TMaCS: A hybrid template matching and classification system for partially-automated particle selection

Jianhua Zhao, Marcus A. Brubaker, and John L. Rubinstein

Version  Post-print/Accepted Manuscript


Copyright/License  This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

How to cite TSpace items

Always cite the published version, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the author manuscript from TSpace because you cannot access the published version, then cite the TSpace version in addition to the published version using the permanent URI (handle) found on the record page.

This article was made openly accessible by U of T Faculty. Please tell us how this access benefits you. Your story matters.
TMaCS: A hybrid template matching and classification system for partially-automated particle selection

Jianhua Zhao 1,2, Marcus A. Brubaker 3,4,*, and John L. Rubinstein 1,2,5,*

1Molecular Structure and Function Program, The Hospital for Sick Children Research Institute,
2Department of Medical Biophysics, University of Toronto, 3 Department of Computer Sciences, University of Toronto, 4 Toyota Technological Institute at Chicago, 5 Department of Biochemistry, University of Toronto. * corresponding authors

Short title: TMaCS: A program for partially-automated particle selection

Editorial correspondence to:
John Rubinstein
The Hospital for Sick Children Research Institute
555 University Avenue
Toronto, Ontario Canada M5G 1X8
Telephone: 001-416-813-7255, Fax: 001-416-813-5022
Email: john.rubinstein@utoronto.ca
Abstract

Selection of particle images from electron micrographs presents a bottleneck in determining the structures of macromolecular assemblies by single particle electron cryomicroscopy (cryo-EM). The problem is particularly important when an experimentalist wants to improve the resolution of a 3D map by increasing by tens or hundreds or thousands of images the size of the dataset used for calculating the map. Although several existing methods for automatic particle image selection work well for large protein complexes that produce high-contrast images, it is well known in the cryo-EM community that small complexes that give low-contrast images are often refractory to existing automated particle image selection schemes. Here we develop a method for partially-automated particle image selection when an initial 3D map of the protein under investigation is already available. Candidate particle images are selected from micrographs by template matching with template images derived from projections of the existing 3D map. The candidate particle images are then used to train a support vector machine, which classifies the candidates as particle images or non-particle images. In a final step in the analysis, the selected particle images are subjected to projection matching against the initial 3D map, with the correlation coefficient between the particle image and the best matching map projection used to assess the reliability of the particle image. We show that this approach is able to rapidly select particle images from micrographs of a rotary ATPase, a type of membrane protein complex involved in many aspects of biology.

Keywords: cryo-EM; software; particle picking; particle selection; automated; machine learning; template matching; single particle
Introduction

Single particle electron cryomicroscopy (cryo-EM) is an increasingly important tool for the structural analysis of large protein complexes. In cryo-EM, images of randomly oriented macromolecular assemblies embedded in a thin layer of vitrified buffer are imaged at liquid nitrogen temperature or colder with a transmission electron microscope. The images of individual protein particles correspond approximately to 2D projections of the 3D Coulomb potential of the protein complex. In order to avoid damage to high-resolution features in the specimen, micrographs are recorded with a low electron exposure (Glaeser, 1971) resulting in images with a low signal-to-noise ratio (SNR) at high resolution. By determining the orientations of the protein particles in images, the images can be combined to calculate a 3D map of the protein complex.

Initial maps of protein particles may be calculated with as few as several thousand particle images. In principle, it should be possible to calculate 3D maps with ~3 Å resolution from fewer than 13,000 particle images (Henderson, 1995). However, larger numbers of particle images often allow improved resolution in 3D maps (Armache et al., 2010) due to a variety of sources of image deterioration (Henderson, 1992) and, for some complexes, the inability to determine particle orientations precisely (Baker et al., 2012; Henderson et al., 2011). In a few cases, near-atomic resolution maps of highly-symmetric virus particles have been generated after averaging millions of the asymmetric units of the virus particles (Yu et al., 2011; Zhang et al., 2010). Collection of large numbers of particle images presents a bottleneck in the process of determining 3D structures. At the initial stages of map calculation it is often beneficial to manually select particle images: this process gives the microscopist an opportunity to inspect
closely the quality of images and become familiar with attributes of the protein complex under investigation. However, manual particle image selection is a tedious task and the edifying effect of manual selection is outweighed by its disadvantages when tens or hundreds of thousands of particle images are needed.

