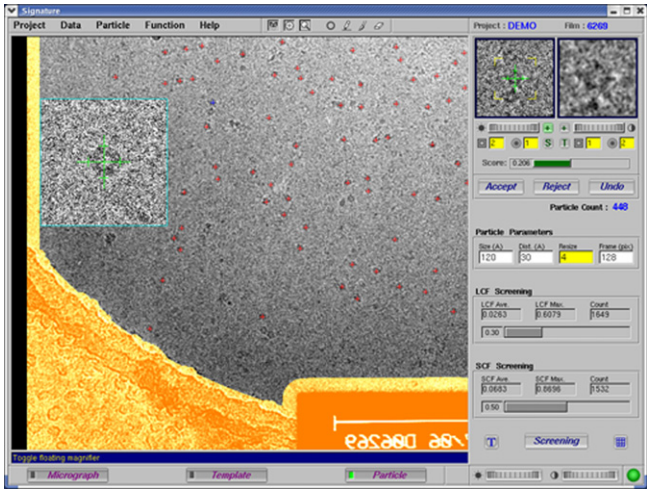

Fig. 2. The graphic user interface of SIGNATURE. The image mask, the particle selection and a floating magnifier are shown overlying the micrograph display.

screening is complete, the selected particles can be sequentially displayed on the GUI at various scales and filtering levels for the user to make the accept–reject decision. Another feature of SIGNATURE is an integrated "particle stack editor", which can be used to display and modify (extract, delete, resize, etc.) the selected particles. The output from the editor is automatically synchronized with the particle annotation on the micrograph. At the current implementation, the selected particles can be exported to files as both coordinates and image stacks. SIGNATURE has been tested on Linux (32/64-bit applications), Mac OS X, and MS Windows platforms. The program website (including a user manual) is at www.brandeis.edu/~jzchen/Signature.

SIGNATURE has been validated using both synthetic and experimental electron micrographs. The synthetic dataset provides the ground-truth that can be rigorously controlled. In the precision test, synthetic micrographs and particle templates are generated from 2D projections of the EM model of *N*-ethyl maleimide sensitive factor (NSF) (Fürst et al., 2003) at various orientations. The NSF molecule is about 120 Å in diameter, and the map voxel size is 3.5 Å. Seven projections (15° apart along the polar angle) are calculated and put into a template stack (Fig. 3). Four of them (#1, #3, #5 and #7) are used to produce the synthetic micrographs—multiple copies of each template are shifted and rotated (between 0° and 180°), then randomly patched into an empty image array (without overlapping). Gaussian noise at various levels (SNR measured by the ratio of variance between signal and noise) is subsequently introduced. The three unused projections (#2, #4 and #6) serve as "decoy templates" in the test—a robust algorithm should be able to unambiguously match particles in the micrograph to their respective templates. In the screening, the rotational search step is set at 1°, ranging from 0° to 360°. The result is summarized in Fig. 4.

Because of image pixelization and noise, it is conceivable that minor misalignment will occur when compared to the ground-truth. Since the particle diameter corresponds to

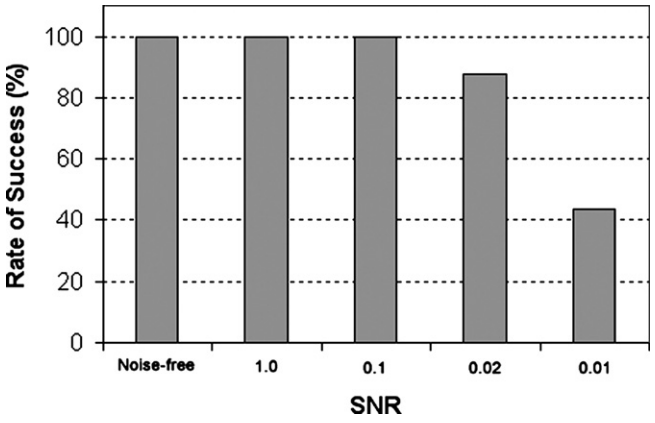

Fig. 4. Precision test on synthetic EM micrographs.

~35 pixels in the micrograph, an error within 3-pixel in translation and 5-degree in rotation will be deemed acceptable. It is observed from the chart that the algorithm performs well for SNR as low as 0.1: all particles are correctly recovered from the micrograph and are related to the appropriate templates. When the SNR decreases to 0.02, templates 1, 2 and 3 sometimes cannot be differentiated. Still, particles themselves are identified according to the tolerance level. Because the SNR of experimental cryo-EM micrographs is normally at or above this level (Frank and Al-Ali, 1975), and is much higher for negative-stain samples, this method is sufficiently accurate and is applicable to experimental EM micrographs. The algorithm has been further tested with an even lower SNR at 0.01. At that point, severe misalignments occur, together with both false-negative and false-positive selections. Although the error can be mitigated to some extent by experimenting with the LCF and SCF thresholds, at the level where signal is overwhelmed by noise, the "truth" is no longer well defined.