Computers are often suited better to highly-repetitive tasks, such as particle image selection, than are humans. Several groups have implemented automated particle image selection algorithms with varying levels of success (Zhu et al., 2004). Many of these approaches have been reviewed elsewhere (Langlois and Frank, 2011; Zhu et al., 2004) and a full review is beyond the scope of this article. Programs have included generative approaches, such as template matching, that attempt to identify particles in micrographs based on similarity to reference images (Chen and Grigorieff, 2007; Roseman, 2003; Volkmann, 2004); unsupervised approaches that attempt to recognize particles in micrographs by characteristics that distinguishes them from the background and other objects (Adiga et al., 2004; Voss et al., 2009); and discriminative methods that attempt to learn to distinguish particles from non-particles in a micrograph based on positive and negative example images (Arbelaez et al., 2011; Langlois et al., 2011; Mallick et al., 2004; Ogura and Sato, 2004; Sorzano et al., 2009).

In our laboratory, we use cryo-EM to study membrane protein complexes such as rotary ATPases (Baker et al., 2012; Benlekbir et al., 2012; Lau and Rubinstein, 2010; Lau et al., 2008; Lau and Rubinstein, 2012; Rubinstein et al., 2003). These complexes present unique challenges in cryo-EM for several reasons. The complexes can be relatively small by cryo-EM standards (600 to 900 kDa), which results in low SNRs in micrographs. The complexes are intrinsic
membrane proteins and consequently require detergents for solubilization, which further reduces image contrast (Rubinstein, 2007; Schmidt-Krey and Rubinstein, 2010). Finally, although the rotary ATPases are asymmetric, they possess a pseudo-two-fold axis between their soluble \( F_1/V_1/A_1 \) regions and their membrane-bound \( F_0/V_0/A_0 \) regions, and consequently the ‘middle’ of the protein particles are regions of low protein density rather than high protein density, as is typical for other protein complexes. The challenge in identifying good particle images from a micrograph means that two or more different people tasked with selecting particle images from a micrograph often do not select the same set of images. These complications also make the computational problem of automated particle image selection more challenging. In order to automate the particle image selection process, we tested many of the available software packages but were not able to establish conditions that recovered particle images sufficiently well to satisfy our requirements. Previous data used as a test-set for particle image selection programs (Zhu et al., 2004) were often higher contrast than rotary ATPase images, making it difficult to assess which programs might be useful for selecting particle images from our micrographs.

Rather than attempting to adapt an existing particle image selection program to the rotary ATPases images, we elected to build an algorithm that would be well suited to the problem. As a result, we have developed an approach to particle image selection for the situation where a preliminary 3D map already exists for the protein complex under investigation. The program uses the same file formats and conventions as the program Frealign (Grigorieff, 2007) and our own image analysis programs (Benlekbir et al., 2012; Lau and Rubinstein, 2012) allowing the program to be easily integrated into the map refinement pipeline. A flow chart of the steps performed by the program is shown in Fig. 1. ‘Template images’ are generated from the initial
3D map and used for exhaustive searching of micrographs by calculation of cross-correlation functions. A low cross-correlation threshold is used to identify a large set of candidate particle images, including many non-particle images, from each micrograph. A subset of these candidate particle images is labeled manually as particles or non-particles in a purpose built graphical user interface (GUI) and these data are used to train a machine learning classifier employing a support vector machine (SVM) (Cortes and Vapnik, 1995). Machine learning algorithms offer a data-driven approach to discriminating between classes of images, as opposed to attempting to manually encode rules to distinguish a particle image from a non-particle image. Support vector machines implicitly find a hyperplane that separates the two classes of images in a high dimensional and potentially non-linear feature space. SVMs are efficient to train, fast to evaluate and, given sufficient data, capable of learning highly complex classifiers. The classifier is iteratively refined by interactive correction of its output within the GUI. The remaining candidate particle images are then labeled as particles or non-particles by the algorithm. In a final stage of the analysis, the particle images identified by machine learning are subjected to projection matching to obtain a correlation coefficient between the particle image and its best matching projection of the initial 3D map. The correlation coefficient from projection matching is used to rank the selected particles from most reliable to least reliable. Due to the combination of template matching and machine learning for classification, we have named our program TMaCS (template matching and classification system).

Analysis of the behavior of TMaCS shows that it performs almost as well as a human particle image selector for rotary ATPase images, with several independent human selectors opting to discard fewer than 6% of the images selected by the program. By using this approach,
moderately sized datasets can be rapidly extended into much larger datasets with little need for repetitive work by the experimenter. New datasets can also be acquired automatically for samples that are closely related to an existing 3D map, such as a structure of a complex with a new co-factor or binding-partner attached. Therefore, the TMaCS approach will help an experienced microscopist generate large datasets of particle images that can be incorporated into 3D map calculation with minimal human intervention.

**Methods**

*Data and programming languages*

Images of keyhole limpet hemocyanin from *Megathura crenulata* (KLH) were acquired from the National Resource for Automated Molecular Microscopy (nramm.scripps.edu) (Zhu et al., 2003). KLH is ~8 MDa with dimensions of ~300 x 300 x 400 Å. Images of the V/A-ATPase from *Thermus thermophilus* were from data recorded to calculate a 9.7 Å map of the complex (Lau and Rubinstein, 2012). V/A-ATPase is ~600 kDa with dimensions of ~100 x 100 x 200 Å. Image pretreatment and template matching were performed with new programs written in modern Fortran and compiled with the gfortran compiler from the GNU compiler project. The interactive particle-labeling program and GUI were implemented in Python and the machine learning algorithm was implemented in Python.

*Pretreatment of images*

Micrographs of KLH and V/A-ATPase were 2048 x 2048 pixels and ~13000 x ~12000 pixels with pixel sizes of 2.2 Å and 1.4 Å, respectively. Micrographs were downsized 4 x 4 and 2 x 2 for the KLH and V/A-ATPase datasets, respectively. For the V/A-ATPase images, micrographs
were masked by setting regions of high and low local variance to the average micrograph pixel value (Chen and Grigorieff, 2007). Once masking parameters, which are adjustable in the program, were selected manually with a few micrographs, they could be applied without change to all micrographs in the dataset. This process involved setting the values of the pixels in 240 x 240 pixel blocks to the micrograph average where the variance was below 0.02 and above 1.1 times the image variance. This local variance thresholding was applied twice in order to better distinguish desirable and undesirable regions in micrographs. Low frequency gradients in the V/A-ATPase micrographs were removed by subtracting from each pixel the average value of a 500 x 500 pixel box centered on that pixel.

Template matching

Projections of 3D maps of KLH from *M. crenulata* (http://ami.scripps.edu/redmine/projects/ami/wiki/KLH_dataset_I) and the V/A-ATPase from *T. thermophilus* (Lau and Rubinstein, 2012) (EMD-5335) were created with the program Genproj_Fspace (J. Rubinstein, unpublished software), which uses the Euler angle convention from Spider (Frank et al., 1996) and Frealign (Grigorieff, 2007). A template image of the KLH complex was generated by averaging the $\theta = 90^\circ$, $\phi = 0^\circ$ projection of the complex with the $\theta = 90^\circ$, $\phi = 36^\circ$ projection of the complex. In-plane rotations were generated with $\psi \in \{0,180^\circ\}$ in steps of $10^\circ$. The same procedure used to generate template images for KLH was performed for the V/A-ATPase, except that the Euler angles used were $\theta = 90^\circ$ and $\phi \in \{0,360^\circ\}$ in steps of $10^\circ$ to create the initial template with $\psi \in \{0,360^\circ\}$ in steps of $10^\circ$ to produce 36 in-plane rotated template images. Normalized cross-correlation functions between each template image and the pretreated micrographs were calculated by padding the template image with zeros to the
size of the entire micrograph. Cross-correlation functions were calculated including spatial frequencies between 500 and 20 Å for the KLH micrographs and 300 and 20 Å for the V/A-ATPase micrographs. The highest correlation coefficient among all the templates for each position in the micrograph was recorded in a combined cross-correlation map. Micrograph positions with cross-correlation coefficients above 0.01 and 0.05 were windowed from the untreated micrographs in 272 x 272 pixel and 256 x 256 pixel boxes for KLH and V/A-ATPase, respectively, and used as candidate particle images. For suppression of high cross-correlation values near to cross-correlation peaks during particle windowing, suppression radii of 300 and 140 Å for KLH and V/A-ATPase were used, respectively.

Classification by machine learning

Initially, 100 candidate V/A-ATPase particle images from the template-matching algorithm were labeled with the GUI to assign particle images and non-particle images. Images were bandpass filtered in Fourier space, masked using a circular mask with a radius of 60 pixels, downsized to 32 x 32 pixels, and the mean of the image was subtracted from each pixel in the image. Principal component analysis (PCA) (Jolliffe, 2002) was performed by singular value decomposition (SVD) (Press et al., 1992) using the linear algebra algorithm numpy.linalg.svd in NumPy (Oliphant, 2007). PCA functions as a dimensionality reduction step when the number of training examples is smaller than the number of pixels in each image and whitens the data, removing correlations between pixels, when the number of training examples is greater than or equal to the number of pixels in each image. The PCA coefficients for each image were then used to train a binary support vector machine (SVM) classifier (Cortes and Vapnik, 1995) with a radial basis function kernel using an implementation of the algorithm from the Machine Learning Toolkit
(MILK) in Python (Luis Pedro Coelho, http://packages.python.org/milk/). Subsequently, sets of 100 candidate particle images were displayed to the user with a particle or non-particle label assigned by the learning algorithm. These labels were manually corrected until 1018 candidate particle images and 5182 non-particle images had been inspected. These manually approved labels were used to train the classifier for the classification of the remaining candidate particle images.