The program has also been tested on the cryo-EM images of Keyhole Limpet Hemocyanin (KLH), the dataset used in the "particle selection bakeoff" organized by Zhu et al. (2004). There are 82 defocus pairs in total, and only

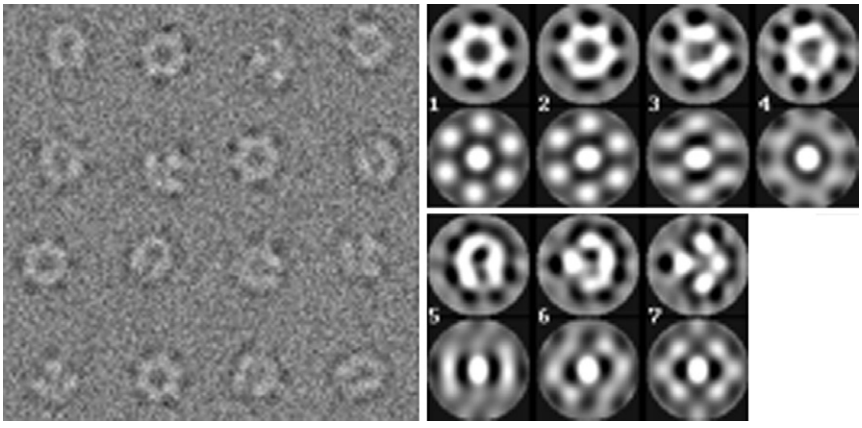

Fig. 3. One synthetic micrograph at SNR = 0.10 and seven template images. Templates 1, 3, 5, and 7 are used in the synthetic micrograph. The ACF of each particle template is displayed under the image.

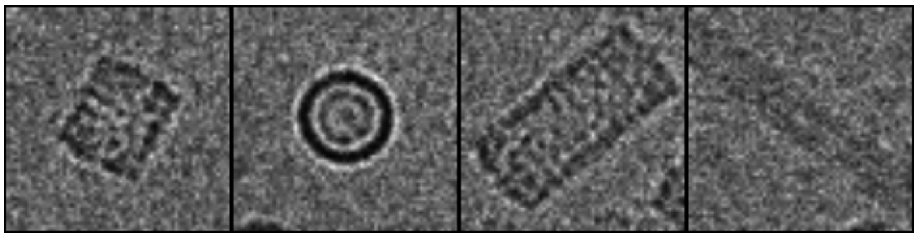

Fig. 5. Particle templates cropped out directly from the cryo-EM test images of KLH. From the left, KLH side-view (the target), KLH top-view, aggregated KLH particles, and TMV segment. The last three cases are used as "traps" to reduce false-positives.

the far-from-focus images are used in our test. Both top-views and side-views of the KLH particles are present in the images, together with TMV filaments and varying degree of contamination. The goal is to identify only the clean side-views of KLH. Particle templates are extracted directly from the test images using an editing function of SIGNATURE (Fig. 5). In addition to the target (KLH side-view), a few other instances—a KLH top-view, an aggregated KLH side-view, and a TMV segment—are also included in the template set. After the algorithmic screening, SIGNATURE can exclusively write out those particles matched to the side-view of KLH for subsequent analysis. Although SIGNATURE provides versatile manual editing functions, only algorithmic screening is applied in this test.

Following the evaluation procedure used in the bakeoff, the particle screening result from SIGNATURE is compared with the manual annotation (by Mouche) included in the test dataset. For LCF = 0.15, SCF = 0.20, 20 Å inter-particle distance, and 3° interval in rotational search, the average false-positive-rate (FPR) over the 82 images is 12.9%, and the false-negative-rate (FNR) is 9.8%. The LCF/SCF parameters are identified by experimenting on the first three images to produce the best result. Because of the uneven quality and contrast of the testing images, the fixed LCF/SCF thresholds may not be optimal in some cases. We identified about one dozen such images and adjusted the thresholds on an individual basis (still without any manual editing). This extra, but minor, effort improves FPR to 10.7%, and FNR remains unchanged. These results compare favorably to the published statistics from the bakeoff (average FPR = 21.7%, average FNR = 16.2%, quoted from Table 2, column-4, Zhu et al., 2004). In another test designed to determine the effectiveness of the SCF screening, the SCF stage is turned off (LCF = 0.15, SCF = 0.0 for all cases), upon which FPR rises to 15.5%, and FNR drops slightly to 8.1%. This indicates that SCF indeed contributes to the quality improvement in the algorithmic particle selection.