**Ranking of images by projection matching**

After completion of template matching and machine learning to select particle images, the output particle images were subjected to a final ranking by projection matching. V/A-ATPase images were downsized to produce 64 x 64 pixel images (5.6 Å/pixel) and the average perimeter pixel value was subtracted from each image. The best possible cross-correlation coefficients between each image and 144000 different projections of a V/A-ATPase cryo-EM map (Lau and Rubinstein, 2012) were determined by projection matching with the program Search_Fspace (Benlekbir et al., 2012).

**Computational performance**

All steps in the program were performed using a home built workstation with an Intel Core i7 2.93 GHz processor and 3.9 GB RAM memory running the Ubuntu 11.04 operating system. Pretreatment and template matching required ~1 sec/micrograph and ~10 min/micrograph for the KLH and V/A-ATPase datasets, respectively. Training of a classifier with V/A-ATPase candidate particle images required ~7 min and classification of 7408 test images required ~10 min. Manually training the SVM classifier can be expected to take an experienced microscopist
approximately one day of work. Projection matching with Search_Fspace required ~13 sec/particle image.

**Results**

*Overview and definitions*

The TMaCS particle image selection approach consists of the sequential application of three algorithms: a template matching algorithm, a machine learning algorithm, and a projection matching algorithm. For clarity, we define *template images* as the images derived from the initial 3D map that are used by the template matching algorithm to search the micrographs and produce a combined cross correlation function for each micrograph. The template matching algorithm generates many *candidate particle images*. The machine learning algorithm labels the candidate particle images from template matching as either *particle images* or *non-particle images*. Finally, the projection matching algorithm ranks the images labeled by machine learning as particle images in order of reliability.

*Template matching*

Two sets of micrographs were used to characterize the template matching image selection algorithm. The first set of micrographs was comprised of images of keyhole limpet hemocyanin (KLH) recorded on a charge-coupled device (CCD) at 120 kV (Zhu et al., 2003). In the KLH images (e.g. Fig. 2A) regions corresponding to protein particles have relatively high SNRs, are easy to detect by eye, and appear fairly homogeneous. Side views of KLH particles in these images were manually labeled previously (Zhu et al., 2003) in order to allow development of particle image selection algorithms. The second set of micrographs were from cryo-EM of
detergent-solubilized V/A-ATPase particles recorded on photographic film at 200 kV and digitized with a densitometer (Lau and Rubinstein, 2012). Film images often contain features such as a text label, regions of the carbon support substrate, and defects from developing the film. A particularly dramatic example of a film with these artifacts is shown in Fig. 2Bi (artifacts indicated by arrows). This second set of micrographs is typical of the data acquired in our laboratory for the structural analysis of F- (Baker et al., 2012; Lau et al., 2008; Rubinstein et al., 2003), V- (Benlekbir et al., 2012), and V/A-ATPases (Lau and Rubinstein, 2010; Lau and Rubinstein, 2012). Images of the particles in the micrographs have lower SNRs than the KLH images and it is often ambiguous which densities correspond to acceptable particle images. A set of acceptable particle images in these micrographs were labeled previously during a project that led to the determination of the V/A-ATPase structure at subnanometer resolution (Lau and Rubinstein, 2012).

Micrographs were masked as described in the Methods section to remove areas of carbon, large areas of ice contamination, film development defects, and text labels on the film (Fig. 2Bii, artifacts indicated by arrows). Template images from the KLH map (Fig. 2C) and V/A-ATPase map (Fig. 2D) were generated from the map projections as described in the Materials and Methods section. A larger number of templates can be easily incorporated in the template matching procedure at the expense of increased computation time. The map projection process included multiplication of the Fourier transform of the projection with an estimated contrast transfer function (CTF) for the microscope to account for the effects of the microscope’s point-spread function (Grigorieff, 2007). Here, a constant underfocus of 3 μm was used for all images, but defocus values specific to each micrograph could be incorporated into the analysis.
However, we expect that this change would not affect the accuracy of the template matching.

Cross-correlation functions were generated between the micrographs and each of the template images. For each micrograph, a combined cross-correlation function was calculated that gives, at every position, the maximum value from all of the template-specific cross-correlation functions at that position. Examples of these combined cross-correlation functions are shown in Fig. 3Ai and Bi for KLH and V/A-ATPase, respectively. This approach to template-based particle image selection inherently ranks the candidate particle images from each dataset by assigning each candidate particle image with a cross-correlation score. Candidate particle images were selected from the micrographs by systematically extracting windows in order of cross-correlation coefficient. After each extraction, nearby maxima were suppressed with a mask that had a diameter approximately equal to the diameter of the protein complex. Candidate particle images were extracted until either no more candidates with correlation above an arbitrary cutoff could be found or a specified maximum number of candidate particle images per micrograph were extracted. An example of the output from this step of the algorithm is shown in Fig. 3. For the KLH data (Fig. 3Aii), every candidate particle image extracted from a single micrograph is shown in order of decreasing correlation with the templates. For the V/A-ATPase (Fig. 3Bii), every 400th extracted candidate image from a single micrograph is shown in order of decreasing correlation with the templates. As can be seen from the figure, the KLH particle images correspond to side views of KLH particles in the earlier, higher correlation coefficient, candidate images. In comparison, the selected V/A-ATPase candidate images correspond primarily to features in the image that are not V/A-ATPase particles. In order to assess quantitatively the precision of this template-based approach against the previous manual labeling
of particle images, precision-recall curves were plotted for datasets of 81 micrographs of KLH and a single V/A-ATPase micrograph. Precision-recall curves (Langlois and Frank, 2011) plot the ability of a program to correctly classify a particle image against the fraction of labeled particle images that are retrieved from the image. In these curves, precision is defined by

\[
\text{precision} = \frac{TP}{TP + FP},
\]

where TP is the number of true positives, or true particle images, selected, and FP is the number of false positives, or features that are selected that do not correspond to particles. Recall, the fraction of available true particle images retrieved from the micrographs, is defined by

\[
\text{recall} = \frac{TP}{TP + FN},
\]

where FN is the number of false negatives, or particle images that the program fails to identify in the micrographs. Precision recall curves were generated from the template matching algorithm by varying the correlation threshold at which a candidate particle images was accepted as a particle. As can be seen from the precision-recall curves in Fig. 3Aiii and Biii, the template-based method works significantly better for the KLH images than for the V/A-ATPases images. Based on these results, it is apparent that the template-based particle image selection method is sufficient for identifying KLH particle images from micrographs, as seen previously with other particle image selection programs used with this dataset (Chen and Grigorieff, 2007; Roseman, 2004; Volkmann, 2004), but it is not sufficient for selection of V/A-ATPase particle images.
Machine learning with a support vector machine

In order to better identify particle images from the micrographs of V/A-ATPase, a machine learning approach was implemented. The output of the template-based particle image selection program with a low cross-correlation threshold was used to generate 256039 candidate particle images from 306 micrographs of V/A-ATPase. From these candidate particle images, a test set consisting of 7408 candidate particle images was set aside for characterizing the algorithm. A purpose-written graphical user interface (GUI) driven program was developed to aid the use of the machine learning algorithm. A screenshot from the GUI is shown in Fig. 4. With the GUI, 100 randomly drawn particle images from the training set are displayed on the computer screen and manually labeled by the user as particles (Fig. 4 green boxes) or non-particles (Fig. 4 red boxes). The images are masked, bandpass filtered in Fourier space, downsized, and normalized prior to further processing. The program performs principal component analysis (PCA) (Jolliffe, 2002) so that each of the images is described by their principal component coefficients. The principal component coefficients are used to train a classifier using a support vector machine (SVM) (Cortes and Vapnik, 1995). SVMs are kernel-based algorithms that maximize the distance from a decision boundary to the ‘support vectors’ in a high dimensional feature space, while minimizing the error associated with the decision boundary for inseparable datasets. The classification score is the signed distance to the SVM decision boundary. Images with a classification score greater than zero are classified as particles, and those with a classification score less than zero are classified as non-particles.

A subsequent set of 100 candidate particle images are then drawn randomly from the remaining
candidate images and classified as particle images or non-particle images before being displayed by the GUI. The user manually corrects the classification where necessary, thereby providing more information for further training of the classifier. This process of PCA, training of the classifier, and correction of the classification is repeated until diminishing returns on training are observed, in this case with 1018 images labeled as particle images and 5182 images labeled as non-particle images being inspected or corrected by the user. The number of particle images required for optimal training of the classifier will likely depend on the nature of the image, with fewer training examples required for images that have high SNRs.

After training, the classifier was used to label the candidate particle images in the test set. Out of the 7408 candidate particle images in the test set, 1446 images were labeled by the classifier as particle images. Fig. 5A shows every 10th image selected as a particle image by the machine learning algorithm. By default, candidate particle images that receive a positive score from the SVM are classified as particles while negative scores are classified as non-particles. However, precision recall curves can be generated from the SVM classifier by varying the default decision threshold of 0 to include fewer or more images as particles. A precision-recall curve for the SVM classification of particle images (Fig. 5C) shows reasonable precision with recall when compared to the manually selected dataset from the corresponding micrographs. In the curve, precision decreases slowly with recall until ~ 80 % recall where there is a sharp drop in precision. Inspection of the particle images associated with recall values of 80 % or higher revealed many particle images that were selected off center by the template matching algorithm. These off-center particle images tended to have particularly poor contrast, which likely led to correlation of the reference images with noise during template matching. It is probable that these
particle images would also behave poorly during 3D map refinement by projection matching. Template matching with micrographs of KLH, which had high SNRs, did not lead to selection of off-center particle images.