SIGNATURE has already been put into practical use in single-particle EM structural analysis. Since its beta release in 2004, several research groups have applied the program for particle screening in studying macromolecular systems, which include HPV (M. Wolf, personal communication), DNA origin recognition complex (X. Zhang, personal communication), ribosomes (C.M. Spahn and R. Beckmann, personal communication), human transferrin receptor (Y. Cheng, personal communication), Arp2/3 complex (O. Sokolova, personal communication), and exon junction complex (M.E. Stroupe, personal communication).

3. Discussion

When LCF is close to 1.0 in template-matching, the false-positive rate is normally quite low. However, in EM particle screening, because the image SNR is well below 1.0, LCF is around 0.2 at the best and a simple LCF thresholding is insufficient in detecting good particles. To improve the accuracy, we have introduced an SCF function to complement the LCF function. Since LCF measures pixel-wise intensity correlation, and SCF measures overall shape similarity, the combined function can reduce false-positives and produce more reliable particle datasets.

By definition, particle selection entails only a binary decision: YES or NO. The template-matching algorithm provides further information regarding the in-plane rotation of a particle candidate. Therefore, when the rotational search is done using a small step size, the method implemented in SIGNATURE can also be used for particle alignment. 2D projections of a 3D density model from an initial reconstruction can be used as templates to screen for a much larger particle dataset. Then, with the known projection orientation and the in-plane rotation, a better model can be built and refined. This is ongoing research and will be reported in a future publication.

Applying SIGNATURE for particle screening requires pre-defined particle templates, which may come from three sources: (1) 2D projections from a known, low-resolution density model; (2) image class-averages of a small dataset selected by manual annotation; and (3) characteristic particle images cropped directly from an electron micrograph (as demonstrated in the test on the KLH particles). Once an initial model is established, new templates can be generated from model 2D projections, and repeated screening can proceed to refine the selection and/or to identify more particles for another reconstruction at higher resolution.

Particle heterogeneity presents a major challenge to high-resolution EM structure determination. The task of sorting particles into homogeneous subsets is often left to a later stage of data processing (for example, particle classification). When models of various structural conformations become available, SIGNATURE can be used to

differentiate particles from a heterogeneous dataset, and therefore, improve the efficiency in the subsequent classification. To achieve this, several template sets originating from various models can be used simultaneously in the program for particle screening. Upon completion, particles matched to a specific template subset can be exported as a homogeneous dataset. This function can also be applied to reduce the false-positives in particle screening: ''trap templates'' can be set and the particles matched to those templates will be discarded automatically.

Acknowledgments

The development and refinement of SIGANTURE have benefited tremendously from user feedback. In particular, we thank Matthias Wolf, Duncan Sosa, Elizabeth Stroupe at Brandeis University, the Walz group at Harvard University, and the Spahn group and the Beckmann group at Humboldt University for their constructive comments and suggestions. The authors gratefully acknowledge financial support from the National Institutes of Health, Grant 1P01 GM-62580. J.Z.C. designed and implemented SIGNATURE. N.G. was supported by a research fellowship for the Humboldt Foundation. N.G. thanks Christian Spahn and his colleagues for hosting his sabbatical.

References

Frank, J., Al-Ali, L., 1975. Signal-to-noise ratio of electron micrographs obtained by cross correlation. Nature 256, 376–379.

Fürst, J., Sutton, R.B., Chen, J.Z., Brünger, A.T., Grigorieff, N., 2003. Electron cryomicroscopy structure of *N*-ethyl maleimide sensitive factor at 11 Å resolution. The EMBO Journal 22 (17), 4365–4374.

Henderson, R., 1992. Image contrast in high-resolution electron microscopy of biological macromolecules: TMV in ice. Ultramicroscopy 46, 1–18.

Nicholson, W.V., Glaeser, R.M., 2001. Review: automatic particle detection in electron microscopy. Journal of Structural Biology 133, 90–101.

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

Zhu, Y., Carragher, B., Potter, C.S., 2003. Automatic particle detection through efficient hough transforms. IEEE Transactions on Medical Imaging 22 (9), 1053–1062.

Zhu, Y., Carraghera, B., Glaeserb, R.M., Fellmanna, D., Bajajc, C., Bernd, M., Mouchea, F., Haase, F., Hallf, R.J., Kriegmang, D.J., Ludtkeh, S.J., Mallickg, S.P., Penczeki, P.A., Roseman, A.M., Sigworth, F.J., Volkmannl, N., Potter, C.S., 2004. Automatic particle selection: results of a comparative study. Journal of Structural Biology 145, 3–14.