Although significantly better than the template-based approach shown in Fig. 3Biii, the machine learning classification appears to have, at best, ~80 % precision in the precision-recall curve. However, inspection of the output particle images (e.g. Fig. 5A) shows that an overwhelming majority are particle images that an experienced human selector who has worked with rotary ATPase particles would designate as ‘good’ particle images. The reason for this apparent discrepancy is that, as described previously, even two experienced human microscopists will often not select the same particle images from a micrograph but will deem each other’s selections as appropriate. A more relevant statistic for an automated particle image selection algorithm in this situation is what fraction of the output images are acceptable to a human particle image selector. Here, precision retains the same definition as above but false positive particles are identified by inspection. To test this form of precision for the algorithm, we had three experienced rotary ATPase microscopists inspect the output of the machine learning algorithm and manually label undesirable, or false positive, images. Images could be labeled as undesirable for any reason, including crowding of particles, an apparent defect in the particle, or an apparent defect in the image. With this approach, the number of false negative particles cannot be determined and consequently recall is not known. The results of this analysis are shown in Fig 5D as the average precision as a function of particle number, where the particle images are ranked according to their output score from the classifier. As can be seen from Fig 5D, at most ~6 % of particle images were labeled as undesirable, or false positives, by the
average microscopist. Examples of undesirable particle images are shown with red borders in
Fig. 5A. Analysis of the accuracy of particle orientation determination (Baker et al., 2012;
Henderson et al., 2011) has shown that 25 to 50 % of rotary ATPases particle images have
orientations that are incorrectly determined, but these errors only affect the appearance of the
map in the worst cases (Baker et al., 2012). Misaligned particle images and inclusion of
undesirable images should affect the final 3D map in similar ways, suggesting that including 6 %
of undesirable images in a map should not significantly affect the map. The observation that the
three different microscopists labeled different numbers of particle images as false positives (116,
36, and 114) illustrates the subjectivity of manual particle selection for these images. It is also
worth noting that the total number of particle images identified by the machine learning
algorithm, an average of ~140 per micrograph, corresponds well with the average of ~150
particle images selected by a microscopist in the initial analysis of the data (Lau and Rubinstein,
2012). This number of particles was obtained with the default SVM classifier score of 0 used as
the decision threshold. The SVM=0 threshold corresponds to a precision of 65 % with 70 %
recall as assessed by comparison with the manually selected dataset.

**Ranking of particle images by projection matching**

As can be seen from Fig. 5A, the undesirable particle images in the dataset after classification,
although accounting for at most 6 % of the dataset, are spread throughout the last 40 % of the set
when particle images having a classification score greater than zero are arranged according to
their scores. In order to test if a better ranking could be obtained so that a fixed fraction of
poorly ranked particle images could be discarded, we subjected the candidate particle images to a
round of projection matching with Search_Fspace (Benlekbir et al., 2012). The projection
matching utilized nearly 150,000 projections of the 3D map, a number of reference images that would be prohibitively slow if used in the initial template matching step in TMaCS. Particle images were rearranged according to their correlation with an optimal matching projection of the reference 3D map. Again, the three experienced microscopists identified undesirable, or false positive, particle images in the dataset. The result of this analysis is shown in Fig. 5B, where the undesirable particle images are shown with red borders. As can be seen from Fig. 5B, ranking of particle images by their correlation against a matching projection of a 3D map is better able to discriminate desirable images from undesirable images as indicated by the clustering of the undesirable images towards the last 20% of the dataset. It is again interesting to note that the same experienced microscopist, when presented with the same particle images in a different sequence, often does not choose the same number as undesirable, further emphasizing the difficulty of this particle image selection task.

**Discussion**

The process we have described for partially automated particle image selection easily lends itself to batch processing. Although some initial investment is required to establish a data processing pipeline, such as the generation of template images, determination of threshold values for cross-correlation, and training of a classifier, these preparatory steps need only to be performed once. Subsequently, micrographs can be processed quickly to increase the size of the particle image dataset. As with manual particle image selection, there is some danger that the TMaCS approach will be limited by preconceived notions of what particles ‘should’ look like and will consequently select only a subset of particle images, possibly distorting the final 3D map. Therefore, this approach does not remove the requirement for a reliable starting 3D map,
generation of a sufficiently diverse series of reference templates, and familiarity with the
different views of the 3D complex when training the machine learning classifier.

Classification has been used previously to improve the quality of output from template matching
particle selection algorithms (Roseman, 2004). Similarly, a variety of machine learning
algorithms have been employed in the task of automated particle image selection (Arbelaez et al.,
2011; Langlois et al., 2011; Mallick et al., 2004; Ogura and Sato, 2004; Sorzano et al., 2009). Of
these approaches, only two groups attempted to select small asymmetric particles by cryo-EM
that could be considered as challenging as the rotary ATPase particles used here. The other
algorithms were tested on particle images that came from cryo-EM of large and/or symmetric
particles or negative stain EM. Ogura and Sato (2004) characterized their algorithm with a 200
kDa sodium channel while Albelaez et al. included images of the 900 kDa RNA polymerase II in
their test data. Ogura and Sato (2004) provided little quantitative evidence to support the
effectiveness of their neural-network approach to particle selection. The approach of Arbelaez et
al. (2011) for selecting cryo-EM images of RNA polymerase II was moderately successful with
~50 % precision at 70% recall.

Most of the earlier approaches attempt to use machine learning to identify particles directly from
micrographs, with positive examples selected interactively (Arbelaez et al., 2011; Mallick et al.,
2004; Ogura and Sato, 2004; Sorzano et al., 2009). Negative examples were either selected
interactively (Mallick et al., 2004; Ogura and Sato, 2004), taken as all regions in a micrograph
not selected by the user (Sorzano et al., 2009), or derived from the selected positive examples
and their surrounding regions (Arbelaez et al., 2011). Overall, the approach we describe here is
most similar to that of Langlois et al (2011) in that we use one algorithm to generate candidate particle images and machine learning algorithm to classify them. Consequently, both TMaCS and the approach of Langlois et al. (2011) can provide both positive and realistic negative examples of particle images for optimal training of the classifier. Langlois et al (2011) used a modified Difference of Gaussians (DoG) picker to provide candidate particle images to an AffinityRank learning algorithm. Our approach differs in that the user interacts with the machine learning algorithm to improve the quality of the classification iteratively. With our program also, any other candidate particle image generating approach could be used to provide images to the machine learning algorithm. However, in our experience, the highly successful DoG picker approach (Langlois et al., 2011; Voss et al., 2009) tends to select individual F₁/V₁/A₁ regions or F₀/V₀/A₀ regions from rotary ATPases, rather than the centers of the particles, and is therefore not suited to this particular problem. This issue makes the template-matching algorithm a better choice for rotary ATPases, and presumably for other protein complexes that have their centre of mass in the middle of a low-density portion of the complex. Furthermore, a final projection-matching step in TMaCS provides a cross-correlation score that can be used to exclude the non-particle images that were erroneously labeled as particles by the learning algorithm. This procedure can increase the accuracy of particle image selection and speed up subsequent 3D map refinement.

The accuracy of TMaCS will depend on the quality of the micrographs and particle images. For micrographs with high SNRs, template matching is often sufficient for particle image selection, as we demonstrated with KLH. Band-pass filtering images can have a strong effect both for template matching and classification. Other program options, such as masking and downsizing
images, affect computational speed but appear to have little effect on the accuracy of particle image selection. The TMaCS approach is also not expected to be particularly sensitive to inaccuracies in the 3D map used to generate template images for template matching. This tolerance is apparent from the way we produced template images by averaging 36 different side views of the V/A-ATPase.

Although TMaCS was developed for selecting rotary ATPase particle images, we predict that the program will be useful for many other specimens. Template matching and classification are both general techniques that can be easily adapted to a wide range of specimen images. The ability to separate the different steps in TMaCS further increases the flexibility of the system and facilitates its use for particle selection for other specimens. However, we expect that the components of the approach may become computationally infeasible with structurally heterogeneous and impure samples. Template matching would require an impractical number of templates to identify structurally heterogeneous particle images in micrographs. Classification by SVM would similarly require an impractical number of training examples when it is required to distinguish between a structurally heterogeneous protein and a similarly sized impurity in the sample.

In combination, machine learning algorithms for automated particle image selection are beginning to provide the tools necessary to obtain large cryo-EM datasets for high-resolution structure determination and analysis of heterogeneous protein structures. TMaCS provides one such route to obtaining a large dataset, performing almost as well as a human particle selector. The system has now been used with several different datasets in our laboratory and could be
applied easily to other datasets. The program includes a user friendly GUI to facilitate the interactive training of the SVM classifier and all steps of the process are easily implemented into a pipeline for map refinement with Frealign and our own compatible programs.

The TMaCS software has been made available through the website:

http://www.sickkids.ca/research/rubinstein/

**Author contributions**

JZ wrote the computer programs and analyzed the data. MAB contributed parts of the machine learning program. MAB and JLR conceived and supervised the research. JZ, MAB, and JLR wrote the manuscript.

**Acknowledgements**

We thank Drs. Ryan Lilien and Peter Rosenthal for a critical reading of the manuscript. JZ was supported by a studentship from the Hospital for Sick Children. MAB was supported by a postdoctoral fellowship from the Natural Science and Engineering Research Council of Canada. JLR was supported by a New Investigator Award from the Canadian Institutes of Health Research and an Early Researcher Award from the Ontario Ministry of Research and Innovation. This research was supported by the Natural Science and Engineering Research Council Discovery Grant 401724-12. Images of the V/A-ATPase were acquired during research funded by Canadian Institutes of Health Research Grant MOP 81294.

**References**


Glaeser, R.M., 1971. Limitations to significant information in biological electron microscopy as a result of radiation damage. J Ultrastruct Res 36, 466-482.


**Figure Captions**

**Figure 1.** Flowchart describing the steps in TMaCS. Projections are generated from a reference 3D map and used to perform template matching with a masked micrograph. A classifier is iteratively trained in an interactive user interface with a subset of the candidate particle images selected from template matching. Once trained, the classifier is used to identify particle images in the rest of the dataset. Projection matching is then used to rank the selected particle images from most to least reliable.

**Figure 2.** Pre-processing of micrographs and generation of template images for template matching. (A) An example micrograph from the KLH dataset shows the desired side views of KLH (rectangular views), undesired top views of the complex (circular views), and the tobacco mosaic virus particles used as a magnification standard (elongated rod shapes). (B) An example of an atypical low-quality micrograph from the V/A-ATPase dataset before (i) and after (ii) variance thresholding and local background subtraction. The variance thresholding removes micrograph text labels, regions of carbon support substrate, ice contamination, and film developing artifacts from micrographs, indicated by the arrows. (C) Template images derived from a 3D KLH cryo-EM map (http://ami.scripps.edu/redmine/projects/ami/wiki/KLH_dataset_I) used for template matching. (D) Templates images derived from a 3D cryo-EM map of V/A-ATPase (Lau and Rubinstein, 2012). (Scale bars: 300 Å).

**Figure 3.** Template matching search of micrographs. (Ai) A combined cross-correlation map
from the KLH micrograph from Fig. 1A with the KLH template images from Fig. 1C shows several clear maxima corresponding to side views of the KLH particles. (ii) Candidate particle images selected from the micrograph are primarily side views of the complex (green bordered images) with a few undesirable views of the complex (red bordered images) extracted with smaller correlation coefficients. (iii) A precision-recall curve constructed against a manually-selected list of particle images from 81 micrographs shows high precision with recall. (Bi) A combined cross-correlation map from the V/A-ATPase dataset shows many more peaks of nearly equivalent intensity than the KLH combined cross-correlation map. (ii) Representative output from the cross-correlation search of one micrograph showing every 400th candidate particle image reveals that the majority of selected regions from the micrograph do not correspond to views of the V/A-ATPase complex. (iii) A precision-recall curve constructed against a manually-selected list of V/A-ATPase particle images from a single micrograph shows poor precision with recall. (Scale bars: 300 Å).

**Figure 4.** Graphical user interface (GUI) for training the support vector machine (SVM) classifier. Stacks of candidate particle images are imported into the program and 100 images are drawn randomly from all of the candidates and displayed. These initial images are labeled manually by clicking images to turn red-bordered images (non-particles) to green-bordered images (particles), and vice-versa. Subsequent windows containing 100 candidate particle images are then displayed with red- or green-borders already tentatively assigned in order to continue training the classifier. Once a sufficient training set has been inspected, the remaining candidate particle images are labeled by the classifier.
Figure 5. Classification of candidate particle images from template matching of ten V/A-ATPase micrographs. (A) Representative output from classification of images showing every 10th image that was automatically labeled as a ‘particle’, ranked in order of classification score. Most of the images correspond to typical V/A-ATPase particle images (manually labeled with green borders) while a small fraction of images still correspond to undesirable images (manually labeled with red borders). The first undesirable image to appear in this ranking is at position 981 out of 1440 particle images. The scale bar corresponds to 300 Å and applies to part B also. (B) The particle images were subjected to projection matching against the original 3D map and ranked according to their resulting cross-correlation score. In this ranking, the first undesirable image occurs at position 1330 out of 1440, suggesting that removal of the worst 8% of particle images in this ranking would eliminate almost all of the remaining undesirable images. (C) A precision-recall curve constructed against a manually-selected list of particle images appears to indicate a lower precision than what is expected from the output in parts A and B, probably due to different particle images being selected in the manual particle image selection and automated selection. (D) Particle image selection precision was assessed in a subjective manner by having three experienced microscopists inspect the 1440 images labeled as particles and either agree with or reject the label. The average interpolated precision was plotted against particle number in the ranked list (D). The curve shows high precision with particle number, consistent with what is observed from the output images in 4A and 4B. The error bars correspond to one standard deviation.
Figure 2
Figure 4